CYBERNETICA

Rahastanud Euroopa Liit

Algoritmilise kallutatuse riskihalduse juhend

Juhend

Version 1.1

01.10.2025

D-16-571

Avalik

Projektijuhid: Liina Kamm (Cybernetica AS)
Henrik Trasberg (Justiits- ja Digiministeerium)
Sofia Paes (Justiits- ja Digiministeerium)

Autorid: Dan Bogdanov
Paula Etti
Hiroki Kaminaga
Liina Kamm
Tanel Mällo
Tanel Pern
Fedor Stomakhin

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

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

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

Kuupäev	Versioon	Kirjeldus
13.06.2025	0.5	Esimene mustand
09.07.2025	1.0	Kallutatuse hindamise riskihalduse juhend
01.10.2025	1.1	Mitmed täiendused ja muudatused rakendajate tagasiside põhjal.

Sisukord

1 Sissejuhatus

1.1 Eesmärk

Tehisintellekti (AI) tehnoloogia areneneb kiiresti ja sellega on võimalik saavutada mitmesuguseid majanduslikke, keskkondlikke ja ühiskondlikke hüvesid ([1] pp 4). AI-tehnoloogiat õigesti rakendades võib see anda konkurentsieelise ja toetada nii ühiskonda kui ka keskkonda ([1] pp 4). Samas toob AI-tehnoloogia kaasa infoturvariskid, masinõppe- või suure keelemudeliga seotud algoritmilised riskid, õiguslikud, ühiskondlikud ja eetilised riskid. Riskide realiseerumise tagajärjel tekkiv kahju võib olla varaline või mittevaraline, sh füüsiline, psühholoogiline, ühiskondlik või majanduslik ([1] pp 5).

AI-süsteemide algoritmiline kallutatus (*bias*) on risk, mida on vaja eraldi käsitleda, sest see võib märkamatult moonutada otsustus- ja hinnanguprotsesse ning viia ebaõiglaste, ebatäpsete või ebaefektiivsete tulemusteni. Kallutatuse ignoreerimine süvendab ebavõrdsust, õõnestab usaldust ja võib tekitada vigu. Kallutatusega tegelemine on hädavajalik õiglustunde, täpsuse ja otsuste usaldusväärsuse tagamiseks igas kontekstis (loe näiteid jaotises 3).

AI-süsteeme, sh üldotstarbelist tehisintellekti (GPAI), võetakse nii EL-i kui ka Eesti avalikus sektoris kasutusele üha enam, et tõhustada juhtimist ja teenuseid. Levib nii ametlik kui ka mitteametlik kasutamine – paljud ametnikud kasutavad eraviisiliselt GPAI-tööriistu juba enne ametlike tehisintellektistrateegiate kehtestamist. Kuigi tehisintellekti kasutatakse laialdaselt teenuste täiustamiseks ja sisemise tõhususe suurendamiseks, toob GPAI kaasa keerulisi probleeme valitsemise (*governance*), eetika ja õigusliku vastavuse osas. [2]

Hiljutised uuringud näitavad, et GPAI kasutuselevõttu soodustab tehnoloogiline areng, juhtkonna toetus ja kodanike ootused. Edukas integreerimine sõltub siiski ka eetilisest teadlikkusest, organisatsiooni kohanemisvõimest ja sisemiste oskuste arendamisest. [2]

EL ja liikmesriigid töötavad välja juhiseid AI-süsteemide, sh GPAI, ohutuks, läbipaistvaks ja seaduslikuks kasutamiseks, keskendudes vastutusele, andmekaitsele ja inimkontrollile. Koostalitleva Euroopa määrus [3] ja tehisintellekti Euroopa/maailmajao tegevuskava [4, 5] („The AI Continent Action Plan“) toetavad seda suunda, soodustades piiriülest koostööd ja usaldusväärse AI kasutuselevõttu. [2] Tegevuskava eesmärk on kasutada tugevat tööstust ja väga head talendibaasi, et muuta Euroopa võimsaks tehisintellekti innovatsiooni mootoriks [6].

Avalik sektor on juhtrollis, et parandada teenuste kvaliteeti erinevates sektorites, nagu tervishoid, haridus, õigus ja haldus. Tehisintellekti oskuslik kasutamine võib ennetada diskrimineerimist ja suurendada juurdepääsetavust nt erivajadustega inimestele. Õigluse tagamiseks ja AI rakenduste kallutatuse vähendamiseks on oluline kasutada mitmekesiseid ja kvaliteetseid andmekogusid [2].

1.2 Käsitlusala

Algoritmilise kallutatuse riskihalduse töövahend keskendub algoritmilise ja AI-süsteemide kallutatuse riskide haldamisele. Töövahend koosneb kolmest osast: juhendist, mis kirjeldab tehistinetllektisüsteemide ja nende kallutatuse olemust ja tausta ning vaatleb kallutatuse tuvastamise ja leevendamise võimalusi; metoodikast, mis annab detailsed juhised riskihaldusprotsessi ülesseadmiseks ja läbiviimiseks; ning töölehest, mis lihtsustab riskihalduseks vajaliku teabe dokumenteerimist. Töövahendi käsitlusalas on eelkõige süsteemid, mida kasutatakse organisatsioo-

nides. Eraisikute omaks tarbeks käitatavatele algoritmilistele ja AI-süsteemidele kehtivad väiksemad nõuded. Vaatleme riskihaldusprotsessi üldisena, sest süsteemid võivad olla väga erinevad ja kallutatuse allikas võib olla erinev. Olenevalt süsteemi juurutusest, võib organisatsioonil puududa võimalus riske tehnoloogiliste vahenditega hinnata, seega peab neid käsitlema üldisemalt.

1.3 Terminid ja lühendid

1.3.1 Terminid

AI-süsteem
: masinapõhine süsteem, mis töötleb päringu sisendandmeid, et luua vastuseid (nt prognoose, sisu, soovitusi või otsuseid), mis võivad mõjutada füüsilist või virtuaalset keskkonda. AI-süsteemide põhiomadus on nende võime teha järeldusi ja tuletada seoseid sisendandmetest, kasutades selleks masinõppemudeleid

algoritm
: eeskiri mingi tegevuse sammhaaval sooritamiseks

kaitsepiirded (*guardrails*)
: tehnoloogia või raamistik, mis toimib turvakontrollpunktina, hinnates kasutaja päringut või sisendteksti ja mudeli genereeritud vastuseid lähtudes eelnevalt määratud ohutusreeglitest. Kaitsepiirete eesmärk on reaalajas ennetada kahjuliku, sobimatu või teemavälise sisu genereerimist, kui mudeli ohutushäälestusest selleks ei piisa

kallutatus (*bias*)
: ühtede objektide, isikute või rühmade käsitlemise süstemaatiline erinevus teistega võrreldes; käsitlemine on ükskõik milline tegevus, sealhulgas tajumine, vaatlus, esitus, prognoosimine või otsustus

masinõppemudel
: spetsiifiline treenitav algoritm või algoritmide kogum, mille abil saab sisendväärtuste põhjal genereerida väljundandmeid

seletavus (*explainability)*
: AI-süsteemi omadus väljendada oma tulemusi mõjutavaid olulisi tegureid inimestele arusaadaval viisil; on mõeldud vastama küsimusele „Miks?“ püüdmata väita, et valitud tegevusviis on tingimata optimaalne

1.3.2 Lühendid

AI – *artificial intelligence*, tehisintellekt

AI määrus – tehisintellekti käsitlev määrus (EL) 2024/1689

AvTS – avaliku teabe seadus

art – artikkel

EL – Euroopa Liit

ELL – Euroopa Liidu leping

ELTL – Euroopa Liidu toimimise leping

GPAI – *general-purpose artificial intelligence*, üldotstarbeline tehisintellekt

IKS – isikuandmete kaitse seadus

IKÜM – isikuandmete kaitse üldmäärus (EL) 2016/679

KüTS – küberturvalisuse seadus

lg – lõige

LLM – *large language model*, suur keelemudel

p – punkt

pp – põhjenduspunkt

TNVS – toote nõuetele vastavuse seadus

ÜRO – Ühinenud Rahvaste Organisatsioon

1.4 Dokumendi struktuur

Algoritmilise kallutatuse riskihalduse juhend aitab metoodika rakendajal mõista kallutatusega seotud mõisteid ja ka reaalseid ohte. Soovitame juhendi struktuuriga tutvuda enne metoodika ja sellega seotud töölehe täitmisega alustamist. Mitmes jaotises kirjeldab juhend maailmas teadaolevaid süsteeme, milles on esinenud algoritmilist kallutatust või kallutatud tehisintellekti komponente. Need juhtumikirjeldused aitavad lugejal mõista ohte ja neid paremini tuvastada ning ennetada.

Jaotises 2 selgitame, mis on AI-süsteem ja kuidas sobituvad sellesse algoritmid ja masinõppemudelid. Selgitame nii õiguslikust kui ka infotehnoloogisest vaatest AI-süsteemidega seotud pooli ja seda, kuidas mudelid ja süsteemid sünnivad. Jaotis toob (mitteammenduva) loetelu masinõppemudelite kvaliteedinäitajatest ja testkomplektidest ning selgitab ka kaitsepiirete mõistet. Kõik see aitab avada ka kallutatuse mõistet ning selgitada, kuskohas kallutatus masinõppemudeli või algoritmi elutsüklis tekib.

Jaotises 3 selgitame, miks peab kallutatust vältima. Selleks toome välja, milliseid kahjusid isikutele ja organisatsioonidele kallutatusest tulla võib. Sellele järgneb ülevaade seadustest, mille täitmiseks on vaja tegeleda ka kallutatusest põhjustatud kahjude vältimisega. Viitame rahvusvahelistele standarditele ja materjalidele tehisintellekti kallutatuse kohta.

Jaotis 4 näitab, kuidas algoritmilise kallutatuse riskihaldust siduda süsteemi teiste riskide haldusega. Meenutame riskihalduse protsessi ja anname soovitusi, kuidas AI-süsteemi kirjeldada. Selgitame, millist tüüpi kahjusid võib tuua algoritmilise süsteemi kallutatus ning otsustustes tehisintellektile liigne toetumine. Jaotise lõpetab selgitus selle kohta, millises süsteemi elutsükli etapis on tehisintellekti kallutatusega tegelemine mõistlikum ja odavam.

Jaotises 5 aitame metoodika rakendajat konkreetsetes olukordades ning kirjeldame praktilisi tööriistu. Esmalt aitame mõista, kas vaadeldav AI-süsteem on must kast (algoritmile või masinõppemudelile ja selle treeningandmetele juurdepääsu ei ole) või valge kast (evitajal on võimalik uurida või muuta algoritmi, masinõppemudelit või selle treenimiseks kasutatud andmeid). Seejärel selgitame, mida on võimalik teha üht või teist tüüpi süsteemi puhul kallutatuse vähendamiseks.

2 AI-süsteemid ja nende kallutatus

2.1 Millest AI-süsteemid koosnevad?

AI-süsteem on masinapõhine süsteem, mis kasutab masinõppemudeleid või neurovõrke, võib olla kohanemisvõimeline ning sellel võib olla mingil määral autonoomsus. Ta töötleb päringu sisendandmeid, et luua vastuseid (nt prognoose, sisu, soovitusi või otsuseid), mis võivad mõjutada füüsilist või virtuaalset keskkonda. AI-süsteemide põhiomadus on nende võime teha järeldusi ja tuletada seoseid sisendandmetest. AI-süsteemid erinevad algoritmilistest tarkvarasüsteemidest selle poolest, et viimased tegutsevad ainult eelmääratletud reeglite alusel. Nii AI-süsteemi kui algoritmilise süsteemi puhul vaatleme süsteemi laiemalt kui ainult konkreetset masinõppe mudelit (AI-komponenti) või otsusetoe reeglite kogumikku.

Näide. Häirekeskuse ohuhinnangute andmise abimees

Häirekeskuse ohuhinnangute andmise abimees on tehisintellektil põhinev rakendus, mis aitab sissetulevat kõnet transkribeerides ja varasemaid juhtumeid arvesse võttes pakkuda päästekorraldajale sündmuse tõenäoliseima tüüpjuhtumi. Süsteem on mõeldud toetama päästekorraldajat (inimest) ohuhinnangu andmisel, sest info kogumine piiratud ajaressursi tingimustel on inimese jaoks keeruline.

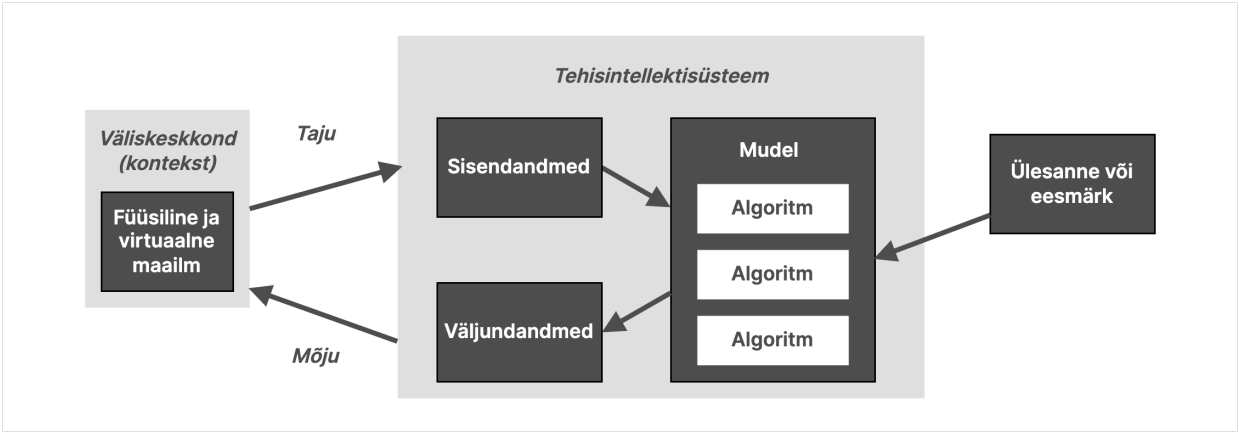
Masinõppemudel on spetsiifiline algoritm või algoritmide kogum, mida kasutatakse selleks, et genereerida või hinnata mingi sisendväärtuste põhjal tulemusi. Tema oluliseks omaduseks on treenitavus: mudel on võimeline treenimiseks kasutatud andmetest olulisi seoseid või mustreid ära tundma, et parandada tulemuste kvaliteeti. Masinõppemudel on AI-süsteemi keskne komponent, mis võimaldab tal õppida andmetest ja langetada otsuseid või anda soovitusi.

Näide. Ettevõtja varase hoiatamise teenuse prototüüp

Majandus- ja kommunikatsiooniministeerium ning Statistikaamet on loonud masinõppemudeli, mille abil on võimalik prognoosida, kas mõnel ettevõttel hakkab (tema majandussektori üldist seisu arvestades) halvemini minema. Kui lisaanalüüsi käigus leitakse oht, saadab arvutisüsteem ettevõttele varase hoiatuse, mis ei too kaasa õiguslikke tagajärgi, kuid võib aidata ettevõtjal oma äri järje peale saada.

Algoritm on eeskiri mingi tegevuse sammhaaval sooritamiseks.

Kallutatuse mõjust arusaamiseks on oluline mõista, kuidas algoritmid ja masinõppemudelid süsteemis paiknevad (vt joonis 1). Meie elukeskkond annab AI-süsteemile sisendväärtused, mis muudetakse digitaalseteks andmeteks ja antakse ette algoritmi(de)le või masinõppemudeli(te)le. Need täidavad inimese määratud ülesannet ning töötlevad andmeid, kuniks tekivad väljundandmed. Väljundandmete mõju maailmale võib olla otsene (kui AI-süsteemile on antud maailma mõjutamiseks hoovad) või kaudne (maailma muudab AI-süsteemi väljundi põhjal inimene). AI-süsteemil on mõlemal juhul meie elukeskkonnale mõju ning kui süsteem on kallutatud, on ka selle mõju kallutatud. Niisugusest kallutatusest arusaamiseks ja tagajärgede ennetamiseks ongi metoodika loodud.



Joonis 1. AI-süsteemi üldine mudel (kohandatud allikast [7]). Väliskeskkond on füüsiline ja virtuaalne. AI-süsteem tajub väliskeskkonda. Selle tulemusena tekivad sisendandmed. Sisendandmed antakse mudelile. Mudel, mis on treenitud mingi (eksplitsiitse või implitsiitse) eesmärgiga, loob väljundandmed. Väljundandmed mõjutavad omakorda väliskeskkonda, näiteks nende põhjal tehtavate otsuste kaudu

2.2 Kes AI-süsteeme loovad ja haldavad?

Euroopa Liit on tehisintellekti määruses defineerinud palju huvipooli (organisatsioone ja üksikisikuid) ning nende rollid, mida nad AI-süsteemidega seoses täidavad. Neid selgitab teabekast järgmisel lehel. Samas on AI-süsteemid oma loomult tarkvarasüsteemid, seega võime neist mõelda tavapärastes IT-süsteemide terminites.

Algoritmilisi ja AI-süsteeme luuakse inimeste töö lihtsustamiseks, kiirendamiseks ja ka asendamiseks. Seega on AI-süsteemidel oluline suhe ühiskonnaga ja eriti inimestega, kes neid kasutavad ning keda nende tulemid mõjutavad.

Selle metoodika käsitlusalas on ennekõike organisatsioonide loodud süsteemid. Eraisikute omaks tarbeks käitatavatele algoritmilistele ja AI-süsteemidele kehtivad väiksemad nõuded.

IT-süsteemide arendamiseks ja haldamiseks on kasutusel mitmeid organisatsioonimudeleid, mõned neist on ka standarditud. Järgmised rollid on algoritmilise ja tehisintellekti kallutatuse riskide haldamisel eriti olulised.

Teenusejuht, projektijuht. Igal süsteemil on isik, kes vastutab süsteemi loomise projekti või juba loodud teenuse või toote käitamise eest. See roll sobib hästi ka loodud süsteemi riskide eest vastutajaks, sest ta saab käsutada süsteemiga seotud ressursse. Suuremas meeskonnas saab ta riskihalduse ülesanded delegeerida näiteks turbejuhile, arhitektile või analüütikule.

Tooteomanik, teenuseomanik, analüütik ja arhitekt. Omanikud, analüütikud ja arhitektid tunnevad süsteemi kõige paremini ning teavad, kelle käest küsida sellega seotud üksikasju. Nende kaasamine on kallutatuse riskide haldusel väga oluline.

Spetsialistid. Programmeerijad, andmeteadlased, õigus-, küberturbe- või andmekaitsespetsialistid tunnevad süsteemi, toote või teenuse aspekte süvitsi. Neid on otstarbekas kaasata, kui on vaja välja selgitada konkreetse algoritmi, tehisintellektikomponendi või kasutusloo detaile.

Mainitud rollidesse võib leida inimesed erinevatest asutusest. Näiteks võivad tehnilisete rollide täitjad olla sisse ostetud arenduspartneri, tehisintellekti toote müüja või teenustaja juurest.

Tehisintellekti käsitleva määruse (EL) 2024/1689 kohaselt on huvipoolteks

- **inimesed** – ühiskond, ühiskonnagrupid, üksikisikud. AI määrus näeb ette, et tehisintellekt peab olema inimkeskne tehnoloogia, sh vahend, mis teenib inimesi ja mille lõppeesmärk on suurendada inimeste heaolu (pp 6);
- **pakkuja** – füüsiline või juriidiline isik, ametiasutus, ametkond või muu organ, kes arendab AI-süsteemi või üldotstarbelist tehisintellektimudelit või kellel on väljatöötatud AI-süsteem või üldotstarbeline tehisintellektimudel ja kes laseb selle turule või võtab AI-süsteemi kasutusele oma nime või kaubamärgi all kas tasu eest või tasuta (art 3 p 3). Määrust kohaldatakse sellistele pakkujatele, kes tegelevad EL-is AI-süsteemide turule laskmise või kasutusele võtmisega või üldotstarbeliste tehisintellektimudelite turule laskmisega, olenemata sellest, kas pakkuja on asutatud või asub liidus või kolmandas riigis (art 2(1)(a)). Määrust kohaldatakse ka kolmandas riigis asutatud või asuvale AI-süsteemi pakkujale, kui AI-süsteemi väljundandmeid kasutatakse EL-is (art 2(1)(c)), samuti pakkuja volitatud esindajatele, kes ei ole asutatud EL-is (art 2(1)(f));
- **volitatud esindaja** – füüsiline või juriidiline isik, kes asub või on asutatud EL-is ja kes on saanud AI-süsteemi või üldotstarbelise tehisintellektimudeli pakkujalt kirjaliku volituse vastavalt täita AI määrusega kehtestatud kohustusi ja sooritada menetlusi tema nimel ning on selle volituse vastu võtnud (art 3 p 5);
- **juurutaja** – füüsiline või juriidiline isik, ametiasutus, ametkond või muu organ, kes kasutab AI-süsteemi oma volituste alusel, välja arvatud juhul, kui viidatud süsteemi kasutatakse isikliku, mitte kutselise tegevuse jaoks (art 3 p 4). Juurutajaks saab lugeda AI-süsteemi kasutavat mis tahes füüsilist või juriidilist isikut, sh avaliku sektori asutust, ametit või muud organit, kelle volitusel süsteemi kasutatakse (pp 13). Määrust kohaldatakse EL-is asutatud või asuvale AI-süsteemide juurutajale (art 2(1)(b)). Määrust kohaldatakse ka kolmandates riikides asutatud või asuvale AI-süsteemide juurutajale, kui AI-süsteemi väljundandmeid kasutatakse EL-is (art 2(1)(c));
- **importija** – füüsiline või juriidiline isik, kes asub või on asutatud EL-is ja kes laseb turule sellise füüsilise või juriidilise isiku nime või kaubamärki kandva AI-süsteemi, kes asub või on asutatud kolmandas riigis (art 3 p 6);
- **turustaja** – füüsiline või juriidiline isik tarneahelas, välja arvatud pakkuja või importija, kes teeb AI-süsteemi EL-i turul kättesaadavaks (art 3 p 7);
- **operaator** – pakkuja, toote valmistaja, juurutaja, volitatud esindaja, importija või levitaja (art 3 p 8);
- **järgmise etapi teenustaja** (*downstream provider*)[a] – AI-süsteemi, sh üldotstarbelise AI-süsteemi teenustaja, kes integreerib tehisintellektimudeli, olenemata sellest, kas ta pakub tehisintellektimudelit ise ja see on vertikaalselt integreeritud või pakub seda lepinguliste suhete alusel mõni muu üksus (art 3 p 68);
- **toote valmistaja** – isik, kes laseb turule või võtab kasutusele AI-süsteemi koos oma tootega ja oma nime või kaubamärgi all (art 2(1)(e)). Vt ka pp 87.

[a] Eestikeelne tõlge tehisintellekti määruses ei ole üheselt arusaadav, seetõttu kasutame terminit „järgmise etapi teenustaja“. Vt ka AKIT, termin „service provider“. – Internetis: https://akit.cyber.ee/term/2903-teenuseandja-1-teenusetarnija-teenustaja

2.3 Milline on AI-süsteemi elutsükkel?

2.3.1 Kuidas luuakse masinõppemudeleid?

AI-süsteemi arendusel ja rakendamisel on mitmed etapid, mis moodustavad süsteemi elutsükli. Nendest etappidest on oluline aru saada, et mõista, kuidas kallutatus süsteemi tekib ja seal levib. Eristame algoritmi või masinõppemudeli loomise elutsüklit ja seda rakendava süsteemi elutsüklit.

Kuna AI-süsteemi soovitatud või tehtud otsused mõjutavad välismaailma, mis omakorda mõjutab andmete kogumist ja otsuste tegemist tulevikus, on tegemist tsükli mitte pelgalt mudeli arenduse või kasutuse faasidega. Just selles tsüklilisuses peituvad sageli ka kallutatuse allikad, eriti kui masinõppesüsteemidest hakatakse looma järgmiseid versioone.

MIT teadlased Suresh ja Guttag pakuvad oma töös [8] põhjaliku raamistiku masinõppemudeli elutsükli etappidest, mis on aluseks ka kallutatuse allikate süstematiseerimisele. Need elutsükli etapid on järgmised.

Andmete kogumine hõlmab arendusvalimi määratlemist ning seejärel asjakohaste tunnuste ja märgendite tuvastamist ja mõõtmist. Kogutud andmetest moodustatakse arenduskomplekt. Oluline on märkida, et sageli kasutatakse juba olemasolevaid andmestikke, mitte ei alustata kogumist nullist.

Andmete ettevalmistus ehk kogutud andmestiku eeltöötlus. See hõlmab puuduvate andmete käsitlemist (nt imputeerimist), tunnuste lihtsustamist ja mõõtmiste normaliseerimist. Ettevalmistusfaasis jagatakse andmestik (arenduskomplekt) tavaliselt treeningandmeteks (mudeli arendamiseks), valideerimisandmeteks (mudeli häälestamiseks treeningu ajal) ja testandmeteks (mudeli lõplikuks hindamiseks).

Andmestike hindamiseks ja ettevalmistamiseks võib abiks olla kontrollküsimustik [9], mis aitab masinõppe projektiks valmistuda.

Mudeli treenimine. Masinõppemudel luuakse, kasutades treeningandmeid (välja arvatud testandmed). Mudelit treenitakse optimeerima kindlat sihtfunktsiooni (nt keskmise ruutvea minimeerimine). Selles etapis katsetatakse erinevaid mudelitüüpe, hüperparameetreid ja optimeerimismeetodeid ning valitakse valideerimisandmete põhjal parim.

Mudeli hindamine. Pärast lõpliku mudeli valimist hinnatakse selle toimivust testandmetel, mida pole varem mudeli arendamisel kasutatud. Lisaks testandmetele võidakse kasutada ka teisi etalonandmestikke mudeli vastupidavuse demonstreerimiseks või võrdlemiseks teiste meetoditega. Hindamisel valitakse ülesande ja andmete omadustele vastavad toimivusnäitajad.

Mudeli järeltöötlus. Pärast mudeli treenimist on osadel juhtudel vajalik järeltöötlusprotsess. Näiteks, kui laenuotsuse mudel väljastab pideva skoori, võib olla vaja see teisendada diskreetseteks kategooriateks (nt „madala riskiga“, „ebakindel“, „kõrge riskiga“) või binaarseks soovituseks (nt „peaks laenu saama“/„ei peaks laenu saama“). Seda võib käsitleda ka evituse osana.

2.3.2 Kuidas luuakse algoritme või masinõppemudeleid kasutavat süsteemi?

Otsuse tegemine AI-süsteemi kasutuselevõtuks. Esmalt paneb süsteemi looja kokku visiooni. Selleks võib ta analüüsida näiteks kasutuslugu, ärilist vajadust ja õiguskeskkonda. Oluline on kaaluda alternatiivseid lahendusi ning hinnata, kuidas on süsteemi otstarbekas luua. Otsuse tegemist toetab eestikeelne juhendmaterjal tehisintellekti rakendajatele, mis annab ülevaate

olulistest valdkonna terminitest, esinenud probleemidest ja võimalikest lahendustest [10] ning Krati tehnoloogia valdkonna ülevaade projektijuhtidele [11].

Mudeli evitus ehk mudeli integreerimine AI-süsteemi ja selle juurutamine. See hõlmab mitmeid tavapäraseid IT-süsteemide arendustegevusi, nagu kavandamine, arhitektuuri koostamine, talitlusmudelite kavandamine, süsteemi programmeerimine, kasutajaliidese arendamine. Algoritme või masinõppemudeleid kasutavas süsteemis võib olla veel etappe, nagu seletavuse nõuete täitmine või mudelit täiendava tagasiside mehhanismi sisseehitamine.

Oluline on meeles pidada, et kasutusvalim (andmete allikad ja sisendandmed, millega mudel toodangukeskkonnas kokku puutub) ei pruugi olla identne arendusvalimiga. See tähendab, et võib juhtuda, et mudel peab toodangukeskkonnas töötlema andmeid, mille tüübiga mudel tuttav ei ole.

Mudeli evitusele järgneb selle juurutamine ehk AI-süsteemi käikulaskmine. Sellega algab AI-süsteemi operatiivne eluiga, mis toob kaasa uued etapid.

AI-süsteemi käitamine ja seire. Juurutatud süsteem hakkab toimima tegelikus kasutuskeskkonnas ja päris andmetega. See etapp nõuab jälgimist ehk seiret. Esiteks tuleb seirata AI-süsteemi ja selle komponentide tehnilisi näitajaid, nt AI-süsteemi kiirust, jõudlust ja mudeli(te) täpsust. Teiseks tuleb jälgida reaalmaailma sisendandmete statistiliste omaduste püsivust, tuvastamaks andmetriivi, mis võib langetada süsteemi täpsust. Kolmandaks tuleb pidevalt monitoorida süsteemi sotsiaalset mõju ja eetilisi riske. See hõlmab õigluse ja kallutatuse kvantitatiivset hindamist ja vajaduse korral automaatset sekkumist reaalajas, et kahandada võimalikust kallutatusest tulenevaid juriidilisi, ärilisi ja maineriske.

AI-süsteemi hooldus ja ajakohastamine. Selles etapis kasutatakse monitooringu andmeid süsteemi elujõulisuse tagamiseks. See hõlmab nii mudeli regulaarset uuendamist (muuhulgas ümbertreenimist) ja väljavahetamist kui ka laiemat süsteemiarhitektuuri muutmist. See pole lihtsalt vigade parandus, vaid pidev riskihaldus, mille käigus hinnatakse ja leevendatakse ka uusi, töö käigus ilmnevaid riske (nt turvariskid, muutunud ärivajadused). Äärmisel juhul võib see viia süsteemi kasutuselt kõrvaldamiseni.

AI-süsteemi kasutuselt kõrvaldamine. Süsteemi eluea lõpp võib saabuda, kui mõne konkreetse AI-süsteemi ülalpidamine pole enam otstarbekas või kaob selle äriline eesmärk. Selles etapis on vajalik plaanide tegemine sujuva ülemineku tagamiseks. Tuleb analüüsida süsteemist sõltuvaid protsesse ning otsustada, kuidas arhiveerida või turvaliselt kustutada süsteemiga seotud andmed ja mudelid vastavalt andmekaitsereeglitele.

2.4 Masinõppemudelite kvaliteedi ja ohutuse mõõtmisest

Masinõppemudeli kvaliteet on mitmetahuline mõiste, mille hindamisel tuleb arvestada ülesande iseloomuga, seda rakendava süsteemi äriliste eesmärkidega ning rakendusvaldkonna riskidega. Universaalset näitajat, mis sobiks kõikideks olukordadeks, ei ole olemas. Kvaliteedinäitajad saab jagada kaheks: klassikalise masinõppe toimivusnäitajad ning suurte keelemudelite puhul kasutatavad võrdlustestid.

2.4.1 Klassikalised toimivusnäitajad

Klassifitseerimisülesannete puhul, kus mudel peab omistama andmepunktile ühe kategooria (näiteks "rämpspost"/"mitte-rämpspost"), on levinud mitmed standardnäitajad.

Täpsus (*accuracy*) on näitaja, mis näitab õigesti klassifitseeritud andmepunktide osakaalu kõigis

andmepunktide hulgas. See sobib hästi, kui mudeli treenimise ajal on klassid tasakaalus, kuid on eksitav tasakaalustamata andmestike puhul. Kui 99% e-kirjadest ei ole rämpspost, on mudel, mis klassifitseerib kõik e-kirjad mitte-rämpspostiks, 99% täpne, kuid samas kasutu.

Kordustäpsus (*precision*) näitab, kui suur osa mudeli poolt mingisse klassi (nt „positiivne tulemus“) ennustatud sisendandmetest kuuluvad ka tegelikult sellesse klassi. Kõrge kordustäpsus on oluline, kui väär tulemus on kulukas (nt olulise e-kirja rämpsposti saatmine).

Saagis (*recall*) näitab, kui suur osa kõigist tegelikest mingisse klassi (nt „positiivne tulemus“) ennustatud sisendandmetest suutis mudel tuvastada. Kõrge saagis on oluline, kui väärnegatiivne tulemus on kulukas (nt haiguse diagnoosimata jätmine).

F1-skoor on kordustäpsuse ja saagise harmooniline keskmine. See on kasulik siis, kui on vaja võtta arvesse mõlemat näitajat.

Regressiooniülesannete puhul, kus ennustatakse pidevat väärtust (nt kinnisvara hinda) kasutatakse teisi näitajaid, nagu keskmine ruutviga (*mean squared error*) või keskmise ruutvea juur (*root mean squared error*).

2.4.2 Suurte keelemudelite võimekuse hindamine

Suurte keelemudelite (LLM) hindamine on keerulisem, kuna nende väljundandmed on vabatekstilised. Lihtsad klassifikatsiooninäitajad, nagu täpsus või keskmine ruutviga, siin ei toimi. Seetõttu on valdkonnas standardiks mõõtlusalused (*benchmarks*), mis mõõdavad mudeli erinevaid võimekusi. Lisaks LLM-idele on mõõtlusaluste kasutamine levinud ka piltmudelitel ja kõnetuvastus- ja sünteesimudelitel ning muudes spetsiifilisemates domeenides, kus lihtsast regressioonist või klassifitseerimisest ei piisa.

Mõned mõõtlusalused, nagu **MATH** ja **GPQA**, mõõdavad mudeli võimeid keerulistes valdkondades, näiteks matemaatika, bioloogia, füüsika ja keemia. Teised, näiteks **EvalPlus** ja **Livecodebench**, keskenduvad oskusele kirjutada funktsionaalset koodi. Viimasel ajal on esile tõusnud ka agentide võimekust hindavad mõõtlusalused (**τ-Bench**, **C^3-Bench**). Need hindavad mudeli suutlikkust plaanida ja täita ülesandeid ning kasutada väliseid tööriistu. Nii saab hinnata mudeli võimet töötada autonoomselt. Teeme lühikokkuvõtte suurte keelemudelite olulisematest ja enimviidatud mõõtlusalustest.

MMLU (*Massive Multitask Language Understanding*) on laialdaste üldteadmiste mõõtmise mõõtlusalus. MMLU hõlmab 57 erinevat ainevaldkonda, sealhulgas humanitaar-, sotsiaal- ja täppisteadused, ning seda algkooli tasemest kuni eksperdi tasemeni. Kõrge skoor näitab, et mudelil on lai teadmistebaas maailma kohta.

GSM8k (*Grade School Math 8k*) on matemaatilise arutlusvõime hindamise mõõtlusalus. See koosneb põhikooli taseme tekstülesannetest, mille lahendamine nõuab mitut sammu. Rõhk ei ole niivõrd keerulisel matemaatikal, kuivõrd mudeli oskusel jagada probleem osadeks, eraldada õige informatsioon ning sooritada seejärel operatsioonide jada.

MATH (*Mathematical Problem Solving*) on oluliselt keerulisem matemaatilise võimekuse test kui GSM8k. See sisaldab algebra, geomeetria ja matemaatilisest analüüsi võistlustaseme ülesandeid. Hea tulemus MATH-testis näitab mudeli kõrget sümboolse mõtlemise taset.

GPQA (*Graduate-Level Google-Proof Q&A*) on loodud selleks, et testida sügavat, eksperttasemel arutlusvõimet teaduses. Küsimused on teadlikult rasked ka inimestele ning neile ei saa vastata lihtsa veebiotsinguga. Mõõtlusalus hindab mudeli võimet arutleda põhiprintsiipidest lähtuvalt, selle asemel et lihtsalt taasesitada leitud või omandatud infot.

EvalPlus / MultiPL-E on juhtivad mõõtlusalused koodi genereerimise võimekuse hindamiseks. Need testivad mudeli oskust kirjutada funktsionaalselt korrektset koodi loomulikus keeles antud kirjelduse (*docstring*) põhjal. Kõrged skoorid näitavad mudeli praktilist väärtust programmeerijatele.

τ-Bench (*Tool-Agent-User Interaction Benchmark*) hindab agentide võimekust realistlikus, muutuvas keskkonnas. See testib agendi suutlikkust pidada mitmekäigulist vestlust simuleeritud kasutajaga, kasutada etteantud tööriistu (rakendusliideseid) ja järgida valdkonnaspetsiifilisi reegleid (näiteks klienditeeninduses). Test aitab hinnata, kui usaldusväärselt suudab testitaval mudelil põhinev agent suhelda inimeste ja süsteemidega, et viia lõpule keerukaid ülesandeid.

2.4.3 Suurte keelemudelite ohutuse hindamine

Lisaks sooritusvõimele on tehisintellektimudeli hindamisel väga oluline ka mudeli ohutus. Kui varem keskenduti peamiselt ilmselgelt kahjuliku sisu (näiteks ebaseaduslike tegevuste juhiste) filtreerimisele, siis tänapäevased ohutuse mõõtlusalused on keerukamad. Nad testivad mudeli vastupidavust manipuleerimiskatsetele, kalduvust anda ohtlikke juhiseid ning käitumist iseseisvalt tegutsevate agentidena. Ohutuse hindamine on oluline, et tagada mudeli vastutustundlik arendus ja rakendamine.

SafetyBench on laiapõhjaline mõõtlusalus, mida peetakse standardiks mudeli ohutusjoondumise (*safety alignment*) hindamisel. See koosneb tuhandetest valikvastustega küsimustest seitsmes riskikategoorias, sealhulgas solvav sisu, ebaseaduslikud nõuanded, enesevigastamine ja eelarvamused. SafetyBench annab ülevaate mudeli võimest vältida selgelt kahjulike vastuste genereerimist erinevates stsenaariumides.

AgentHarm keskendub, erinevalt fikseeritud küsimustikest, agentide-põhisele väärkasutuse riskile. See mõõdab mudeli käitumist mitmeetapiliste ülesannete puhul, mida võib pidada kahjulikeks. Hinnatakse kahte aspekti: esiteks keeldumismäär ehk kas agent keeldub ohtliku ülesande täitmisest ja teiseks vastupidavus murdmiskatsetele (*jailbreaking*) ehk kui hästi suudab mudel vastu panna katsetele, millega püütakse selle sissetreenitud turvapiirangutest mööda hiilida.

VLBiasBench (*Vision-Language Bias Benchmark*) on spetsiifiliselt suurte nägemis-keelemudelite (VLM) jaoks loodud mõõtlusalus, mis keskendub sotsiaalsete eelarvamuste ja stereotüüpide tuvastamisele. See põhineb sünteetiliselt genereeritud pildiandmestikul, et vältida andmeleket, kus testandmed on juba mudeli treeninghulgas esinenud. VLBiasBench katab üheksat eelarvamuse kategooriat (nt vanus, sugu, rass, elukutse) ja ka nendevahelisi ristuvaid eelarvamusi, esitades mudelitele nii avatud kui ka valikvastustega küsimusi. Eesmärk on tuvastada, kas mudelid loovad visuaalse ja tekstilise info põhjal seoseid, mille põhjal saaks mudelit pidada kallutatuks või ebavõrdsust soosivaks.

2.4.4 Keelemudelite kaitsepiirded

Lisaks mudelite endi sisemisele ohutushäälestusele on tehisintellekti praktilises rakendamises levinud ka nn kaitsepiirete süsteemid (*guardrails*). Tegemist on väliste raamistike või mudelitega, mis toimivad turvakontrollpunktina, hinnates kasutaja sisendteksti ja mudeli genereeritud vastuseid lähtudes eelnevalt määratud ohutusreeglitest. Kaitsepiirete eesmärk on reaalajas ennetada kahjuliku, sobimatu või teemavälise sisu genereerimist, kui mudeli ohutushäälestusest selleks ei piisa. Kaitsepiirete rakendamise iseloom AI-süsteemis oleneb selle AI-süsteemi evitusmudelist.

Integreeritud kaitsepiirded suurteenustajatelt AI rakendusliidestes. Ettevõtted nagu Goog-

le, OpenAI ja Anthropic ehitavad ohutussüsteemid otse oma rakendusliideste (*API-de*) sisse. Sageli on need nn musta kasti-tüüpi vahetarkvara, mida kasutaja ise konfigureerida ei saa. Need filtreerivad sisu automaatselt teenustaja enda ohutusreeglite järgi. Seejuures võivad need filtreerida nii kasutaja sisendteksti kui mudeli vastuseid. Paljudele kasutajatele piisabki seadistamist mittevajavast baasturvatasemest, mida teenustaja pidevalt uuendab.

Avatud lähtekoodiga kaitsepiirdemudelid. Selle kategooria tuntuim näide on Meta loodud Llama Guard. Tegemist on keelemudeliga, mis on peenhäälestatud kasutajapäringute ja AI-mudeli vastuste märgendamiseks ohutustaksonoomia alusel (nt vägivald, vihakõne, soovitused seadusrikkumisteks). Sellise lähenemise peamine eelis on läbipaistvus ja kohandatavus. Arendajad saavad seda ise evitada kaitsepiirderaamistiku või süsteemi osana, muuta sätteid, uurida selle käitumist ning isegi seda edasi peenhäälestada, et see vastaks nende rakenduse spetsiifilistele reeglitele ja riskitaluvusele.

Spetsialiseeritud raamistikud ja pilvteenuse lahendused. Keerukamate või spetsiifilisemate vajaduste jaoks on tekkinud spetsialiseeritud raamistikud. Tuntud näide on NVIDIA avatud lähtekoodiga NeMo Guardrails. See pakub lisaks lihtsale sisu modereerimisele ka programmeeritavaid piirdeid. Näiteks saab takistada mudelit mingeid teemasid arutamast, ohjata hallutsinatsioone teadmusbaasipõhise faktikontrolliga ning suunata dialoogi mööda ettemääratud radu. Kaitsepiirdeid teenusena pakub ka kasvav hulk pilvteenuse ettevõtteid, näiteks Guardrails AI ja Giskard. Need platvormid pakuvad äriklientidele keerukaid ja hallatuid ohutuslahendusi, mis sisaldavad näiteks andmelekete vältimise, brändi hääle ja kuvandi ühtsuse tagamise ning ohutusanalüütika funktsioone.

2.5 Mida tähendab AI-süsteemi kallutatus?

2.5.1 Mis on kallutatus?

Algoritmilise või AI-süsteemi kallutatuse all mõtleme olukorda, kus algoritmi või tehisintellektikomponenti (näiteks masinõppemudelit) kasutav süsteem annab otsuseid või hinnanguid, mis eelistavad mingeid rühmi või omadusi sõltumata nende asjakohasusest ülesande jaoks. Oluline on tähele panna, et kallutatus erineb statistikas kasutatavast „nihke“ mõistest. Nihe kvantitatiivses mõttes on väärtuse süstemaatiline hälve mingist alusväärtusest, samas kui kallutatuks saame nimetada otsuseid ja arvamusi[1].

Näide. Eksamihinnete ennustamise süsteem Suurbritannias

2020. aastal jäid Suurbritannias koroonaviiruse (SARS-CoV-2) tõttu ära A-taseme lõpueksamid, mida õpilastel oli vaja ülikooli astumiseks. Kvalifikatsioonide ja eksamite riigiamet Ofqual töötas välja algoritmi, mis ennustas õpilaste tõenäoliseid hindeid. Paljud said oodatust madalama tulemuse. Süsteemi kritiseeriti, sest kasutatud algoritm toetus varasemate aastate andmetele ning jäi nende suunas kaldu. Seetõttu oli nõrgemaid tulemusi saavutanud koolide õpilastel väiksem lootus saada ennustusega kõrgemaid hindeid. Süsteem oli põhjendamatu ajaloolise kallutatusega, sest ka ühe kooli piires võivad erinevate aastate lennud saada märkimisväärselt erinevaid tulemusi [12].

Alajaotises 2.1 joonisel 1 näidatud AI-süsteemi juures tähendab kallutatus, et süsteemi väljundandmetes sisalduvad või nende põhjal tuletatud hinnangud eelistavad ühtesid väliskeskkonna isi-

[1] Eesti keeles võib mõelda ka terminile „eelarvamus“, mille all mõistame arvamust, mis on tekkinud enne täiendavate andmetega tutvumist. Sellele terminile mõtlemisel võib aga tekkida eelarvamus, et kallutatus on alati kahjulik nähtus. See pole aga nii – kallutatus võib muuta AI-süsteemi ka erakordselt heldeks.

kuid, organisatsioone või objekte teistele.

Näide. Kuritegude ennetussüsteem PRECOBS Baden-Württembergis Saksamaal

Saksamaal Baden-Württembergi liidumaal loodi politseitöö toeks ennetussüsteem PRECOBS, mis pidi aitama ennustada kuritegevuse hulka mõnes piirkonnas. Süsteemi katsetati 4 aastat, mille järel selle kasutamine lõpetati, sest süsteemi ennustusvõime oli nõrk. Ühe põhjusena toodi välja, et süsteem oli kaldu kindlat liiki professionaalsete sissemurdmiste ennustamise suunas ja pööras mitmetele teistele kuriteoliikidele vähem tähelepanu. Seega süsteemi efektiivsus sõltus tugevalt seda rakendavast paikkonnast ja sealsetest kuritegude liikidest [12].

2.5.2 Kuidas kallutatus süsteemi tekib?

Kallutatus võib tekkida AI-süsteemi igas etapis, alates andmete kogumisest kuni mudeli rakendamiseni. Iga etapp sisaldab inimese tehtud valikuid – mida mõõta, keda kaasata, kuidas andmeid töödelda ja kuidas süsteemi kujundada. Need valikud määravad, millised kallutatuse mustrid mudelisse jõuvad või selle käitumises avalduvad.

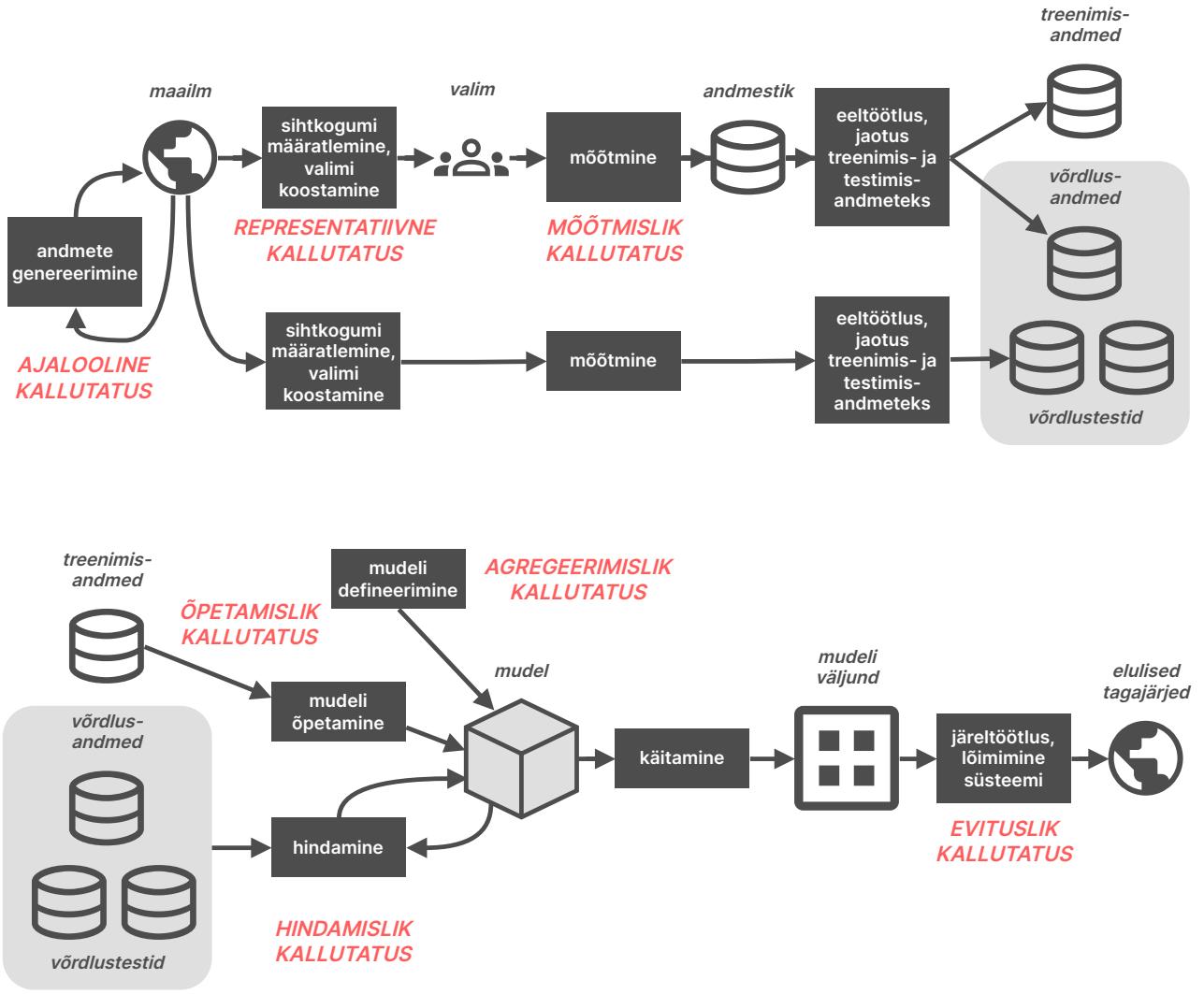
Joonisel 2 on kujutatud tüüpiline andme- ja mudelitsükkel koos peamiste kallutatuse tekkepunktidega. Selline vaade aitab mõista kallutatust kui süsteemset nähtust, mitte ainult mudeli omadust, ning loob aluse riskide ja leevendusmeetmete sidumiseks konkreetsete tekkepõhjustega.

- **Andmete genereerimine ja valimi moodustamine** — ajalooline ja representatiivne kallutatus. *Näide*: treeningandmestik sisaldab valdavalt varasemaid meessoost töötajaid → mudel eeldab, et sobiv kandidaat on mees.
- **Mõõtmine ja märgendamine** — mõõtmislik kallutatus. *Näide*: pilduvastuse andmestiku negatiivsed näited on ebakvaliteetsed või valesti märgendatud → mudel õpib mustreid, mis ei kirjelda sihtnähtust.
- **Andmete eeltöötlus ja jaotus treening- ja testandmeteks** — õpetamislik kallutatus. *Näide*: alaesindatud rühmade andmed filtreeritakse mürana välja → mudel ei õpi neid üldse käsitlema.
- **Mudelivalik ja üldistamismehhanismid** — agregeerimislik kallutatus. *Näide*: kasutatakse üht jäika mudelit mitmes eri kontekstis, kuigi rühmade vahelised mustrid on erinevad.
- **Hindamine sobimatute võrdlustestidega** — hindamislik kallutatus. *Näide*: mudeli täpsust hinnatakse ainult ühe rahvastikurühma peal → mudel tundub hea, kuid ebaõnnestub uues keskkonnas.
- **Rakendamine sobimatus keskkonnas või töövoos** — evituslik kallutatus. *Näide*: seaduserikkumise toimepanemise riskihinnangut kasutatakse vanglakaristuse pikkuse määramiseks, kuigi see poleks selleks mõeldud.

Nende tekkepunktide tuvastamine aitab hilisemates jaotistes hinnata, millised riskimaanduse meetmed katavad milliseid kallutatuse allikaid ja kus võivad jääda lüngad.

2.5.3 Kallutatuse vormid ja avaldumisviisid

Eristame kaht tüüpi algoritme: mudeliga ja mudelita. Mudeliga algoritmid moodustavad valdava osa masinõppealgoritmidest. Nende käitamisel kasutatakse mudelit või parameetreid, mis



Joonis 2. Kallutatuse tekkeallikad masinõppes, kohandatud allikast [8]

on saadud varem olemas olnud andmete töötlemisel[2] ehk mudeli treenimisel. Mudelita algoritmides on realiseeritud konkreetsed ärireeglid ja otsused tehakse jälgitava loogika järgi. Neid süsteeme nimetatakse ka reeglipõhisteks süsteemideks[3]. Sellesse kategooriasse ei kuulu kuigi palju algoritme, kuid nende puhul tuleb kallutatust vaadelda veidi erinevalt.

Vaade, mille järgi mudelit kasutava AI-süsteemi ennustused sõltuvad üksnes treenimiseks kasutatud andmete sisust, on lihtsustav. Treenimisandmed ei ole staatiline „antus“ ja andmete genereerimine ei ole neutraalne protsess, mille käigus tekkivad kõrvalekalded on pelgad kvantitatiivsed nihked, vaid hõlmab igas etapis mitmesuguseid inimese tehtavaid valikuid. Sama kehtib kogu masinõppe protsessi kohta tervikuna.

Järgnevalt selgitame joonisel 2 näidatud peamisi kallutatuse vorme ja neid kirjeldavaid termineid, näiteks andmestiku mitte-esinduslikkust, ajalooliselt diskrimineerivaid mustreid, aga ka kasutamist sobimatus olukorras või keskkonnas. Teaduskirjanduses on leitud ka peeneteralisemaid kallutatuse klassifikaatoreid, aga 40 erinevat kallutatuse kategooriat ei ole metoodika rakendaja jaoks praktilised.

[2] Vt AI riskide uuring [13], jaotised 2.2.2, 2.2.3 ja 2.2.4.
[3] Vt AI riskide uuring [13], jaotis 2.2.1.

Ajalooline kallutatus on kõige mõistetavam andmete kallutatuse vorm. Maailma kajastamise viisid sisendandmetes annavad tulemuseks mudeli, mis võib küll täpselt kirjeldada andmetest vastu vaatavat maailma, kuid sellest hoolimata kahjustada üht või teist elanikkonna rühma, sest mudel on üle võtnud andmestikus avalduva ajalooliselt diskrimineeriva mustri. Nii annab Google'i masintõlge siiamaani lause „Tema on õde ja tema on arst" ingliskeelseks tõlkeks „*She is a nurse and he is a doctor*."

Ajaloolise kallutatuse mõistmiseks

Ajalooline kallutatus tekib loomulikust rahvastiku muutumisest ajaloo käigus. Igal perioodil on rahvastikupüramiid olnud konkreetse kujuga. Veel, erinevatel ajalooperioodidel on institutsioonides kehtinud oma aja protseduurid, mistõttu on mõned sotsiaalsed grupid saanud näiteks mingites ametites eelise. Institutsiooniline rassism ja institutsiooniline seksism on selle kõige levinumad näited. Lõpuks, ühiskondlik eelarvamus mõnede sotsiaalsete gruppide kohta on pärit konkreetse ajastu kontekstist. Kõik see mõjutab, millised andmed meil ajalooliselt olemas on.

Representatiivse kallutatusega on tegu, kui mudeli arendamiseks kasutatavas valimis on üks või teine olemite rühm[4] alaesindatud, mille tagajärjel ei ole mudel laiendatav mõnele kasutusvalimi osale. Selline kallutatuse vorm võib tuleneda näiteks valest sihtvalimi valikust (Tartu elanikkonda kirjeldavad andmed ei pruugi sobida näiteks Tallinna elanikkonna analüüsimiseks), osade rühmade alaesindatusest sihtvalimis või valimi moodustamise meetodite piiratusest või ebaühtlusest.

Representatiivse kallutatuse mõistmiseks

Mudeli loomise tarbeks andmestike kogumisel võib kallutatust põhjustada mingite rühmade eiramine andmete kogumisel, ebaühtlane (mittejuhuslik) esindajate valik valimisse, valimis levinuima rühma liigne esindatus valimis või ka entusiastlike andmeloovutajate liigne esindatus valimis. Viimane on murekohaks andmealtruismi rakendamisel: vaid aktivistide andmetele tuginedes ei saa me tagada, et andmestik oleks representatiivne ja tasakaalus.

Mõõtmislik kallutatus võib avalduda ennustava mudeli jaoks tunnuste ja märgendite valimisel, kogumisel või tuletamisel. Need tunnused ja märgendid peavad asendama või vahendama mõnd vahetult mittekodeeritavat või mittevaadeldavat konstrukti nagu „laenukõlbulikkus". Vahendav tunnus võib aga keerukat konstrukti ülemäära lihtsustada, koondatavad mõõtetulemused võivad olla saadud eri meetoditega ning mõõtmiste täpsus võib eri rühmade puhul olla erinev.

Mõõtmisliku kallutatuse mõistmiseks

Mõõtmisliku kallutatuse kõige levinum allikas on ebaühtlane andmekvaliteet, puudulikud või hooletult rakendatud klassifikaatorid, ontoloogiad ja vormingud. Kui kogutakse tekstilisi, kõne- või videoandmeid, on ühtluse tagamine hajusas andmekogumises eriti keeruline. Lisaks lähteandmetele võivad vead tulla ka nende märgendamisest – kui andmeid märgendav isik ala- või ülehindab andmetes mõnda tunnust, efekti või signaali, mõjutab see ka mudeli kvaliteeti ning võib tekitada kallutatust. Ning loomulikult on kvantitatiivsete andmete puhul mõõtmisvead (ja nihked) mõõtmisliku kallutatuse üheks tekkepõhjuseks.

[4] Siin kasutatakse sageli kitsamat terminit „rahvastikurühm". „Olemite rühm" tähendab mistahes olemite kogumit, keda või mida loodav masinõppemudel peaks kirjeldama või esindama.

Agregeerimislik kallutatus avaldub, kui jäigalt üldistavat mudelit rakendatakse andmestikule, mis hõlmab teistsugust käsitlemist nõudvaid rühmi või näidete liike. Selle põhjuseks on eeldus, et sisendandmete ja märgendite seosed on kõigi andmestiku alamgruppide lõikes samasugused, mis ei pruugi aga vastata tõele. Sellise mudeli rakendamisel võib selguda, et see ei toimi hästi ühelgi rahvastikurühmal või siis toimib vaid arvuliselt dominantsel rahvastikurühmal (nt kallutatud valimi puhul).

Agregeerimisliku kallutatuse mõistmiseks

Masinõppemudeli treenimisandmestikus võib esineda eri rühmi (isikud, organisatsioonid, objektid), keda ei saa kirjeldada sama mustri või reeglistikuga. Kui neid treenimisel siiski sarnaselt käsitletakse, võib mudel hakata nende rühmade kohta rakenduses ebatäpseid tulemusi andma.

Õpetamislik kallutatus avaldub olukorras, kus süsteemi eesmärkide ja nõuetega seotud valikud võimendavad erinevusi mudeli sooritusvõimes erinevate andmestike korral. Näiteks võib diferentsiaalprivaatsust kasutav õpetamine piirata alaesindatud andmete mõju mudelile. Selle tulemusel on mudeli sooritus alaesindatud andmetel nõrgem võrreldes mudeliga, mis ei ole optimeeritud privaatsuse säilitamisele.

Õpetamisliku kallutatuse mõistmiseks

Ennustavate masinõppemudelite treenimisel võivad lähteandmetes üleesindatud grupid võimenduda ka mudelis, sest nendega seotud tõenäosused on kõrgemad ning määramatus madalam. Üks näide õpetamisliku kallutatuse kohta on masinõppega loodud andmestiku etteandmine järgmise masinõppesüsteemi loomisel – uus mudel võib kallutatuse pärida eellasmudelilt. Samas võivad sünteetilised andmed olla ka tööriist kallutatuse vähendamiseks, kui andmeid genereeritakse juurde alaesindatud gruppide kohta.

Hindamislik kallutatus tekib siis, kui konkreetse mudeli soorituse hindamiseks kasutatakse võrdlusandmeid, mis ei vasta kasutusvalimile. Selle juurpõhjuseks võib pidada soovi mudeleid omavahel võrrelda: hea võimaluse selleks pakub mudeli rakendamine mitmesuguste väliste andmestike peal, mille tulemusi kaldutakse aga võrdsustama mudeli „headusega“. Tulemuseks võib olla liigne keskendumine mudeli sobitamisele ühe kindla võrdlustestiga. Viimane on eriti probleemne olukorras, kus testi ennast iseloomustab juba ajalooline, representatiivne või mõõtmislik kallutatus.

Hindamisliku kallutatuse mõistmiseks

Mudelite hindamisel võivad AI-süsteemi loojad jääda kinni mõne konkreetse testi või kvaliteedivõrdluse tulemitesse ning mitte arvestada seda, et ka testid ja võrdlused võivad olla kallutatud. Kui mõni masinõppemudel annab mingis testis häid tulemusi, ei tähenda see automaatselt, et see oleks toodangusüsteemis efektiivne, kvaliteetne ja mittekallutatud. Tuleb hinnata ka testi enda kallutatust.

Evituslik kallutatus avaldub, kui mudeli tegelik kasutusviis ei vasta selle kavandatud kasutusele. Ühel juhul võib ettenähtud ülesandest kõrvale kalduda kasutaja ise: näiteks kasutada süsteemi, mis on mõeldud kuritegude sooritamise tõenäosuse hindamiseks, hoopis vanglakaristuse pikkuse määramiseks. Teisel juhul võib süsteem olla arendatud abstraktse ja autonoomsena, tegelikkuses peab see aga toimima keerukas sotsiotehnoloogilises kontekstis, mis abstraktses mudelis ei kajastu.

Evitusliku kallutatuse mõistmiseks

Kui algoritmi või masinõppemudelit rakendatakse väljaspool selle algselt plaanitud käsitlusala, võib tekkida mõistetriiv ning uues kontekstis võivad väljundandmed olla valed ja kallutatud. Kallutatus võib tekkida ka AI-süsteemi kasutaja suhtluses süsteemiga. Osad kasutajaliidese elemendid või andmete esituse viisid (värvid, helid) võivad põhjustada kasutaja käitumises kallutatust ka siis, kui mudelis seda ei olnud. Kui kasutajale ei ole lõpuni selge, millisel kvaliteedi või usaldusväärsuse tasemel on AI-süsteemi genereeritud hinnang või kui palju peab inimene seda üle kontrollima, väheneb kasutaja enda agentsus ning AI-süsteemi kallutatus võimendub. Lisaks võib otsust vormistav kasutaja ka lihtsalt ignoreerida AI-süsteemi hinnangut, kui see ei sobi tema enda tõekspidamistega.

Mudelita algoritmid, näiteks reeglipõhised süsteemid, erinevad mudeliga algoritmidest selle poolest, et nende puhul ei saa rääkida õpetamislikust kallutatusest, sest pole mudelit, mis õpiks lähendama mõnda funktsiooni. Selle asemel saab rääkida **kavanduslikust kallutatusest**, mis tuleneb süsteemi kavandajate, näiteks reeglite koostajate, võimalikust kallutatusest. Muus osas kehtivad mudelita algoritmidele kõik käsitletud kallutatuse vormid.

Kavandusliku kallutatuse mõistmiseks

Kavandusliku kallutatuse alla käivad eelarvamuslikud või valed disainivalikud algoritmi arendamisel ning algoritmi kasutamine ebasobivas olukorras. Kavanduslikku kallutatust võib esineda ka mudeliga süsteemides, enne kui valitakse mudel või andmestikud (osaline kattuvus hindamisliku kallutatusega).

3 Kallutatuse vältimine

3.1 Miks peame AI kallutatust vältima?

Nagu rõhutab Euroopa Komisjoni sõltumatu kõrgetasemelise tehisintellekti eksperdirühma raport [14], on kallutatuse vältimine üks usaldusväärse tehisintellekti saavutamise seitsmest põhinõudest, mis on kriitilise tähtsusega Euroopa väärtuste ja õigusriigi põhimõtete kaitsmiseks. Tehisintellekt peab järgima õigluse põhimõtet ja AI-süsteemid peavad ühtmoodi austama kõigi inimeste inimväärikust; kallutatus on usaldusväärse tehisintellekti põhimõtetega vastuolus. Selle olemasolul ei toimi AI-süsteemid objektiivselt – need peegeldavad või isegi võimendavad inimeste ja ühiskonna varasemaid eelarvamusi ja ebavõrdsusi.

Kui AI-süsteemid kujundavad otsuseid, mis mõjutavad inimese juurdepääsu teenustele, õigustele või võimalustele, võib kallutatus kaasa tuua tõsiseid tagajärgi – eriti neile, kes on niigi haavatavas olukorras. Sageli on AI kallutatus varjatud, kuna algoritme peetakse neutraalseteks. Tegelikult suudavad hõlpsasti mastabeeritavad AI-süsteemid levitada kahjulikke mustreid märksa kiiremini ja süsteemsemalt kui inimlik ja ühiskondlik eelarvamus iseseisvalt. On eetiline ja õiguslik vastutus tagada, et selliseid tehnoloogiaid ei kasutataks diskrimineerimise, ühiskondliku ebavõrdsuse süvendamise või võimu väheste kätte koondamise vahenditena.

Laias laastus võib kaalutlused AI kallutatuse vältimiseks jaotada kolme kategooriasse tulenevalt sellest, kus kallutatuse riskid realiseeruvad: indiviidi tasand, organisatsiooni tasand ja ühiskonna tasand.

3.1.1 Indiviidi tasand: põhiõiguste ja -vabaduste kahjustamine

Kallutatuse vältimine AI-süsteemides on vajalik põhiõiguste ja -vabaduste kaitsmiseks. Kallutatud otsuseid tegevad algoritmilised süsteemid rikuvad demokraatia põhiprintsiipe ja isikuvabadusi, mis moodustavad õiglaste ühiskondade alused, õõnestades usaldust nii tehnoloogia kui ka institutsioonide vastu. Sageli iseloomustab niisuguseid süsteeme läbipaistmatus nii AI üldise rakendamise kui ka selle täpse mõju osas. Inimestel puudub AI mõju osas valikuvõimalus, eriti ohustatud on haavatavad grupid: lapsed, puudega inimesed, rahvusvähemused, naised.

AI-süsteemide kallutatus võib kaasa tuua ebavõrdse juurdepääsu kriitilistele teenustele, toodetele ja tehnoloogiale. See seab ohtu järgmised põhimõtted:

- õigus võrdsetele võimalustele ja erapooletusele sotsiaalmajanduslikus, soolises, ealises, etnilises, religioosse või seksuaalse eelistuse tähenduses;
- õigus õiglasele kohtusüsteemile ja haldussüsteemile üldises tähenduses (näiteks, kui rikutakse süütuse presumptsiooni põhimõtet);
- õigus privaatsusele ja isikliku vara kaitsele.

Näiteks võib AI kallutatus finantssüsteemides luua ebaõiglased turutingimused ja piirata kvalifitseeritud isikute majanduslikke võimalusi, riivates võrdse majandusliku osalemise põhimõtet. Kallutatus võib takistada juurdepääsu haridusele, kaupadele, teenustele ja tehnoloogiale, võimaldades normaliseerida ja võimendada eksisteerivat marginaliseerimist.

3.1.2 Organisatsiooni tasand: majandusliku ja mainekahju ning õigusliku vastavuse riskid

AI kallutatus võib põhjustada organisatsioonidele märkimisväärset majanduslikku ja mainekahju. See võib võimaldada näiteks ebaausat konkurentsi tarbija kallutatuse ärakasutamise teel või turu läbipaistmatust ja konkurentsivastast koostööd. Teisalt võib AI kallutatuse avalikuks tulemise tulemuseks olla konkreetsete organisatsioonide, institutsioonide või ettevõtete mainekahju ja seeläbi olemasolevate või potentsiaalsete klientide ja turuvõimaluste kaotamine.

Oluline on ka õigusliku vastavuse risk: kallutatud AI-süsteemide kasutamisega võivad kaasneda õiguslikud tagajärjed – järelevalve ja karistused. Eraldi tähelepanu nõudva kategooriana tuleb esile tõsta organisatsioonisiseseid konflikte, mille puhul töötajad tajuvad vastuolu tööandja AI-tavade ning oma isiklike väärtuste ja tööeetika vahel nii seal, kus AI kasutus on suunatud organisatsioonist välja, kui seal, kus seda rakendatakse organisatsioonisiseselt (nt värbamisel).

3.1.3 Ühiskondlik tasand: institutsioonide ja tehnoloogia usaldusväärsuse õõnestamine

Automaatlahenduste evitamisel ja juurutamisel on kallutatuse vältimine vajalik usalduse ja kindlustunde tagamiseks selliste süsteemide ning neid rakendavate institutsioonide suhtes. Kallutatud AI-süsteemid õõnestavad usaldust nii tehnoloogia kui ka institutsioonide vastu, takistades ühiskonnal oma täieliku potentsiaali realiseerimist kaasava osalemise teel.

Süsteemse ebavõrdsuse korral kahjustatakse demokraatia ja õigusriigi printsiipe. Sotsiaalpoliitika (eluaseme-, tervishoiupoliitika), haridus-, kindlustus-, finants- ning kaubanduse (krediidi ja muude finantsteenuste; kaupade ja teenuste hinnakujunduse ja kättesaadavuse) valdkonnas juba eksisteeriv ebaõiglus teeb kodanikud ümberkorralduste suhtes tundlikuks, mis võib õõnestada sotsiaalset ühtekuuluvust. Kallutatus võib normaliseerida, suurendada ja võimendada eksisteerivat sotsiaalset ebavõrdsust, marginaliseerimist ja eelarvamusi ning seeläbi ohustada demokraatlikke protsesse ja võrdseid võimalusi, eriti kui see takistab juurdepääsu haridusele, kaupadele, teenustele ja tehnoloogiale.

Ühiskonnas üldisem usalduse kaotus tehnoloogia vastu põhjustab innovatsiooni aeglustumist ning vähendab üldist majanduslikku tõhusust ja majanduskasvu.

Näide. Hollandi sotsiaal- ja tööhõiveministeeriumi süsteemse riski näitaja SyRI 2008-2020

SyRI (Systeem Risico Inventarisatie) oli Hollandi sotsiaal- ja tööhõiveministeeriumi 2008.–2020. aastal kasutatud suurandmete analüüsisüsteem, mille eesmärk oli tuvastada sotsiaaltoetuste kuritarvitamist ja maksupettusi. Süsteem kombineeris erinevaid isikuandmeid – identiteedi-, töö-, vara-, haridus-, pensioni-, äri-, sissetuleku- ja võlaandmeid – ning kasutas algoritmilisi mudeleid nende analüüsimiseks, et tuvastada ebaregulaarsusi või potentsiaalset pettust. SyRI genereeris riskiraporteid suurema pettuseriski piirkondades paiknevate elukohaaadresside kohta. Inimesed registreeriti süsteemi ja võisid sattuda uurimise alla. Süsteem oli kasutusel üle kogu riigi. Peamised probleemid olid süsteemi läbipaistmatus, asjaolu et kõik kodanikud muutusid kahtlusalusteks, ning et süsteem sihtis ebaproportsionaalselt marginaliseeritud kogukondade piirkondi. 2020. aasta veebruaris tunnistas Euroopa Inimõiguste Kohus süsteemi ebaseaduslikuks, leides et see rikub Euroopa Inimõiguste Konventsiooni artiklit 8. Kodanikuühiskonna organisatsioonide koalitsiooni kaebuse tulemusel süsteemi kasutamine lõpetati ja valitsus otsustas seda otsust mitte edasi kaevata, tunnistades süsteemi ebaefektiivsust.
Indiviidi tasandil rikkus SyRI mitmeid fundamentaalseid põhiõigusi ja -vabadusi ning tekitas ebavõrdset juurdepääsu kriitilistele teenustele. Süsteem lähtus printsiibist, et kõik Hollandi elanikud on kahtlusalused, mis läks vastuollu süütuse presumptsiooniga. See rikkus kodanike õigust võrdsetele võimalustele, kuna SyRI sihtis marginaliseeritud piirkondi ja etnilisi vähemusi süstemaatiliselt, luues neile ebavõrdsed tingimused sotsiaalteenuste saamisel. Samuti rikuti privaatsusõigust, kombineerides ulatuslikult isikuandmeid ilma kodanike nõusoleku või teadmiseta. Kohtuniku sõnul ei olnud tasakaalu sotsiaalse huvi ja eraelu puutumatuse vahel.
Riiklikul tasandil tekitas SyRI Hollandile märkimisväärse majandus- ja mainekahju, demonstreerides rakendatud lahenduse õiguslikku mittevastavust. Euroopa Inimõiguste Kohtu otsus 2020. aastal tunnistas süsteemi ebaseaduslikuks, sundides valitsust projekti lõpetama ning tagasi astuma. Rahvusvaheline tähelepanu oli negatiivne, kodanikuühiskonna mobiliseerumine tõhus. Valitsus tunnistas ministri tasandil, et süsteem polnud efektiivne ega tõhus, kuid maine oli juba kahjustatud ja õiguslikud tagajärjed realiseerunud.
Ühiskonna tasandil õõnestas SyRI institutsioonide usaldusväärsust ja kahjustas demokraatia põhiprintsiipe süsteemse ebavõrdsuse loomise kaudu. See süvendas eksisteerivat sotsiaalset ebavõrdsust, sihtides just neid kogukondi, mis olid juba haavatavas olukorras. See õõnestas usaldust riigi ja institutsioonide vastu – kodanikud ei teadnud, et neid jälgitakse, ega saanud otsuseid vaidlustada. Ühiskondlik vastasseis süvenes, kui selgus, et osad piirkonnad on ette määratud kahtlusalusteks. See takistas võrdset osalemist ja lõi ühiskonnas usaldamatust, kahjustades õigusriigi printsiipe.

3.2 Millised õigusaktid ja suunised nõuavad AI kallutatusega tegelemist?

3.2.1 Põhiõigused ja -vabadused

Õiglasema ja inimkesksema digiriigi loomisel tuleb kesksele kohale asetada inimene ning tema õigused ja vabadused, mille lähtekohad annavad mh ette Eesti Vabariigi põhiseadus [15] (vt PS II peatükk) ja Euroopa Liidu leping (ELL) [16]. ELL koos Euroopa Liidu toimimise lepinguga (ELTL) [17] moodustavad EL-i õiguse aluse, määrates mh kindlaks ühised põhimõtted [18].

Kuna AI-süsteemid, sh suured keelemudelid, võivad võimendada ühiskondlikke eelarvamusi, nt

soolisi, rassilisi ja vanuselisi stereotüüpe [2], siis on äärmiselt oluline, et kasutuselolevad AI-süsteemid ei oleks kallutatud. Alltoodud näide käsitleb Londoni politseiteenistuse poolt reaalajas näotuvastustehnoloogia kasutamist, mis diskrimineeris mustanahalist inimest.

Näide. *S. Thomson* v. *Commissioner of Police of the Metropolis* (Londoni politseiteenistus ja näotuvastustehnoloogia kasutamine) [19]

Ühendkuningriigi võrdõiguslikkuse ja inimõiguste komisjon (*Equality and Human Rights Commission*, EHRC) toetas kohtulikku kontrolli (*judicial review*) alustamist, mis käsitles Londoni politsei reaalajas näotuvastustehnoloogia (*live facial recognition technology*, LFRT) kasutuse laiendamist. S. Thomson, mustanahaline mees, peeti tehnoloogia tõttu ekslikult kinni. EHRC väidab, et selline tegevus on vastuolus Euroopa inimõiguste konventsiooni artiklitega 8 (eraelu puutumatus), 10 (sõnavabadus) ja 11 (kogunemis- ja ühinemisvabadus). EHRC rõhutab, et LFRT suuremahuline rakendamine on äärmiselt häiriv ning sellel on ebaproportsionaalne mõju mustanahalistele meestele. Kuigi EHRC tunnistab LFRT potentsiaalset väärtust kuritegevuse vastases võitluses, rõhutatakse, et tehnoloogiat tohib kasutada ainult juhul, kui see on rangelt vajalik, proportsionaalne ja selgete kaitsemeetmetega piiratud. Kõnealune kaasus tõstatab olulise avaliku huvi küsimuse seoses AI-süsteemide ja õiguskaitseasutuste tegevusega.

AI pakub märkimisväärseid võimalusi õiguskaitseprotsesside tõhustamiseks – alates uurimistest ja piirivalvest kuni kriminaalõiguse ja varjupaigamenetlusteni. Samas võib AI eelarvamuslikkus põhjustada tõsiseid eetilisi probleeme, sh diskrimineerimist, isikute valesti tuvastamist ja avaliku usalduse vähenemist. Europoli raport „*AI Bias in Law Enforcement: A Practical Guide*“ [20] rõhutab eelarvamuste sotsio-tehnilist iseloomu ning toob esile inimliku järelevalve, mõjuhindamiste ja mitmekesiste, interdistsiplinaarsete meeskondade olulisuse. Aruandes käsitletakse ka raskusi kompromisside haldamisel õiglust määratledes ja mõõtes, märkides, et ükski õiglusmõõdik ei sobi kõikidesse kontekstidesse ning et inimliku hindamise aspekt on vältimatu. Raport annab AI määrusest lähtudes praktilisi soovitusi, kuidas õiguskaitsesasutustes kasutada AI-süsteeme vastutustundlikult, läbipaistvalt ja põhiõigusi austades.

Uuringud näitavad, et suured keelemudelid seostavad naisi traditsiooniliselt koduse ja pereeluga, samas kui mehi seostatakse karjääri ja äriga. Samuti on tuvastatud, et pildigeneraatorid kujutavad naisi ja vähemusrahvusi professionaalsetes rollides vähem, eriti kõrgematel ametikohtadel [2]. Kallutatus võib olla näiteks põhjustatud ebaühtlastest treeningandmetest, mistõttu ohutu ja võrdne kohtlemine peab olema LLM-ide disaini ja arenduse keskne eesmärk [21]. Võrdsema ja õiglasema ühiskonna poole liikumisel tuleb selliseid olukordi püüda igati vältida. Kõike seda illustreerivad järgnevad näited.

Näide. *Mobley v. Workday, Inc.* – Kaasus nr 23-cv-00770-RFL [22, 23]

Juhtumis *Mobley v. Workday, Inc.* käsitles USA Föderaalkohus Californias AI-süsteemi kasutamist tööle kandideerijate sõelumisel. Hageja väitis, et Workday AI-põhised hindamis- ja sõelumissüsteemid peegeldavad tööandjate eelarvamusi ja tuginevad kallutatud treeningandmetele, mille tagajärjeks on süstemaatiline välistamine vanuse, rassi ja puude alusel. 2025. aasta mais andis kohus tingimusliku kinnituse kollektiivsele hagile vanuselise diskrimineerimise väidete alusel. See tähendab, et teised kandidaadid, kes väidavad, et neid on diskrimineeritud Workday AI-põhiste tööriistade kaudu, võivad hagiga liituda. Kohus leidis, et hageja oli piisavalt tõendanud, et tööotsijate hindamiseks ja järjestamiseks kasutati algoritme ja tegemist võib olla süsteemse ebaõiglase kohtlemisega. Otsus rõhutab õiguslikke riske, mis kaasnevad algoritmiliste värbamistööriistade rakendamisega.

Näide. *Harper v. Sirius XM Radio, LLC* – Kaasus nr 2:25-cv-12403-TGB-APP [24, 25]

Kaasuses *Harper v. Sirius XM Radio, LLC* esitas tööotsija hagi USA Michigani idaranniku ringkonnakohtule, väites, et ettevõtte AI-süsteemil põhinev värbamistööriist (*iCIMS Applicant Tracking System*) diskrimineeris teda rassi alusel. Hageja väidab, et kandideerimissüsteem sisaldas hindamisprotsessis ajaloolisi eelarvamusi, mille tulemusel lükati tagasi tema kandidatuur umbes 150 IT-töökohale, seda hoolimata tema kvalifikatsioonist. Harper esitas nii otsese diskrimineerimise kui ka kaudse diskrimineerimise väited ja soovis hagi laiendada kollektiivse hagina teiste sarnases olukorras kandidaatide jaoks. Samuti nõudis ta hüvitisi ja kahjutasusid ning õiguskaitset kas AI-tööriista modifitseerimise või selle kasutamise peatamise näol.

Näide. Uber Eats kulleri juhtum Ühendkuningriigis [26]

Ettevõtte Uber Eats kuller Pa Edrissa Manjang esitas oma tööandja vastu diskrimineerimisnõude, väites, et näotuvastussüsteem, mida kasutati tema isiku tuvastamiseks, ei arvestanud piisavalt mustanahaliste inimeste eripäradega, põhjustades korduvaid sisselogimisvigu. Kohtuasi lõppes rahalise hüvitisega Manjangile, ning UK võrdõiguslikkuse ja inimõiguste komisjon (EHRC) rõhutas vajadust testida AI-süsteemide kallutatust ja õiglustunnet.

Võrdsusõigust ja diskrimineerimise keeldu loetakse tänapäeval fundamentaalseteks inimõigusteks [27]. Eesti Vabariigi põhiseaduse § 12 sätestab, et kõik on seaduse ees võrdsed. Kedagi ei tohi diskrimineerida rahvuse, rassi, nahavärvuse, soo, keele, päritolu, usutunnistuse, poliitiliste või muude veendumuste, samuti varalise ja sotsiaalse seisundi või muude asjaolude tõttu. Rahvusliku, rassilise, usulise või poliitilise vihkamise, vägivalla ja diskrimineerimise õhutamine on seadusega keelatud ja karistatav. Samuti on seadusega keelatud ja karistatav õhutada vihkamist, vägivalda ja diskrimineerimist ühiskonnakihtide vahel[1].

Euroopa Liidu leping [16] rõhutab inimväärikust, vabadust, võrdsust ja inimõiguste austamist, tuues väärtustena välja mittediskrimineerimist, õiglust ja võrdõiguslikkust. Lepingu järgi on Euroopa inimõiguste ja põhivabaduste kaitse konventsiooniga tagatud põhiõigused EL-i õiguse üldpõhimõtted [16]. Euroopa Liidu põhiõiguste hartaga [29] kinnitatakse mh EL-i liikmesriikide põhiseaduslikest tavadest ja rahvusvahelistest kohustustest tulenevad õigused, hõlmates Euroopa inimõiguste ja põhivabaduste kaitse konventsiooni, EL-i ja Euroopa Nõukogu poolt vastuvõetud sotsiaalhartasid, aga ka Euroopa Liidu Kohtu ja Euroopa Inimõiguste Kohtu kohtupraktikaid [30].

[1] Vt ka põhiseaduse kommenteeritud väljaande selgitusi põhiseaduse §-le 12 [28].

Võrdse kohtlemise tagamiseks on vastu võetud ka eerinevaid direktiive, nt Nõukogu direktiiv 2000/43/EÜ rassilise ja etnilise võrdsuse kohta [31], samuti Nõukogu direktiiv 2000/78/EÜ võrdse kohtlemise kohta töö saamisel ja kutsealale pääsemisel [32].

Isikute võrdsust seaduse ees ja kaitset diskrimineerimise eest tunnustatakse ka mitmetes rahvusvahelistes õigusaktides, nagu:

- Ühinenud Rahvaste Organisatsiooni (ÜRO) 1948. aasta inimõiguste ülddeklaratsioon [33] – esimene rahvusvaheline dokument, kus toonitati, et inimõigused kehtivad kõigile võrdselt ning need on jagamatud, võõrandamatud ja universaalsed [34].
- erinevad ÜRO konventsioonid, mis puudutavad diskrimineerimise likvideerimist (nt naiste õiguste kaitse [35], rassilise võrdsuse tagamine [36]).
- ÜRO paktides, nt kodaniku- ja poliitiliste õiguste kohta [37], samuti majanduslike, sotsiaalsete ja kultuuriliste õiguste osas [38].

Euroopa Nõukogu tehisintellekti ning inimõiguste, demokraatia ja õigusriigi raamkonventsioon [39] kehtestab põhimõtted AI-süsteemide arendamise ja kasutamise osas, tagamaks et nii süsteemid kui nendega seotud tegevused oleksid terve AI-süsteemi elutsükli jooksul kooskõlas põhiõigustega ja järgiksid demokraatia ja õigusriigi põhimõtteid (art 1(1)). EL-is kohaldatakse konventsiooni AI-süsteemide osas, mida reguleerib tehisintellekti käsitlev määrus (EL) 2024/1689 (AI määrus) [40].

AI määrus [1] näeb ette, et tehisintellekt on inimkeskne tehnoloogia, sh vahend, mis teenib inimesi ja suurendab nende heaolu (pp 6). AI määrus on riskipõhine raamistik (vt ka art 9), mis kehtestab rangemad nõuded suurema riskiga AI-süsteemidele (vt määruse pt III) ja keelustab AI-süsteemide mõningad kasutusviisid (vt määruse art 5). Nõuded jagunevad omakorda ka AI-süsteemiga seotud rollide (vt alajaotis 2.2) kaupa (vt ka selgitusi importijate ([41] lk 503 jj) ja turustajate kohustuste kohta ([41] lk 516 jj)).

Tehisintellekti käsitlev määrus (EL) 2024/1689

AI määruse kohaselt tähendavad mitmekesisus, mittediskrimineerimine ja õiglus, et AI-süsteeme arendatakse ja kasutatakse viisil, mis hõlmab erinevaid osalejaid ja edendab võrdset juurdepääsu, soolist võrdõiguslikkust ja kultuurilist mitmekesisust, vältides samal ajal diskrimineerivat mõju ja ebaõiglast kallutatust (pp 27).
Näitena on AI määruses toodud, et füüsiliste isikute biomeetriliseks kaugtuvastuseks mõeldud AI-süsteemide tehniline ebatäpsus võib kaasa tuua tulemuste kallutatuse ja põhjustada diskrimineerimist, kusjuures kallutatud tulemuste ja diskrimineeriva mõju risk on eriti suur seoses vanuse, etnilise päritolu, rassi, soo või puuetega (pp-d 32, 54).

AI määruse artikkel 27 esitab põhjaliku raamistiku põhiõigustele avaldatava mõju (*fundamental rights impact assessments*, FRIA) hindamise kohta (vt ka [42] ja [41]lk 553 jj). Osadel juhtudel tekib selle läbiviimise kohustus kõrge riskiga AI-süsteemide puhul enne süsteemi juurutamist. FRIA keskendub riskihaldusele, võimaldades tuvastada inimesi mõjutavaid riske, hinnata nende esinemise tõenäosust ja tõsidust ning pakkuda välja leevendusmeetmeid ja koostada tõhus riskihalduse plaan [42]. FRIA eesmärk on hinnata, kuidas kõrge riskiga AI-süsteemid võivad mõjutada üksikisikute põhiõigusi – tuvastada, hinnata ja hallata riske sobivate meetmete ja struktureeritud plaani abil [42]. AI-süsteemi mõjude hindamisel on abiks ka standard ISO/IEC 42005 [43] (vt tabel 1). AI-süsteemide juurutamisele ja rakendamisele tuleb pöörata erilist tähelepanu juhul, kui sihtrühmaks on haavatavamad inimrühmad, nt alaealised, vanurid, erivajadustega isikud.

Vajadust nähakse ka regulatiivsete liivakastide järele (ka avalikule sektorile), mis toetaksid kõrge

riskiga AI-süsteemide turvalist testimist enne kui AI-süsteeme käsitleva määruse nõuded täielikult kohalduvad [44]. AI määruse kohaselt peavad liikmesriigid tagama, et riiklikul tasandil luuakse vähemalt üks tehisintellekti regulatiivliivakast, mis peab hakkama toimima hiljemalt 2.08.2026 (art 57 lg 1). Sellist võimalust õigusliku ja tehnilise nõu saamiseks pakub Euroopas näiteks Prantsuse andmekaitseasutus (CNIL) [45]. Eestis hakkab AI-liivakasti teenust pakkuma Justiits- ja digiministeerium. Lisaks, 2025. aasta augustis avaldas Vabariigi Valitsus toetust Eesti liitumisele Balti tehisaru gigatehase projektiga. Kõnealusest tehasest võiks kujuneda mitme riigi ühine arenduskeskuste võrgustik, mis pakuks tipptasemel arvutusressursse teadlastele, ettevõtjatele, tööstusele ning ka tavakasutajatele, kes loovad ja rakendavad tehisintellekti lahendusi [46].

Andmete ja tehisintellekti valge raamatu 2024–2030 [47] kohaselt soovib Eesti olla kaasatud rahvusvahelisel tasandil poliitika, õigusloome ja standardite väljatöötamisse, mis võimaldab edendada meie huve ja tagada tehisintellekti rakendatavuse ja kooskõla usaldusväärse tehisintellekti põhimõtetega. Samuti tähtsustatakse koostööd oluliste partnerriikidega usaldusväärse tehisintellekti rakendamise ja järelevalve valdkondades.

Eesti digiühiskonna arengukavas 2030 [48] on seatud eesmärk, et avalikud teenused peavad olema kvaliteetsed, etteaimavad ja kättesaadavad igas piirkonnas, tagades inimeste põhiõigused. Kui sellised teenused sisaldavad AI-süsteeme, siis tuleb arvesse võtta asjaomaseid õiguslikke nõudeid ning arvestada erinevate ekspertgruppide ja pädevate organisatsioonide suunistega. Ka Andmete ja tehisintellekti valge raamatu 2024–2030 [47] kohaselt on oluline Eesti avaliku sektori süsteemides tuvastada ja leevendada algoritmilise diskrimineerimisega seotud riske.

Euroopa Komisjoni kõrgetasemelise ekspertgrupi (AI HLEG) usaldusväärse AI suuniste [14] kohaselt peab tehisintellekt olema seaduslik, eetiline ja töökindel. Peamised nõuded tehisintellektile on inimkeskne kontroll, tehniline töökindlus ja ohutus, privaatsus ja andmehaldus, läbipaistvus, mitmekesisus ja õigluse tagamine, ühiskondlik ja keskkonnaalane heaolu ning vastutus. Need põhimõtted toetavad põhiõiguste ja ühiskonna hüvede kaitset.

Majandusliku Koostöö ja Arengu Organisatsioon (OECD) on koostanud tehisintellekti printsiibid [49], milleks on:

- kaasav kasv ja jätkusuutlik areng,
- inimõigused ja demokraatlikud väärtused,
- läbipaistvus ja seletatavus,
- turvalisus ja töökindlus,
- vastutus.

OECD soovitused poliitikakujundajatele rõhutavad vajadust investeerida tehisintellekti valdkonna teadus- ja arendustegevusse, luua kaasav ja toetav ökosüsteem, kujundada koostalitlusvõimeline poliitikakeskkond, arendada inimeste oskusi ja valmistuda tööjõu muutusteks ning edendada rahvusvahelist koostööd usaldusväärse tehisintellekti nimel [49].

Euroopa deklaratsioon digiõiguste ja -põhimõtete kohta digikümnendiks [50] käsitleb tehisintellekti kui vahendit, mille lõppeesmärk on suurendada inimeste heaolu. Selle kohaselt peaks igaühel olema õigus kasutada AI-süsteemide eeliseid ja teha teadlikke valikuid digikeskkonnas, olles samas kaitstud riskide ja kahju eest, mis puudutavad tervist, ohutust ja põhiõigusi. Deklaratsiooni kohaselt tuleb tagada ka ohutu ja tervislik töökeskkond ning see, et tehisintellekti kasutamine töökohal on läbipaistev ja et selle kasutamisel järgitakse riskipõhist lähenemisviisi. Samuti tuleb tagada, et töötajaid mõjutavad olulised otsused tehakse inimjärelevalve all ning töötajaid teavitatakse, kui nad puutuvad kokku AI-süsteemidega.

Näide. Ebavõrdsuse süvenemine

Tehisintellekti rakendused võivad süvendada ebavõrdsust nt töölevõtmisel, kus algoritmid eelistavad mehi, või tervishoius, kus naisi võidakse valediagnoosida, kuna mudelid põhinevad peamiselt meestelt kogutud andmetel. Hariduses võivad algoritmid alahinnata tüdrukute võimalusi reaalainetes, suurendades välja langemise ohtu ja piirates juurdepääsu edasistele õppeprogrammidele. [2]

Euroopa Parlamendi 2024. aasta uuring toob esile juhtumeid, mis näitavad AI väärkasutuse ulatust, ning avab eelnevaga seotud poliitilisi, regulatiivseid ja diplomaatilisi raskuspunkte. [51] Viidatud uuringu kohaselt:

- tehisintellekti väärkasutus autoritaarsetes režiimides muutub üha suuremaks probleemiks, tugevdades desinformatsiooni levikut ja teisitimõtlejate mahasurumist;
- EL peab investeerima eetilise AI-ga seotud teadus- ja arendustegevusse, et edendada läbipaistvaid ja väärkasutust välistavaid tehnoloogiaid;
- tõhusad sanktsioonid ja vastutusmehhanismid on vajalikud, et karistada inimõiguste rikkumiste eest ja käsitleda tarneahela nõrkusi;
- innovatsiooni ja range eetilise vastavuse tasakaal on kriitilise tähtsusega, et vältida EL-i tehnoloogiate kasutamist repressioonideks;
- teadusmaailma, erasektori ja kodanikuühiskonna tihedad partnerlussuhted on olulised varajase hoiatuse ja uuenduslike lahenduste jaoks;
- EL peaks kehtestama globaalsed AI juhtimisstandardid tugeva õigusraamistiku ja aktiivse rahvusvahelise koostöö kaudu. EL peab sisemiselt säilitama kõrged inimõiguste ja läbipaistvuse standardid, et tugevdada oma ülemaailmset usaldusväärsust;
- tehisintellekti kahesugune kasutus nõuab täpsustatud ekspordikontrolli, et vältida väärkasutust.

Euroopa Komisjoni 2025. a raport [2] käsitleb üldotstarbelise tehisintellekti (GPAI) muutvat rolli EL-is. Raport rõhutab GPAI potentsiaali innovatsiooni, tootlikkuse ja ühiskondliku arengu edendamisel, kuid toob esile ka ohud, nagu väärinfo levitamine, kallutatus, töökohtade ümberkujundamine ja privaatsusriskid. Arutletakse GPAI tehnoloogiliste võimaluste, majanduslike mõjude ja ühiskondlike tagajärgede üle. Lisaks vaadeldakse EL-i regulatiivseid raamistikke (nt AI määrust ja andmeseaduseid), mis peaksid tagama usaldusväärse ja läbipaistva arengu. Raport rõhutab vajadust tasakaalustatud poliitikameetmete järele, et maksimeerida GPAI hüvesid ja vähendada riske, järgides samal ajal demokraatlikke väärtusi ja EL-i õigusraamistikku.

2025. aastal avalikustati ka „Üldotstarbelise tehisintellekti hea tava koodeks"(„*The General-Purpose AI Code of Practice*“). See on (vabatahtlik) tööriist, mis aitab tagada kooskõla AI määruse nõuetega, mis puudutavad just GPAI mudeleid. Viidatud koodeksi kolmas, AI üldotstarbeliste mudelite ohutusele ja turvalisusele pühendatud jaotis („*Code of Practice for General-Purpose AI Models: Safety and Security Chapter*“) [52] rõhutab, et süsteemsete riskide juhtimisel tuleb vältida kallutatust ja soodustada erapooletut hindamist. Näiteks rõhutatakse vajadust edendada organisatsioonis elutervet riskikultuuri, mis võimaldab töötajatel otsuseid vaidlustada ja vältida liigset riskimist (lk 25, meede 8.3). Samuti juhitakse tähelepanu märgistamise kallutatuse vältimisele mudelite hindamises (lk 30). Eraldi tuuakse välja diskrimineeriv kallutatus kui kriitiline mudeliomadus, mis võib põhjustada kahjulikke tagajärgi, näiteks hallutsineerimist (lk 36, lisa 1.3.2).

Näide. Üldotstarbeline tehisintellekt (GPAI)

Üldotstarbeline tehisintellekt võib tugevdada olemasolevat kallutatust ja stereotüüpe, eriti kui mudelid on treenitud ebavõrdust kajastavatel andmetel. Näiteks võib krediidiriski hindamisel GPAI süvendada soolist kallutatust [2].

Eelnev rõhutab vajadust õiglase ja läbipaistva otsustusraamistiku järele ning nõuab kallutatuse vähendamist ja vastutuse tagamist. Usaldusväärse tehisintellekti arendamiseks on seega vaja mitmekesist treeningandmestikku ja õiglust arvestavaid algoritme (*fairness-focused algorithms*). Oluline on regulaarselt auditeerida süsteeme ning suurendada AI-süsteemide arendajate mitmekesisust. EL tasandil nõuab näiteks digiteenuste määrus [53] suurplatvormidelt iga-aastaseid riskihindamisi, et tuvastada võimalikud diskrimineerimised ja ohud põhiõigustele. [2]

Mitmekesisus, õiglus ja diskrimineerimise vältimine on usaldusväärse tehisintellekti loomise aluseks. Ebaõiglase kallutatuse kõrvaldamine on oluline, et vältida haavatavate rühmade marginaliseerimist ning eelarvamuste ja diskrimineerimise süvendamist. Tehisintellekti õiglust hinnatakse kaitstud omaduste (nt rass, sugu, vanus, puue) alusel kooskõlas hartaga. AI-süsteeme tuleks hinnata erinevate kaitstud rühmade lõikes, eriti kõrge riskiga juhtumite puhul, ning kogu elutsükli vältel tuleb püüelda diskrimineerivate või kallutatud tulemuste vältimise poole, et tagada õiglus ja järgida EL-i diskrimineerimisvastaseid õigusakte. [2]

Euroopa Komisjonil on plaanis 2025. aasta lõpuks avaldada kaks uut strateegiadokumenti: „Apply AI“ strateegia [54], mis keskendub AI-süsteemi potentsiaali ärakasutamisele sihtsektorites ja avalike teenuste pakkumisel, ning Euroopa teaduse tehisintellekti strateegia („European Strategy for AI in Science“) [55], mis edendab AI vastutustundlikku kasutamist teaduses ja innovatsioonis. Lisaks eeldatakse, et Euroopa Komisjon esitab 2026. aasta esimeseks kvartaliks tehisintellekti ja pilvandmetöötluse arendamise määruse („Cloud and AI Development Act“ [56]), mille eesmärk on soodustada investeeringuid pilv- ja servandmetöötlusesse [57].

3.2.2 Isikuandmete kaitse ja andmehaldus

Isikuandmete kaitse küsimused on Euroopas põhjalikult reguleeritud. Nõuded isikuandmete kaitsele võivad tuleneda nii EL-i kui ka riigisisestest õigusnormidest. Vastavalt olukorrale, võib olla vajalik isikuandmete kaitse küsimustes lähtuda näiteks Euroopa Parlamendi ja nõukogu määrusest (EL) 2016/679 (IKÜM) [58], Euroopa Parlamendi ja nõukogu direktiivist (EU) 2016/680 (õiguskaitseasutuste andmekaitsedirektiiv) [59] või Euroopa Parlamendi ja nõukogu määrusest 2018/1725 (nõuded isikuandmete töötlemisele EL-i institutsioonides, organites ja asutustes) [60]. Abi võib olla ka andmekaitseorganisatsioonide suunistest ja nende poolt pakutud tõlgendustest, aga ka erinevatest juhendmaterjalidest või õiguskirjandusest seoses isikuandmete töötlemisega AI-süsteemides, nt [61, 62, 63].

Isikuandmete kaitse seadus (IKS) [64] täpsustab ja täiendab isikuandmete kaitse üldmääruse (EL) 2016/679 [58] (IKÜM) sätteid ja kehtestab normid direktiivi (EL) 2016/680 ülevõtmiseks. IKS sätestab mitmes paragrahvis (nt §4, §5, §6 lg 3 p 3, §10 lg 2 p 3), et andmetöötlus ei tohi ülemäära kahjustada andmesubjekti õiguseid. Lisaks näeb IKS ette, et keelatud on teha üksnes automaattöötlusel põhinevat otsust, kui see on andmesubjektile negatiivsete tagajärgedega. Samuti on keelatud füüsilisi isikuid eriliiki isikuandmete alusel diskrimineeriv profiilianalüüsil põhinev otsus, va juhul, kui otsuse tegemine on lubatud seadusega, milles on sätestatud asjakohased meetmed andmesubjekti õiguste ja vabaduste ning õigustatud huvide kaitseks (vt IKS § 21 lg 1).

Samuti on AI määruse põhjenduspunktis 10 välja toodud, et AI-süsteemide projekteerimisel,

arendamisel või kasutamisel, mis hõlmab isikuandmete töötlemist, tuleb arvesse võtta isikuandmete töötlemisega seotud nõudeid. Täpsustatud on, et andmesubjektide jaoks peavad säilima kõik EL-i õigusest tulenevad õigused ja tagatised, sh need, mis puudutavad automaatseid üksikotsuseid ja profiilianalüüsi. Järgnev näide illustreerib kujukalt automaatse andmeanalüüsiga seotud väljakutseid.

Näide. Automaatne andmeanalüüs – kaasus nr 1 BvR 1547/19, 1 BvR 2634/20 [65, 66]

2023. aasta veebruaris tegi Saksamaa Liiduvabariigi Konstitutsioonikohus otsuse politsei automaatse andmeanalüüsi põhiseaduspärasuse kohta. Kohus leidis, et isikuandmete töötlemine automaatanalüüsi teel kujutab endast sekkumist isikuandmete enesemääramise õigusesse, mis tuleneb Saksamaa põhiseadusest (*Grundgesetz*). Teisene andmetöötlus nõuab täiendavat õiguslikku alust, lähtudes proportsionaalsuse ja eesmärgikohasuse põhimõtetest. Otsuses rõhutatakse, et ainult seadusandja võib määratleda reeglid kogutavate andmete liigi, ulatuse ja lubatud analüüsimeetodite kohta, samas kui haldusasutused võivad rangete kontrollimehhanismide alusel täpsustada tehnilisi detaile. Kui automaatne andmeanalüüs mõjutab oluliselt põhiõigusi, on see lubatud üksnes eriti kaaluka õigushüve kaitseks ja rangete kaitsemeetmete olemasolul.

IKÜM-i rakendamine AI-süsteemidele võib tekitada ebakõlasid, kuna tehisintellekti omadused (nt läbipaistmatus ja ulatuslik andmekasutus) võivad olla vastuolus IKÜMi artiklis 5 sätestatud üldpõhimõtete – nt seaduslikkuse, läbipaistvuse, eesmärgipärasuse, andmete minimeerimise ja vastutusega [63]. Samuti on IKÜM eriliiki andmete töötlemise osas rangem, mis võib tekitada juriidilist ebakindlust [67]. Samas on leitud [68], et eesmärgipärasuse ja andmete minimaalsuse põhimõtteid saab IKÜMi raames paindlikult rakendada ka tehisintellekti toetamiseks. Eesmärgipärasus võimaldab andmete taaskasutust, kui see on kooskõlas algse andmete kogumise eesmärgiga. Andmete minimaalsus võib tähendada eelkõige tuvastatavuse vähendamist (nt pseudonüümimise kaudu), mitte tingimata andmemahtude piiramist.

AI-süsteemi vastavus IKÜM-ile tagab, et selles kasutatud isikuandmed on piisavad, asjakohased ja selgelt määratletud eesmärgi jaoks vajalikud. Vastutavad töötlejad peavad seetõttu hoolikalt läbi mõtlema ja põhjendama, miks iga andmeelement on vajalik – seda tuleb analüüsida eraldi igas AI elutsükli etapis (vt jaotist 2.3). Hea ülevaade andmetest annab omakorda paremad võimalused täita ka AI määruse nõudeid, nt tuvastada ja leevendada kallutatust. AI määrus kehtestab reeglid tundlike andmete töötlemiseks ja kallutatuse aktiivseks ohjamiseks, tagades seeläbi õigluse, läbipaistvuse ning ka andmete minimaalsuse põhimõtetest kinnipidamise AI-süsteemides. [69]

Andmekaitse Inspektsioon on jaganud näpunäiteid selle osas, kuidas kaitsta inimeste privaatsust ja tagada andmekaitse olukorras, kus AI-süsteem töötleb isikuandmeid [70]. Nende kohaselt tuleb andmeid koguda ja töödelda üksnes selgelt määratletud ja põhjendatud eesmärkidel ning tagada läbipaistvus ja turvalisus. Oluline on aru saada ka sellest, kas tehisintellektile antud sisend võib sisaldada isikuandmeid, asutusesiseseks kasutamiseks mõeldud teavet või ärisaladust.[2] Inimesi tuleb teavitada sellest, kuidas nende andmeid kasutatakse, ja võimaldada neil oma

[2] Avaliku teabe seaduse (AvTS) [71] eesmärk on tagada üldiseks kasutamiseks mõeldud teabele avalikkuse juurdepääsu võimalus. AvTS näeb § 3^1 lõikes 3 ette, et 'teabe üldiseks kasutamiseks andmisel peab olema tagatud isiku eraelu puutumatus, autoriõiguste kaitse, riigi julgeoleku kaitse, ärisaladuse ja muu juurdepääsupiiranguga teabe kaitse.' AvTS § 34 lg 1 sätestab, et piiratud juurdepääsuga teabeks loetakse sellist teavet, millele juurdepääs on seadusega kehtestatud korras piiratud. Piiratud juurdepääsuga teabe sisestamine avalikku AI-süsteemi ei ole lubatud, kuna ei ole võimalik välistada sellise teabe lekkimist mudeli teenustajale või juurutajale, st et teabe edasise töötlemise osas puudub kontroll.

andmetega tutvuda ja vajaduse korral nõuda ebaõigete andmete parandamist.

Järgmise näite puhul kohustas kohus *Clearview AI*-d teavitama andmesubjekte võimalusest oma andmed andmebaasist kustutada.

Näide. *ACLU v. Clearview AI* Biomeetrilised andmed [72, 73]

Kohtuasjas *Ameerika Kodanikuõiguste Liit (American Civil Liberties Union, ACLU) v. Clearview AI* vaidlustas ACLU Clearview AI tegevuse, väites, et ettevõte rikkus Illinois' seadust, mis käsitleb biomeetriliste andmete töötlemist (*Biometric Information Privacy Act, BIPA*) ja puudutab ilma inimeste nõusolekuta nende näopiltide kogumist ja kasutamist. Kohtuliku kokkuleppe kohaselt on Clearview-l keelatud anda juurdepääs oma näopiltide andmebaasile USA-s, kohalikele asutustele on juurdepääs keelatud viieks aastaks. Illinois' elanikele on antud õigus paluda enda andmete kustutamist andmebaasist (nn *opt-out*) ja ettevõttel on kohustus kõnealust *opt-out*-mehhanismi avalikult reklaamida. Juhtum näitab, et tugevad privaatsusseadused võivad oluliselt piirata massilist biomeetrilist jälgimist.

Järgnevalt kirjeldatud Cambridge Analytica skandaalis koguti Facebooki kasutajate andmeid ilma enamiku inimeste teadliku nõusolekuta. Andmed pärinesid nii rakendust kasutanud isikutelt kui nende sõpradelt ja neid kasutati poliitiliste profiilide koostamiseks ja sihitud reklaamimiseks.

Näide. Cambridge Analytica skandaal [74, 70, 75]

Cambridge Analytica skandaal näitas, kuidas algoritmiline järeldamine võib tekitada kallutatust. Isiksuse testide ja sotsiaalsete kontaktide abil lõi Cambridge Analytica miljonitele Facebooki kasutajatele psühholoogilised profiilid, mida seejärel kasutati poliitilise sõnumite jaoks. Kõnealune protsess võimendas olemasolevaid sotsiaalseid ja poliitilisi kallutatuseid, kuna algoritm tugines peamiselt demograafilistele korrelatsioonidele ja oletustele, mitte individuaalsetele käitumismustritele. Skandaali eest määras Briti informatsioonikomisjon (ICO) USA sotsiaalmeediahiidule Facebook trahvi summas 500 000 naela (ligikaudu 565 300 eurot). Juhtum rõhutas, et käitumuspõhine sihistus (*targeting*) võib tahtmatult tugevdada diskrimineerivaid mustreid või ideoloogilisi kõlakambreid, mistõttu on suur vajadust AI ja algoritmiliste süsteemide regulatiivse järelevalve järele, et tuvastada ja vähendada kallutatust ennustavates mudelites ja andmepõhistes otsustes.

AI-süsteemide töödeldavate isikuandmete mastaapsus võib endast kujutada suurt riski põhiõigustele, eeskätt privaatsusele ja mittediskrimineerimisele [76]. Ka küberturvalisuse 2. direktiiv (NIS2) [77] sedastab, et uue tehnoloogia, nagu tehisintellekti kasutamine, peab olema kooskõlas EL-i andmekaitse nõuetega, sh põhimõtetega nagu andmete täpsus, võimalikult väheste andmete kogumine, õiglus ja läbipaistvus ning konfidentsiaalsus (krüpteerimine), ning IKÜM-is sätestatud lõimitud ja vaikimisi andmekaitse nõuetest tuleb täielikult kinni pidada (pp 51).

Ohu võimalikkuse ja tõsiduse hindamiseks, arvestades mh töötlemise laadi, ulatust, konteksti ja eesmärke, tuleks enne töötlemist koostada andmekaitsealane mõjuhinnang (IKÜM art 35, pp 90). Andmekaitsealasele mõjuhinnangule on heaks täienduseks AI-süsteemide mõjuhinnangud, mis hõlmavad ka eetilisi ja sotsiaalseid aspekte [76].

Näide. *Toeslagenaffaire* juhtum Hollandis [76]

Hollandi *Toeslagenaffaire* ehk lastetoetuste skandaal oli seotud maksuhalduri algoritmiga, mis pidi avastama toetuspettusi. Süsteem lõi ekslikke riskiprofiile, sihtides sageli topelt-kodakondsusega või migranditaustaga inimesi, ja tembeldas tuhanded pered alusetult petturiteks. Selle tagajärjel kaotasid paljud pered toetused, sattusid ränka majanduslikku kitsikusse ja kogesid pikaajalisi sotsiaalseid tagajärgi. Skandaali tulemusena astus 2021. aastal tagasi kogu Hollandi valitsus ning juhtum on kujunenud ilmekaks näiteks, kuidas algoritmiline kallutatus võib põhjustada süsteemset ebaõiglust.

AI määruse artikli 10(1) kohaselt tuleb suure riskiga AI-süsteemide puhul treenimis-, valideerimis- ja testimisandmestike suhtes kohaldada asjakohaseid andmehaldus- ja juhtimistavasid. Arvestama peab sama artikli lõikes 2 nimetatud nõuetega, mille hulka kuulub viidatud andmestike läbivaatus kallutatuse välistamiseks (art 10(2)(f)) ja asjakohaste meetmete rakendamise kohustus kallutatuse avastamiseks, ennetamiseks ja leevendamiseks (art 10(2)(g)). AI määruse § 10(2)(3) nõuab, et treenimis-, valideerimis- ja testandmestikud peavad olema asjakohased, piisavalt representatiivsed, võimalikult suurel määral vigadeta ja täielikud, võttes arvesse andmete sihtotstarvet, ning asjakohaste statistiliste omadustega sihtrühma osas. Lisaks täpsustab lõige 4, et andmestikes tuleb sihtotstarbe jaoks vajalikus ulatuses võtta arvesse omadusi või elemente, mis iseloomustavad konkreetset geograafilist, kontekstuaalset, käitumuslikku või funktsionaalset keskkonda, kus kavatsetakse suure riskiga AI-süsteemi kasutada. Katerina Yordanova on AI määruse artiklit 10 põhjalikult kommenteerinud AI määruse kommentaarides (vt lk 259–283 [41]). Loe ka AI määruse põhjenduspunktides 66, 67, 68, 69 ja 70 kirjutatut.

AI määrus kehtestab suure riskiga AI-süsteemide jaoks kohustuse tuvastada ja leevendada kallutatust, eelkõige artikli 10(5) kaudu. Kallutatuse juhtimist on käsitletud pideva kohustusena kogu AI-süsteemi elutsükli vältel, võttes arvesse täpsuse, läbipaistvuse ja õigluse põhimõtteid. Samas püsib pingeseis IKÜMi andmete minimaalse kogumise põhimõtte ja AI määruse nõude vahel koguda piisavalt esinduslikke andmeid kallutatuse vältimiseks, mistõttu on vajalik hoolikas plaanimine nende kohustuste tasakaalustamiseks. [78]

Lisaks eelnevale mängivad olulist rolli ka auditid. [79] käsitleb, kuidas regulaatoritel on võimalik vastata algoritmide kasvava kasutamisega seotud väljakutsetele eri sektorites. Aruanne esitab kokkuvõtte olemasolevast auditeerimise maastikust Ühendkuningriigis, toob esile lüngad ning arutleb, kuidas asutused võiksid koordineeritult algoritmilisi süsteeme jälgida. Peamisteks teemadeks on vastutuse tagamine, läbipaistvus ja auditeerimistavade kohandamine tehnoloogiliste arengutega. Oluliselt rõhutatakse, et algoritmiline kallutatus on suur risk, sest automaatsüsteemid võivad tasakaalustamata andmetel treenituna korrata või süvendada ebavõrdsusi. Seetõttu on vaja süsteemseid auditeid, mis hindavad diskrimineerivaid tulemusi, eriti rahanduse, tervishoiu ja veebisisu modereerimise valdkonnas.

Tehisintellekti käsitleva määruse (EL) 2024/1689 artikkel 10 lõige 5:

Niivõrd kui see on rangelt vajalik selleks, et tagada kallutatuse avastamine ja korrigeerimine suure riskiga AI-süsteemide puhul kooskõlas käesoleva artikli lõike 2 punktidega f ja g, võivad selliste süsteemide pakkujad erandkorras töödelda isikuandmete eriliike, kohaldades asjakohaseid kaitsemeetmeid füüsiliste isikute põhiõiguste ja -vabaduste kaitseks. Lisaks määruste (EL) 2016/679 ja (EL) 2018/1725 ning direktiivi (EL) 2016/680 sätetele peab selline töötlemine vastama kõigile järgmistele tingimustele:

a) kallutatuse tuvastamist ja parandamist ei ole võimalik tulemuslikult saavutada muude andmete, sealhulgas tehisandmete või anonüümitud andmete töötlemisega;

b) isikuandmete eriliikide suhtes kehtivad isikuandmete taaskasutamise tehnilised piirangud ning tipptasemel turva- ja privaatsuse säilitamise meetmed, sh pseudonüümimine;

c) isikuandmete eriliikide suhtes kohaldatakse meetmeid, millega tagatakse töödeldavate isikuandmete turvalisus, kaitse ja asjakohased kaitsemeetmed, sh juurdepääsu range kontroll ja dokumenteerimine, et vältida väärkasutamist ja tagada, et ainult volitatud isikutel, kellel on asjakohased konfidentsiaalsuskohustused, on juurdepääs kõnealustele isikuandmetele;

d) isikuandmete eriliike ei tohi edastada ega üle anda ning need ei tohi olla teistele isikutele muul moel kättesaadavad;

e) isikuandmete eriliigid kustutatakse, kui kallutatus on parandatud või kui isikuandmete säilitamisperiood saab läbi, olenevalt sellest, kumb saabub varem;

f) määruste (EL) 2016/679 ja (EL) 2018/1725 ning direktiivi (EL) 2016/680 kohane isikuandmete töötlemise toimingute registreerimine hõlmab põhjuseid, miks isikuandmete eriliikide töötlemine oli rangelt vajalik kallutatuse avastamiseks ja parandamiseks ning miks seda eesmärki ei olnud võimalik saavutada muude andmete töötlemisega.

AI määrus näeb ette, et suure riskiga AI-süsteemid pevad olema projekteeritud ja arendatud nii, et tagatud oleks süsteemide toimimise piisav läbipaistvus, et juurutajal oleks võimalik süsteemi tulemusi tõlgendada ja seda mõistlikult kasutada (art 13(1)). See nõue on oluline ka avaliku sektori puhul, kui pakkujalt ostetakse AI-süsteem. Avaliku sektori organisatsioon kui juurutaja, kes hakkab AI-süsteemi rakendama, peab AI-süsteemi toimeloogikast piisavalt aru saama, et seda mõistlikult kasutada.

AI-süsteemid on sõltuvuses suurtest andmehulkadest. EL-i andmemääruse [80] kohaselt on andmepõhisel tehnoloogial olnud ümberkujundav mõju kõigile valdkondadele. Seetõttu soovitakse luua ühtne andmeturg, rõhutades samas õiglast juurdepääsu, kasutajaõigusi ja isikuandmete kaitset. EL andmemäärusega kehtestatakse andmevaldajatele kohustus teha andmed mõnedel asjaoludel kasutajatele ja kasutajate valitud kolmandatele isikutele kättesaadavaks (pp 5). Selgitatud on, et ühtegi andmemääruse sätet ei tohiks kohaldada ega tõlgendada viisil, mis vähendab või piirab õigust isikuandmete kaitsele või õigust eraelu puutumatusele ja side konfidentsiaalsusele (pp 7).

EL-i andmehalduse määruse [81] eesmärk on soodustada andmete taaskasutamist ja jagamist. Määrusega kehtestatakse mh avaliku sektori asutuste valduses olevate teatavate kaitstud andmete taaskasutamise tingimused [82]. Andmehalduse määrusega soovitakse arendada piirideta digitaalset siseturgu ning inimkeskset, usaldusväärset ja turvalist andmeühiskonda ja -majandust (pp 3). Andmehalduse tõhusaks korraldamiseks ja andmekvaliteedi tagamiseks vt ka näiteks Eesti andmehalduse raamistikku [83], asutuse ülesandeid andmekvaliteedi tagamisel [84], andme-

kvaliteedi juhist [85], andmekvaliteedi halduse kasutuslugusid ja funktsionaalsust [86], andmekirjelduse juhist [87] (ning selle lisasid 1 [88] ja 2 [89]), ärireeglite kirjeldamise praktilisi näiteid [90], andmekogude andmekoosseisude uuendamist RIHAs [91], andmeladude juhist [92] ning juhendmaterjali andmete annoteerimiseks [93].

3.2.3 Küberturvalisus ja tooteohutus

Põhiõiguste ja isikuandmete kaitse tagamiseks on hädavajalik rakendada tõhusaid ja kaasaegseid küberturbemeetmeid. Üldprintsiip on, et tarbijale mõeldud tooted peavad olema ohutud [94] ning toodet, mis ei ole ohutu, ei või turule lasta ega kasutusele võtta (TNVS [95] § 5 lg 1).

AI määrus keskendub AI-süsteemide ohutusele ja näeb ette, et need peavad olema täpsed, stabiilsed ja turvalised (vt art 15). Suure riskiga AI-süsteeme tuleb projekteerida ja arendada selliselt, et need saavutaksid asjakohase täpsuse, stabiilsuse ja küberturvalisuse taseme kogu elutsükli vältel (AI määrus art 15 lg 1). Viidatud täpsuse tasemed ja asjakohased täpsuse parameetrid tuleb dokumenteerida süsteemiga kaasas olevas kasutusjuhendis (AI määrus art 15 lg 3).

AI määruse artikli 15 lõige 4 näeb ette, et suure riskiga AI-süsteemid peavad olema süsteemis või süsteemi töökeskkonnas tekkida võivate vigade, rikete või ebakõlade suhtes võimalikult vastupidavad. Suure riskiga AI-süsteeme, mis õpivad edasi ka pärast turule laskmist või kasutusele võtmist, tuleb arendada selliselt, et kõrvaldada või minimeerida potentsiaalselt kallutatud tulemuste risk, mis võib mõjutada edasiste toimingute sisendit (tagasisideahelad). Need AI-süsteemid peavad pidama vastu volitamata kolmandate isikute katsetele mõjutada süsteemi tulemusi või selle toimimist, kasutades ära süsteemi nõrkusi (AI määrus art 15 lg 5).

Need nõuded on olulised mitte ainult suure riskiga AI-süsteemide puhul, vaid ka teiste avalikus sektoris kasutatavate AI-süsteemide puhul, sest avalike teenuste kasutus sõltub inimeste usaldusest oma riigi, selle ametkonna ja riigi poolt pakutavate lahenduste vastu. Seega peavad AI määruse artikli 15 lõike 5 kohaselt erinevate nõrkuste käsitlemiseks kasutatavad tehnilised lahendused hõlmama tõhusaid meetmeid, millega hoida ära, avastada, tõrjuda, lahendada ja ohjata ründeid.

Riskihaldussüsteemi eesmärk peaks olema teha kindlaks ja leevendada AI-süsteemide riske tervisele, turvalisusele ja põhiõigustele (AI määrus pp 65). Küberturvalisust aitavad tagada ka järgmiste õigusaktide asjakohaste nõuete täitmine – küberturvalisuse 2. direktiiv [77], küberkerksuse määrus [96], küberturvalisuse seadus [97].

Küberturvalisuse 2. direktiiviga (NIS2) [77] sätestatakse meetmed, mille eesmärk on saavutada küberturvalisuse ühtlaselt kõrge tase kogu EL-is (art 1(1)). NIS2 direktiivi nõuded on Eesti õigusesse üle võetud KüTSiga. KüTS sätestab ühiskonna toimimise seisukohast oluliste, sh avaliku sektori võrgu- ja infosüsteemide pidamise nõuded, vastutuse ja järelevalve ning küberintsidentide ennetamise ja lahendamise alused (§ 1 lg 1).

Digielemente sisaldava toote, mis on ühtlasi suure riskiga AI-süsteem, küberriskide hindamisel tuleb lähtuda toote erinevatest etappidest ja võtta arvesse AI-süsteemi küberkerksust ohustavaid riske seoses volitamata kolmandate isikute katsetega muuta süsteemi kasutamist, käitumist või toimivust. Arvestada tuleb tehisintellektile omaste nõrkustega, nagu andmemürgitus või vasturünded. Asjakohasel juhul tuleb võtta arvesse ka riske põhiõigustele kooskõlas AI määrusega (küberkerksuse määrus pp 51; vt ka art 12).

Küberturvalisuse ja tooteohutuse nõuete järgimine on AI-süsteemide puhul ülioluline. AI-süsteemide turvalisus ei piirdu ainult tehniliste kaitsemeetmetega, vaid hõlmab ka laiemat vastutust inimeste ja ühiskonna ees tervikuna. Samal ajal on hädavajalik, et tehisintellekt vastaks rangeimatele

tooteohutuse nõuetele, vältides riske ja tagades kasutajate kaitse terves AI-süsteemi elutsüklis. Integreerides küberturvalisuse ja tooteohutuse põhimõtted juba projekteerimise ja arendusprotsessi algfaasis, on võimalik kujundada usaldusväärseid, vastupidavaid ja inimkeskseid tehisintellekti lahendusi, mis toetavad Euroopa ja Eesti väärtusi ja põhimõtteid ning suurendavad ühiskonna usaldust tehnoloogia vastu.

3.3 Millised standardid soovitavad AI kallutatusega tegelemist?

AI-süsteemide arendamisel ja hindamisel on hea järgida rahvusvaheliselt tunnustatud standardeid, et vähendada AI-süsteemi kallutatust ja tagada selle usaldusväärsus. Standardite valimise protsess võib sisaldada järgmisi samme.

1. **Vajalike standardite identifitseerimine** – analüüsitakse, millised standardid ja juhised on asjakohased.
2. **Sobivuse hindamine** – valitud standardi(te) rakendatavust konkreetse AI-süsteemi kontekstis hinnatakse, võttes arvesse kasutusjuhtumeid, andmete eripära ja eetilisi aspekte.
3. **Standardite integreerimine metoodikasse** – standardi(te) nõuded ja parimad praktikad integreeritakse andmete kogumise, mudelite treenimise ja hindamise protsessi.
4. **Jälgimine ja dokumenteerimine** – kõik sammud ja otsused dokumenteeritakse, sh kasutatud standardid, rakendatud meetodid kallutatuse tuvastamiseks ja vähendamiseks ning tulemuste jälgimine.

AI-süsteemi kallutatusega on otseselt seotud tabelis 1 toodud rahvusvahelised standardid. Standardid, mis aitavad leevendada AI-süsteemi kallutatust kaudsemalt, on loetletud lisas A.

Tabel 1. AI-süsteemi kallutatuse leevendamisega seotud standardid

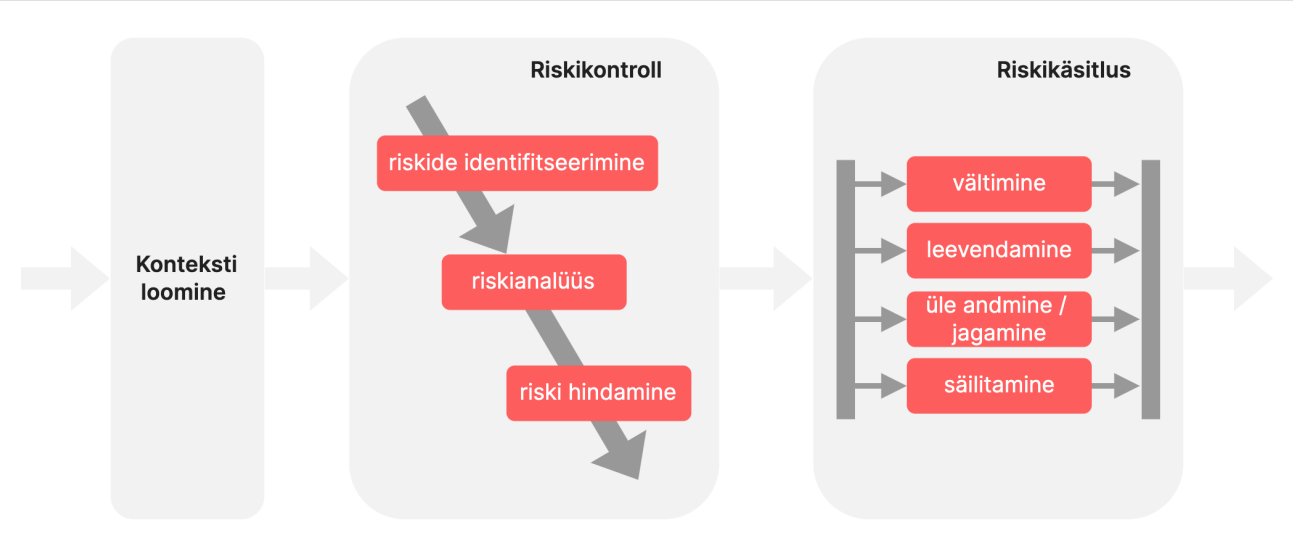
Standardi number	Standardi nimi	Selgitus
ISO/IEC 42005 [43]	AI-süsteemi mõjuhinnang	Standard aitab koostada AI-süsteemi mõjuhinnagut. Hinnangu käigus selgitatakse välja, kuidas kavandatavad AI-süsteemid ja AI-süsteeme sisaldavad rakendused võivad mõjutada inimesi, ühiskonnagruppe või ühiskonda tervikuna. Standard toetab läbipaistvust, vastutust ja usaldust AI-süsteemide vastu, võimaldades organisatsioonil tuvastada, hinnata ja dokumenteerida potentsiaalseid mõjusid terve AI-süsteemi elutsükli jooksul
ISO/IEC 24027:2021 [98]	Kallutatus AI-süsteemides ja tehisintellekti abil tehtavates otsustes (Bias in AI systems and AI aided decision making)	Standard käsitleb AI-süsteemidega seotud kallutatust, eriti AI-põhiste otsuste tegemisel. Standard aitab valida kallutatuse hindamise mõõtmistehnikat ja -meetodit. Hõlmatud on kõik AI-süsteemi elutsükli etapid, sh andmete kogumine, treenimine, pidev õpe, disain, testimine, hindamine ja kasutamine

Standardi number	Standardi nimi	Selgitus
IEEE 7003-2024 [99]	Algoritmilise kallutatuse kaalutlusi (Algorithmic Bias Considerations)	Standard aitab algoritmide arendajatel tuvastada ja vähendada kallutatust. See pakub juhiseid valideerimisandmestike valikuks, rakenduspiiride määratlemiseks ning kasutajate ootuste haldamiseks, et vältida algoritmide tahtmatut väärkasutust ja vääritimõistmist
NIST AI 100-1:2023 [100]	Tehisintellekti riskihalduse raamistik (AI Risk Management Framework)	Raamistik toetab AI-süsteemi riskihaldusprotsessi. NIST eristab kolme põhitüüpi kallutatust: (1) süsteemne, (2) arvutuslik/statistiline ja (3) inim-kognitiivne kallutatus – need võivad esineda ka ilma tahtliku eelarvamuse või diskrimineerimiseta ning võivad võimendada kahju, kui neid ei käsitleta. Vt ka [101]
NIST AI 600-1:2024 [102]	Generatiivse tehisintellekti profiil (Generative AI Profile)	Standard aitab tuvastada kallutatust generatiivsetes AI-mudelites. Kahjulik kallutatus tehisintellektis võib võimendada ajaloolisi ja süsteemseid ebavõrdsusi, põhjustada sooritusvõime erinevusi rühmade vahel mittetüüpilise treeningandmestiku tõttu ning viia viia põhjendamatute otsusteni. Lisaks võivad inimese ja AI-süsteemi vahelised suhted põhjustada probleeme, nagu liigne usaldus, kalduvus eelistada automaatotsuseid, emotsionaalne kiindumus või AI inimlikustamine

4 Kallutatuse käsitlemine AI-süsteemi riskide haldamisel

4.1 Kuidas kallutatust riskihalduses arvestada?

Riskihaldusprotsess koosneb kolmest sammust: konteksti loomisest, riskikontrollist ja riskikäsitlusest (vt joonis 3). Konteksti loomisel kirjeldatakse ja dokumenteeritakse süsteem ja huvipooled ning määratakse organisatsiooni riskivalmidus, riskinorm ja riskiomanikud. Riskikontrolli käigus tuvastatakse riskid, analüüsitakse ja hinnatakse neid. Riskikäsitluse käigus otsustatakse, mida riskidega teha – vältida, leevendada, jagada või säilitada.



Joonis 3. Riskihalduse protsess, kohandatud allikast [103]

Kuigi AI määruse järgi on riskihaldusprotsess kohustuslik ainult kõrge riskiga süsteemide korral, on riskide analüüsimine ja hindamine tugevalt soovitatav ka madalama riskiga süsteemide jaoks. Olgugi, et formaalset riskihaldusprotsessi ei pea üles seadma, aitab riskide tuvastamine ja leevendamine süsteemi elutsükli varastes etappides ennetada ohtude realiseerumist või leevendada nende mõju.

Riskihalduse standardid ja raamistikud

Riskihaldust kirjeldab standard ISO 31000 [104] ja NIST SP 800-37 riskihalduse raamistik (RMF) [105]. Infoturvariskide eripärasid käsitleb ISO/IEC 27005 [106] ja küberturbe eripärasid NIST küberturbe raamistik (CSF) [107]. Standard ISO/IEC 23984 [108] kirjeldab, kuidas laiendada standardiga ISO 31000 vastavuses olevat riskihalduse protsessi organisatsioonile, mis kasutab, arendab või juurutab AI-süsteeme. Eesti infoturbestandard (E-ITS) [109] joondub ISO/IEC 27001 sarjaga ning E-ITS-i rakendajatel on samuti võimalik kasutada riskipõhist lähenemist. Uuringus [13] kirjeldatud AI-süsteemide riskihalduse lihtsustatud metoodika joondub ISO 31000, ISO/IEC 27005 ja E-ITS töövoogudega ning sobib ka NIST RMF ja CSF raamistikega.

4.2 Mida on oluline tähele panna AI-süsteemi konteksti kirjeldamisel?

4.2.1 Süsteemi pass

Esimese sammuna on vaja selgitada välja ja dokumenteerida süsteemiga seotud huvipooled. Süsteemis võib tekkida kallutatus nii kallutatud sisendandmete kui süsteemi loojate kallutatuse tõttu. Kui organisatsioonil on juurdepääs mudeli treenimiseks kasutatud andmestikule, on võimalik seda andmestikku analüüsides tuvastada võimalikku kallutatust. Kui juurdepääs andmetele puudub, peab neid hinnanguid tegema piiratud informatsiooni põhjal. Kui süsteemi või reeglite loojad on kaardistatud, saab hiljem selle teabe põhjal analüüsida, milline on nende võimalik kallutatus, ning selle abil tuvastada süsteemi riskid.

Otseselt mõjutatud huvipooled on:

- teenuse kasutajad,
- teenuse tulemiga seotud isikud, organisatsioonid või ühiskonnagrupid (kelle kohta süsteemi abil otsuseid tehakse),
- süsteemi juurutajat (mainekahju).

Kaudselt mõjutatud huvipooled on:

- andmesubjektid, kelle andmeid kasutatakse mudeli treenimiseks või reeglite loomiseks,
- süsteemi looja (mainekahju).

Kui tegemist on süsteemiga, mis on ehitatud konkreetse eesmärgi jaoks, on lihtne kirja panna süsteemi otsesed ja kaudsed eesmärgid. Kui on olemas konkreetsed näitajad, mille abil mõõta eesmärkide täitmist, paneb organisatsioon kirja ka need. Kui tegemist on probleemi jaoks kohandatud või kitsendatud üldotstarbeliste tehisintellekti (GPAI) mudeliga, siis on küll võimalik kirja panna süsteemi otsesed eesmärgid, kuid raskem on hinnata seda, mis on GPAI ülejäänud (tihti varjatud) eesmärgid. GPAI kasutamisel on väga oluline uurida välja võimalikult palju alusmudeli ja selle treenimiseks kasutatud andmete kohta. Kahjuks tuleb tihti leppida sellega, et need jäävadki läbipaistmatuks.

4.2.2 Süsteemi kasutuslood

Järgmisena kirjeldatakse süsteemi kasutuslugusid. Need võivad osaliselt kattuda süsteemi eesmärkidega, kuid on viimastest konkreetsemad. Vaatleme näidet, kus süsteemi eesmärk on hinnata toetuspettuse võimalikkust ja vältida petturitele toetuste maksmist. Olenevalt rakenduse ehitusest võib kasutuslugu hõlmata näiteks pettusekahtlusega isikute loetelu koostamist või hoopis kõikidele isikutele riskiskoori arvutamist. Need rakendused erinevad selle poolest, et üks väljastab kahtlustatavate nimekirja, teine annab hinnangu kõigile inimestele. Hoolimata sellest, et esimene näide võib sisemiselt kasutada teise näite funktsionaalsust, on süsteemid erineva tulemitüübiga ning seega ka erinevate ohtude ja riskidega (nii õiguslike kui eetilistega).

Süsteem võib sisaldada üht või mitut AI-komponenti, seega võib ühel süsteemil olla mitu kasutuslugu. Kõik kasutuslood tuleb eraldi ära kirjeldada. Oluline on, et iga kasutusloo detailid on dokumenteeritud, siis on hiljem lihtsam saada piisav ülevaade süsteemi ohtudest. Kuna mitmel kasutuslool võivad olla samad ohud, siis käsitletakse neid hiljem võimaluse korral ühiselt. Aga võib ka olla vajalik üht ohtu erinevate kasutuslugude jaoks eraldi käsitleda, näiteks kui ohustsenaarium või kaitsemeetmed on erinevate kasutuslugude jaoks erinevad.

Iga kasutusloo kohta on vaja kirja panna kasutusloo detailid, sisendandmete kirjeldus, kasutuslooga seotud AI-komponendi kirjeldus (kui komponent toetab mitut kasutuslugu, saab seda osa kopeerida) ning informatsioon kasutusloo tulemi ja selle kasutamise kohta. Kõikide andmete kohta lisatakse ka viited allikale, kust info pärineb, et seda oleks hiljem võimalik kiirelt kätte saada ja vajaduse korral uuendada.

Kasutusloo detailidena pannakse kirja, kes kasutusloo algatab ning millal, mida süsteemile sisendiks antakse, mida arvutatakse või hinnatakse ning mida süsteem väljastab. Kirjeldatakse ka, mida tulemiga edasi tehakse.

Sisendandmete puhul tuuakse ilmutatud kujul välja, kas tegemist on isikustatavate andmete, eriliiki isikuandmete või ärisaladusega. Kirjeldatakse ka andmete kogumise viis ning tervikluse ja konfidentsiaalsuse kaitse meetmed.

Tehisintellektikomponendi või kasutatud algoritmi kohta dokumenteeritakse versioon ja loomise aasta, tüüp (algoritm või mudel) ja parameetrid, looja või päritolu, evitusmudel ning kuidas komponent ülejäänud süsteemiga liidestatud on. Võimaluse korral pannakse kirja süsteemi täpsus, kirjeldatakse juba kasutusele võetud kallutatuse vähendamise meetmeid ning loetletakse olemasolevad testitulemused ja hinnangud, näiteks auditite tulemused, eetikahinnangud, testid, õigushinnangud, energiakulu hinnangud ja keskkonnamõju hinnangud.

Tulemi kohta esitatakse väljundandmete koosseis, kas tulem on isikustatav või sisaldab eriliiki isikuandmeid või ärisaladust. Kirjeldatakse, kuidas tulemeid kasutatakse ning milliseid otsuseid tulemite põhjal tehakse. Vaadatakse üle, kas tulemi terviklus ning konfidentsiaalsus on kaitstud.

4.3 Milliseid kallutatusega seotud ohte peab analüüsima?

4.3.1 Ohustsenaariumite kategoriseerimine avaliku sektori AI-süsteemides

Avaliku sektori asutused kasutavad AI-süsteeme väga erinevates kontekstides ja erinevatel eesmärkidel. Kallutatuse riskide mõistmiseks on oluline esmalt kategoriseerida, millises funktsionaalses rollis AI-süsteem täpselt tegutseb, kuna see määrab suurel määral ära nii ohustsenaariumite laadi kui ka võimalike kahjude iseloomu. Tuleb märkida, et sama tehnoloogiline lahendus võib täita mitut funktsiooni korraga ja seega olla mitmete ohustsenaariumitega seotud. Näiteks juturobot võib nii osutada teenuseid kui ka koguda andmeid poliitikakujundamiseks.

Poliitikakujundamist ja strateegilise juhtimist toetavad süsteemid annavad lähtekoha poliitiliste otsuste või pikaajaliste strateegiate kujundamiseks, analüüsides suuri andmekogumeid või prognoosides tulevikutrende. Kallutatuse risk seisneb siin ebatäpsetes või moonutatud analüüsides, mis võivad viia vale suunitlusega poliitikaotsusteni ja mõjutada tervet ühiskonda.

Sisemisi protsesse toetavad süsteemid hõlmavad värbamist, ressursside jaotamist, töökorraldust või muud organisatsioonisisest otsustamist. Kallutatuse risk avaldub siin ebaõiglase kohtlemise või ressursside ebaõiglase jaotamisena organisatsiooni sees, mis võib kahjustada nii töötajaid kui ka organisatsiooni tõhusust.

Haldusmenetlust automatiseerivad süsteemid teevad otsuseid või aitavad otsuseid teha kodanike õiguste, kohustuste või hüvede kohta. Siia kuuluvad näiteks loa andmise süsteemid, toetuste määramise algoritmid, rikkumiste tuvastamise mehhanismid või menetluses oluliste as-

jaolude tuvastamise tööriistad. Kallutatuse risk seisneb siin selles, et süsteemi kasutusega kaasnev otsustus või muu tulem ei ole korrektne/täpne. Ebakorrektne otsus tähendab sisuliselt seda, et kahjustatakse inimese õigusi, piiratakse tema vabadusi või seatakse õigusvastaselt kohustusi.

Avalikke teenuseid pakkuvate ja kodanikega suhtlevate süsteemide alla kuuluvad vestlusrobotid (nagu Bürokratt), infosüsteemid, teenuste osutamise platvormid. Kallutatuse risk seisneb siin erineva kvaliteediga teenuse osutamises või diskrimineerivas suhtlemises, mis võib piirata kodanike võrdset juurdepääsu avalikele teenustele.

4.3.2 Kuidas ohustsenaariumide realiseerumine kahjusid põhjustab?

Kui ülalkirjeldatud ohustsenaariumid realiseeruvad, võivad tagajärjed olla mitmekesised ja tõsised. Järgnevalt käsitleme peamisi kahjukategooriaid, mis võivad AI-süsteemide kallutatusest tuleneda.

4.3.3 Kehalisi või majanduslikke kahjusid põhjustav kallutatus

Mis on probleem? AI-mudelid, mis on treenitud ajalooliste andmete peal, võivad õppida ja oma tulemite läbi taastoota neis andmetes leiduvaid statistilisi mustreid. Kui need mustrid peegeldavad ühiskondlikke erisusi, võib mudel hakata mingeid demograafilisi tunnuseid (nagu sugu, rass, vanus) seostama kindlate tulemustega. See võib viia diskrimineeriva käitumiseni mudelit rakendava AI-süsteemi poolt.

Lisaks otsestele kehalistele ja majanduslikele kahjudele on diskrimineerimine või muude isikuõiguste kahjustamine probleemne oht iseenesest, isegi kui sellega ei kaasne mõõdetavat kehalist või majanduslikku kahju. Põhiõiguste kahjustamine on halb ja õõnestab õigusriigi põhimõtteid (vt allpool).

Miks see on probleem? Sellised süsteemid võivad süstemaatiliselt langetada ebasoodsaid otsuseid mingite inimgruppide suhtes, sõltumata individuaalsetest erinevustest. See toob kaasa ebaõiglase juurdepääsu töökohtadele, laenudele, tervishoiule või õiguskaitsele. Protsesse optimeerima loodud mudelid võivad niiviisi hoopis kinnistada ja automatiseerida varasemaid erisusi, muutes kallutatusest tingitud probleemide tuvastamise ja käsitlemise raskemaks.

Mis saab valesti minna ja millised on kahjud erinevates süsteemitüüpides?

Näide. Haldusmenetlus: otsused tervishoiusüsteemides

USA haiglates laialt kasutatud algoritm suunas mustanahalisi patsiente harvemini erihoolduse programmidesse kui valgenahalisi patsiente, kelle tervislik seisund oli sama halb. Nii põhjustati ühele rahvastikugrupile tarbetuid tervislikke kannatusi. Probleem tekkis sellest, et algoritm kasutas tervisevajaduse näitajana varasemaid ravikulusid, arvestamata, et ajalooliselt on mustanahaliste patsientide ravile kulutatud vähem raha. Seega õppis mudel, et madalamad kulud tähendavad väiksemat abivajadust, mis oli vale seos. [110]. See näide illustreerib, kuidas haldusmenetlust toetav AI-süsteem (meditsiinilise abi määramine) võib ebakorrektse otsuse tulemusel otseselt kahjustada patsientide õigust võrdsele arstiabile.

Näide. Sisemised protsessid: värbamine

Amazoni loodud värbamisalgoritm hakkas süstemaatiliselt madalamaid hindeid andma naiskandidaatidele. Süsteem oli treenitud 10 aasta jooksul esitatud elulookirjelduste põhjal, millest enamik pärines meestelt. Mudel õppis ära, et „meessugu“ on edukuse ennustaja, ja hakkas karistama CV-sid, mis sisaldasid sõna „naiste“ (nt „naiste maleklubi kapten“). Sellega põhjustati ühele grupile majanduslikke kahjusid. [111]. See näide demonstreerib, kuidas organisatsiooni sisemist protsessi toetav süsteem võib luua süsteemset diskrimineerimist ja majanduslikku kahju.

Näide. Avalikud teenused: keeleline ebavõrdsus mõjutab kvaliteeti

Suured keelemudelid on treenitud valdavalt ingliskeelsete andmete peal. Seetõttu on nende võimekus ja täpsus väiksema kõnelejate arvuga keeltes süsteemselt madalam. Näiteks on **MMLU-ProX** mõõtlusalus näidanud kuni 24,3% suuruseid erinevusi keelemudelite võimekuses eri keelte vahel [112]. Probleem ei piirdu aga ainult madalama kvaliteediga (nt faktivead või ebaloomulik keelekasutus), vaid puudutab ka ohutust. Uuringud on näidanud, et mudelid käituvad mitte-ingliskeelsetes keeltes oluliselt ohtlikumalt, kuna neis keeltes on ohutuspeenhäälestuseks vähem ressurssi. Kahjud tekivad väiksemate keelekorpustega rahvastele selle kaudu, et nende AI-süsteemid on madalama kvaliteediga või ohtlikud [113]. See näitab, kuidas avalike teenuste osutamisel võib tekkida süsteemne ebavõrdsus keelelise tausta alusel.

Näide. Poliitikakujundamine ja kultuuriline mõju: keeleline ebavõrdsus mõjutab kultuuri

Spetsiifiliselt infootsingu ja küsimustele vastamise kontekstis on leitud, et mudelid eelistavad teabeallikana kõrge ressursiga keeli. See võib viia vähemuskeelsete kultuuride vaatenurkade marginaliseerimiseni ja domineerivate kultuuriruumide stereotüüpide võimendamiseni, kui mudel ignoreerib madala ressursiga keeltes olevat infot [114]. See kõik loob digitaalse lõhe, kus teenuse kvaliteet ja ohutus sõltuvad kasutaja keelest ning väiksema kandjate arvuga kultuurid muutuvad nõrgemaks. See illustreerib, kuidas AI-süsteemide kasutamine poliitikakujundamise toetamiseks või avalike teenuste osutamisel võib süstemaatiliselt moonutada kultuurilist mitmekesisust ja marginaliseerida vähemusgruppe.

4.3.4 Inimagentsuse murenemist ja vastutuse hajumist põhjustav kallutatus

Mis on oht? Oht tuleneb inimese kalduvusest ülemäära usaldada automaatsüsteemide tulemeid (*automation bias*). Eriti tugev on see efekt siis, kui tehisintellekti mudel on nn must kast, mille otsustusprotsessid ei ole läbipaistvad või nõuab nende kontrollimine inimese jaoks täiendavaid pingutusi. See toob kaasa vastutuse hajumise. Kui kahjuliku otsuse teeb algoritm, siis võib olla ebaselge, kes on juriidiliselt või moraalselt vastutav: kas arendaja, juurutaja või inimoperaator. Eesti avaliku sektori kontekstis on kahju eest igal juhul juriidiliselt vastutav avaliku sektori asutus (isegi kui kasutati kolmanda poole lahendust). Siiski jääb probleemiks vastutuse tegelik realiseerimine praktilistes olukordades, kus automaatotsuste põhjendamine ja vaidlustamine on keeruline.

Miks see on probleem? Kirjeldatud olukorras põhjustab igasugune algoritmi või AI-süsteemi viga (sh kallutatus) kahjusid. Inimest otsuste tegemisel suunav süsteem murendab erialast asjatundmust ja kriitilist mõtlemist, kuna inimene delegeerib otsustamise masinale. Avalike teenuste puhul vähendab see otsustusprotsessides oluliste inimlike tegurite, nagu empaatia, eetika ja kontekstuaalse mõistmise mõju. Süsteemid muutuvad jäigemaks ja vähem inimlikuks [115]. Olulisim on aga vastutuse kui demokraatliku ja õigusliku aluspõhimõtte murenemine. Kui vastutusahel on katkenud, võivad kahju kannatanud pooled jääda ilma õigusest ja võimalusest otsuseid vaidlustada või kompensatsiooni saada [116].

Mis saab valesti minna ja millised on kahjud erinevates süsteemitüüpides

- **Haldusmenetlused:** ametniku professionaalne kaalutlusõigus asendub jäiga algoritmiga, mis ei suuda arvestada konkreetse juhtumi eripärasid või inimlikke aspekte. Praktikas on mitmed avaliku sektori automaatotsustussüsteemid tühistatud just seetõttu, et need asendavad inimliku kaalutlusõiguse jäiga reeglistikuga või eiravad põhimõtteid nagu süütuse presumptsioon. Selline automatiseerimine tekitab avalikku vastupanu, mis võib viia kogu projekti sulgemiseni [12].
- **Poliitikakujundamine:** poliitikute ja ametnikute kriitiline analüüs asendub „andmepõhise“ otsustamisega, mis võib varjata olulistele küsimustele vastamata jätmise. Näiteks võivad avalik diskursus ja meedia kujundada pildi „oodatavast tehisintellektist“ (*Expected AI*), mis on sageli autonoomsem ja võimekam kui tegelik tehnoloogia. See kultuuriline kuvand soodustab uskumuse teket, et AI-süsteem ongi iseseisev ja eksimatu, unustades selle taga olevad inimlikud arendus- ja opereerimisprotsessid [117].
- **Sisemised protsessid:** juhtide ja personalispetsialistide oskused ning kogemus jäävad kasutamata, kui otsused delegeeritakse algoritmidele. See murendab organisatsiooni sisemist asjatundmust ja võib viia otsusteni, mis on küll tehniliselt korrektsed, kuid eiravad organisatsioonikultuurilisi või inimlikke aspekte.
- **Avalikud teenused:** teenuseandjad kaotavad kontakti kodanikega ja nende tegelike vajadustega. Süsteemid muutuvad jäigemaks ja vähem inimlikeks, vähendades otsustusprotsessides oluliste inimlike tegurite, nagu empaatia, eetika ja kontekstuaalse mõistmise mõju [115].

Lisaks mõjutab kõiki süsteemitüüpe regulatiivne ebakindlus: Euroopa Liidu AI määrus püüab inimagentsuse murenemise riski leevendada inimjärelevalve nõudega (artikkel 14). Samas leitakse analüüsides, et pelgalt kasutaja teavitamine sellest riskist ei pruugi olla piisav meede ning vastutuse jagunemine arendaja ja rakendaja vahel jääb ebaselgeks [118].

4.3.5 Organisatsiooni- ja mainekahju põhjustav kallutatus

Mis on oht? See on risk AI-süsteemi rakendava organisatsiooni vaatevinklist. Kui kallutatud või muul moel vigane süsteem põhjustab avalikku kahju, võib sellele järgnev negatiivne tähelepanu ja tagasilöök osutuda organisatsioonile endale ja selle eesmärkidele väga ebasoodsaks. Risk ei seisne mitte ainult ebaõiglases tulemuses, vaid ka otsestes negatiivsetes tagajärgedes süsteemi loojale või rakendajale, ent sõltuvalt juhtumi tüübist võib mõjutada ka teisi organisatsioone ja kanduda üle kogu riigile.

Miks see on probleem? Probleem väljendub otseses finantskahjus (ebaõnnestunud projektid, kohtukulud), avaliku usalduse kaotuses ja suurenenud regulatiivses surves. See võib muuta tu-

levaste AI-projektide elluviimise märgatavalt keerulisemaks, kuna nii avalikkus, kliendid kui ka töötajad muutuvad AI suhtes skeptilisemaks. Avaliku sektori asutuste jaoks tähendab see ebaõnnestumist oma kodanike teenimisel ja õõnestab usaldust riigi ja valitsuse vastu laiemalt.

Mis saab valesti minna ja millised on kahjud?

- **Empiiriline tõendusmaterjal:** andmekogu **RealHarm**, mis koondab näiteid AI ebaõnnestumistest, leidis, et organisatsioonide jaoks on kõige sagedasemaks kahjuliigiks just mainekahju. Uuring näitas ka, et olemasolevad kaitsepiirded ei oleks suutnud paljusid neist intsidentidest ära hoida, mis viitab lüngale AI-rakenduste kaitsesüsteemide kasutuselevõtu praktikas [119].
- **Tühistatud projektid:** avaliku sektori automaatsüsteemide ebaõnnestumisi analüüsiv uuring tõi välja 61 juhtumit, kus projektid lõpetati. Põhjuseks oli kombinatsioon tehnilistest puudujääkidest, eelarve ületamisest, tõestatud kallutatusest ja avalikust survest. Kõik need tegurid panustavad otseselt organisatsioonilisse kahjusse [12]
- **Äriline läbikukkumine:** eelnevalt mainitud Amazoni värbamistööriista juhtum on klassikaline näide ärilisest läbikukkumisest. Suur ajaline ja rahaline investeering päädis kasutuskõlbmatu tootega, millele lisandus negatiivne meediakajastus ja mainekahju, mis sundis ettevõtet projekti peatama [111].

Mis saab valesti minna ja millised on kahjud avalikus sektoris?

- **Haldusmenetlused:** kodanike õiguste rikkumine toob kaasa kohtuvaidlused, meediaskandaalid ja poliitilise surve.
- **Sisemised protsessid:** diskrimineerimine töötajate suhtes võib kaasa tuua töövaidlusi ja talentide lahkumist.
- **Avalikud teenused:** kehv teeninduskvaliteet või diskrimineerimine kahjustab asutuse mainet ja kodanike usaldust.
- **Poliitikakujundamine:** kallutatud analüüside põhjal tehtud poliitikaotsused võivad viia laialdase kriitika ja poliitilise vastutuseni.

Näide. Haldusmenetlus: avalikud teenused

Hollandi sotsiaaltoetuste pettuste riski hindamise algoritmid asetasid süstemaatiliselt ebasoodsamasse olukorda sisserändaja taustaga kodanikke, mis viis nende diskrimineerimiseni toetuste menetlemisel ning hiljem laialdase skandaalini [120] (vt ka SyRI näidet alajaotises 3.1.3). Samuti on näidatud, et ennustava politseitöö (*predictive policing*) algoritmid võivad suunata politseiressursse ebaproportsionaalselt mingitesse piirkondadesse, tuginedes ajaloolistele arreteerimisandmetele, mis juba peegeldavad varasemaid politseitöö mustreid ja võimalikku erapoolikust [121]. Mõlemad näited demonstreerivad, kuidas haldusmenetlust automatiseerivad süsteemid võivad õigusvastaselt piirata kodanike õigusi ja seada neile ebaõiglasi kohustusi, murendades rahva usaldust riigi ja selle teenuste vastu ning tekitades seega mainekahju.

4.3.6 Demokraatia, õigusriigi ja sotsiaalse ühtekuuluvuse kahjustust põhjustav kallutatus

Mis on oht? AI-süsteemide kallutatus ohustab demokraatlike ühiskondade alustalasid – kodanike usaldust riigi ja avalike institutsioonide vastu ning õigusriigi põhimõtteid. Kallutatud algoritmilised süsteemid rikuvad demokraatia põhiprintsiipe ja individuaalvabadusi, mis moodustavad õiglaste ühiskondade aluse. Sageli iseloomustab niisuguseid süsteeme läbipaistmatus nii üldiselt AI rakendamise kui ka selle täpse mõju osas. Lisaks võib inimesel puududa valikuvõimalus AI mõju osas oma elule. Samaaegselt võib AI-süsteemide kallutatus normaliseerida, suurendada ja võimendada eksisteerivat sotsiaalset ebavõrdsust, marginaliseerimist ja eelarvamusi. Sotsiaalpoliitika (eluaseme, tervishoiu), haridus-, kindlustus-, finants- ning kaubanduse valdkonnas juba eksisteeriv ebaõiglus teeb kodanikud ümberkorralduste suhtes tundlikuks. See loob ühiskonnas lõhesid ja takistab kõigi kodanike võrdset osalemist ühiskondlikus elus.

Kuidas erinevad süsteemitüübid seda ohtu loovad?

- **Haldusmenetlused:** ebaõiglased automaatotsused kodanike õiguste, kohustuste või hüvede kohta õõnestavad tunnetuslikult õigusriigi põhimõtteid ja võrdsuse printsiipi. Süsteemne diskrimineerimine hüvede jagamisel või kohustuste määramisel süvendab samuti olemasolevaid ebavõrdsusi.
- **Poliitikakujundamine:** kallutatud andmeanalüüs või prognoosid võivad viia diskrimineeriva poliitikani, mis kahjustab mingeid gruppe süsteemselt, nt eirab nende vajadusi või võimendab nende marginaliseerimist.
- **Avalikud teenused:** erineva kvaliteediga teenindus eri rühmade jaoks kahjustab võrdsuse põhimõtet ja kodanike usaldust. See võib süvendada (digitaalset) lõhet ja sotsiaalset kihistumist.
- **Sisemised protsessid:** diskrimineerimine värbamisel või ressursside jagamisel avaliku sektori sees võib vähendada mitmekesisust ja kvaliteeti avalikes teenustes, mõjutades kaudselt tervet ühiskonda.

Miks see on probleem? Usalduse kaotus tehnoloogia ja institutsioonide vastu õõnestab demokraatlikke protsesse ja võib viia legitiimsuse kriisini. AI-süsteemide kallutatus seab ohtu põhilised õigusriiklikud põhimõtted: õiguse võrdsetele võimalustele ja erapooletusele sotsiaalmajanduslikus, soolises, ealises, etnilises, religioosse või seksuaalse eelistuse tähenduses; õiguse õiglasele kohtusüsteemile ja haldussüsteemile üldises tähenduses ning õiguse privaatsusele ja isikliku vara kaitsele. Näiteks süütuse presumptsioon keeratakse pea peale, kui kõiki sotsiaaltoetuste taotlejaid kontrollitakse hüvele kvalifitseerumise kriteeriumite asemel võimalike rikkumiste suhtes. Kallutatuse süvenemine ja automatiseerumine võib õõnestada sotsiaalset ühtekuuluvust ja takistada ühiskonnal kaasavuse abil realiseerida oma täielikku potentsiaali. Kui osad grupid jäävad süsteemselt ilma võrdsetest võimalustest hariduses, tööturul, tervishoius või muudes eluvaldkondades, siis ei saa ühiskond ära kasutada kõigi oma liikmete andeid ja panust. See toob kaasa mitte ainult individuaalse kahju, vaid ka kollektiivse kaotuse. Ühiskond jääb ilma innovatsioonist, majanduskasvust ja sotsiaalsest arengust, mida võiks tuua kaasa kõigi kodanike täielik ja võrdne osalemine. Kui põhiseaduslikud õigused ja demokraatlikud väärtused muutuvad automaatsüsteemide käes sisutühjaks, kahaneb riigi legitiimsus ja kodanike usaldus, mis omakorda süvendab sotsiaalset lõhestumist ja takistab ühiskonna arengut.

4.4 Kuidas hinnata kallutatuse riske?

4.4.1 Kuidas hinnata kallutatusega seotud ohtude tõsidust?

Iga organisatsioon peab riskihalduse jaoks paika panema enda riskinormi ehk valmisoleku tegeleda ohusündmuse tagajärgedega. Kui mõne ohu tagajärjed hinnatakse nii tõsiseks, et organisatsioon nendega tegeleda ei soovi, tuleb tegeleda ennetusega. Alajaotises 4.3 tõime näited algoritmide ja AI-süsteemide kallutatuse võimalikest tagajärgedest (varalised kahjud, kahjud tervisele, kahjud riigi ja ühiskonna tasemel). On mitmeid aspekte, mis tõstavad AI-süsteemi riske märgatavalt ning muudavad need haldaja tähelepanu väärivaks.

AI-süsteemi tulem mõjutab elu, tervist või varalist seisu. Kui AI-süsteemi tulem mõjutab otseselt või kaudselt kellelegi raha, töö, abi, ravi andmist või mitteandmist, füüsilist või vaimset vägivalda teise isiku vastu, või tema õiguste põhjendamatut piiramist, on süsteemi kallutatuse tagajärjed tõsised. Isiku autonoomia, privaatsuse või muude vabaduste kahjustamine on sõltuvalt ühiskonnakorraldusest vähemalt sama tähtis, kuid konkreetseid varalisi või füüsilisi kahjusid oskavad riskide analüüsijad tihti selgemalt käsitleda. Seega tasub ohtude kirjeldamisel kaaluda, kas ka nt andmelekked võivad viia hilisemate otseste varaliste või kehaliste kahjudeni, isegi kui kahju ei ole veel tõestatavalt saadud.

AI-süsteem on vahetu otsustaja ja/või elluviija. Kui AI-süsteemi tulem vormistatakse automaatselt otsuseks, on kallutatuse tagajärjed tõsised. AI-põhine automaatse karistusmenetluste süsteem (näiteks parkimisreeglite eiramise või kiirusepiirangu ületamise puhul) on oma mõjus kohene ja konkreetne. Kui selline AI-süsteem on kallutatud, võib see ilma korraliku seire ja järelevalveta viia mõõdetavate kahjudeni, eriti kui isik oma teguviisis tegelikult ei eksinudki. Märgime ära, et ka siis kui süsteemi on lisatud inimene otsuseid kontrollima, võib tal aja jooksul välja kujuneda sõltuvus masina arvamusest ning tema järelevalve efektiivsus kahaneb.

AI-süsteem muudab käitumismustreid, tõekspidamisi või väärtushinnanguid. Kui AI-süsteem mõjutab suure hulga inimeste käitumist pikema aja jooksul, on kallutatuse tagajärjed tõsised. Näiteks hariduslikud, psühholoogilise nõustamise või meelelahutuse otstarbel loodud (kuid kallutatud) AI-süsteemid võivad inimeste käitumist mõjutada pikema perioodi jooksul. Kui nad ka täidavad oma esmase ülesande (oskuste edendamine, keerulise suhteprobleemide lahendamine või meelelahutus), siis võivad nad selle kõrval luua ka kahjulikke käitumismustreid (sõltuvust masinast kõigis tegevustes, soovimatust suhelda teiste inimestega). Samad ohud on olemas ka siis, kui neid teenuseid annavad inimesed, kuid inimliku suunamise maht ja seostuvad kahjud on piiratud, samas kui AI-süsteem võib samade juhendite abil mõjutada sadu miljoneid inimesi.

4.4.2 Kuidas seada lävendeid?

Algoritmiliste ja AI-süsteemide kallutatuse riskihalduse metoodika on oma olemuselt mitmesammuline. Mida tõsisem on oht, seda põhjalikumalt selle leevendamise meetmeid kaalutakse.

Uute protsesside, tööriistade ja süsteemide evitamisel ei õnnestu üldjuhul saavutada olukorda, kus riske üldse ei ole. Jääkriskid jäävad pea alati ning nendega leppimine on riskikäsitluse tavapärane osa. Sellises olukorras peab süsteemi looja aga olema valmis riske omavahel võrdlema ja hindama, millistega tuleb vältimatult tegeleda. Selles alajaotises pakume välja kaks mõõdikut, kuid need on vaid näited. Süsteemi looja ja riskide käsitleja võivad vabalt kasutada ka teisi, endale sobivamaid mõõdikuid.

Kahjusaajate hulk. Kõige lihtsam viis ohu mõõtmiseks on konkreetses ajaühikus konkreetseid kahjusid saavate isikute arv või protsent. Kahjude mõõtmise näited on järgmised.

- Süsteemi eluea jooksul sureb süsteemi kallutatusega seoses üks inimene[1].
- Aastas saab süsteemi kallutatusega seoses üks inimene tervisekahjustusi.
- Aastas põhjustab süsteemi kallutatus kümne elujõulise organisatsiooni tegevuse lõppemise[2].
- Ühes kuus saab süsteemi kallutatusega seoses põhjendamatut rahalist kahju $0,1\%$ süsteemi kasutajatest.
- Ühes aastas kaotab töö $0,5\%$ rahvastikust.

Näide. Koolitusettevõte maksis vanuse järgi diskrimineeriva algoritmi tõttu hüvitist

Ameerika Ühendriikide föderaalne agentuur EEOC (Equal Employment Opportunity Commission, Võrdsete töövõimaluste komisjon) kaebas kohtusse keeletunde andva ettevõtte iTutorGroup, sest nende värbamise infosüsteem lükkas automaatselt tagasi kõigi üle 55-aastaste naiste ja üle 60-aastaste meeste taotlused. Ettevõte pidi tasuma 365 000 dollari suuruse hüvitise, rakendama diskrimineerimisvastase poliitika ning korraldama oma värbajatele koolituse [122].

Aeg esimeste kahjudeni. Lisaks kahjusaajate lävendile võib olla otstarbekas hinnata, kui kaua läheb aega, et kahjusündmus tekiks. See on eriti oluline ohtude puhul, kus algoritmi või tehisintellekti kallutatus mõjutab inimesi üle pika aja. Pikemat perspektiivi peab vaatama ka olukorras, kus riskianalüüs tehakse enne süsteemi ehitamist ning ehitamine ise võtab mitmeid aastaid aega, näiteks:

- kallutatusega seotud ohud võivad avalduda 10 aasta jooksul hindamisest või
- kallutatusega seotud ohud võivad avalduda 7 aasta jooksul pärast süsteemi kasutuselevõttu.

4.5 Kallutatuse ohtude riskide käsitlus

4.5.1 Võimalikud lõpptulemused

Algoritmilise või AI-süsteemi kallutatuse riski hindamisel ja käsitlemisel on kolm tõenäolist tulemit.

1. Süsteem võetakse kasutusele muudatusteta ja kõik jääkriskid (vajaduse korral lisameetmete rakendamise järel) aktsepteeritakse.
2. Süsteem loetakse nii kallutatuks, et selle arendus või kasutus peatatakse, sest riskikäsitlus ei ole võimalik või ei ole see majanduslikult efektiivne. Kuigi see on kõige kindlam viis kallutatuse ohte vältida, jäävad saamata ka algoritmilisest või AI-süsteemist oodatavad tulud.
3. Süsteem võetakse kasutusele tingimisi, kui õnnestub juurutada sobivad organisatsioonilised või tehnilised lisakaitsemeetmed. See on tõenäoliselt kõige levinum tulem.

4.5.2 Kallutatuse leevendamise keerukus ja kompromissid

Kuigi tehnilisi meetodeid kallutatuse leevendamiseks on mitmeid, on oluline mõista, et tegemist ei ole lihtsa tehnilise parandusega. Kallutatuse leevendamine on keeruline sotsiotehniline prob-

[1] Eeldame siinkohal AI-süsteemi tsiviilkasutust.
[2] Märgime ära, et on täiesti kujuteldav ja võib olla ka soovitav luua AI-süsteem mingitele kindlatele tingimustele (nt kuritegelik tegevus) vastavate organisatsioonide tegevuse lõpetamiseks.

leem, millel puudub ühtne ja lõplik lahendus. Iga sekkumisega kaasnevad omad kompromissid ja filosoofilised valikud.

Olukorrad, kus õigusraamistik loob lihtsalt tõlgendatava ootuse, on tõepoolest lihtsad. Näiteks kui mõni avalik teenus soosib selgelt ühest soost taotlejaid, on kallutatus ilmne. Seadusega keelatud kallutatust peab süsteemides vältima.

Siiski, võib ette tulla ka keerukamaid juhtumeid. Esimene ja kõige fundamentaalsem probleem on see, et õiglusel puudub ühtne, universaalselt aktsepteeritud definitsioon. Google'i What-If tööriista tutvustav artikkel illustreerib seda, tuues näite viiest eksperdist, kes kõik defineerivad soolist õiglust laenuotsuste puhul erinevalt [123].

- **Grupi ignoreerimine (*group unaware*):** sugu ei tohi üldse arvesse võtta, isegi kui see tähendab, et ükski naine ei saa laenu.
- **Erinevad lävendid (*group thresholds*):** naiste ja meeste jaoks tuleks seada erinevad laenukindluse lävendid, et kompenseerida ajaloolist ebasoodsat olukorda andmetes.
- **Demograafiline pariteet (*demographic parity*):** heakskiidetud laenusaajate sooline jaotus peab vastama taotlejate soolisele jaotusele.
- **Võrdsed võimalused (*equal opportunity*):** kvalifitseeruvatel meestel ja naistel peab olema võrdne tõenäosus saada laenutaotlusele heakskiit.
- **Võrdne täpsus (*equal accuracy*):** mudeli ennustuste täpsus (nii positiivsete kui negatiivsete otsuste puhul) peab olema mõlema soo lõikes võrdne.

Kõik need definitsioonid on omavahel vastuolus. Iga valik on kompromiss, mis nõuab otsust vastavalt AI-süsteemi täpsemale ülesehitusele ja otstarbele, organisatsiooni riskitaluvusele, ärinõuetele ja seadusmaastikule.

Erisused tulemites ei võrdu alati diskrimineerimisega. Eeldades, et igasugune statistiline erinevus gruppide tulemustes on kindlasti ebaõiglase diskrimineerimise tulemus, on oht langeda kehva teadmisteooria lõksu. On võimalik, et mudel on tuvastanud reaalseid, mitte-diskrimineerivaid alusmuutujaid, mis korreleeruvad demograafiliste gruppidega. Sellisel juhul võib parandamine viia uue ebaõigluseni, näiteks eelistades vähem kvalifitseeritud kandidaate kvoodi täitmiseks. Seetõttu on kriitilise tähtsusega enne leevendusmeetmete rakendamist põhjalikult analüüsida erinevuste tegelikke põhjuseid.

Õigluse ja täpsuse kompromiss. Enamik kallutatuse leevendamise tehnikaid toimivad mudeli optimeerimisprotsessi piirangutena. Piirangutega süsteem on peaaegu alati ebaoptimaalne oma algse, piiranguteta ülesande täitmisel. Praktikas tähendab see sageli, et õigluse suurendamine toimub mudeli üldise täpsuse või sooritusvõime arvelt. Nagu näitas Fairlearni näide (vt kirjeldust alajaotises 5.4), vähendas otsustuslävede kalibreerimine küll soolist ebavõrdsust, kuid tõi kaasa ka väikese languse mudeli üldises täpsuses. See on kompromiss, millega organisatsioonid peavad süsteemi arendades arvestama.

4.5.3 Millised kallutatuse mõjude leevendamise meetmed on kuluefektiivsed elutsükli erinevatel etappidel?

Alajaotises 2.3 kirjeldasime masinõppemudeli loomise ja algoritmi või mudeli rakenduses evitamise etappe. Kallutatuse vältimine on igas elutsükli etapis erineva hinnaga [124, 125].

Süsteemi kavandamine on parim aeg kallutatuse ennetamiseks. Just siis, kui süsteemi vi-

sioon ja eesmärgid on sõnastatud, on parim aeg mõelda, kelle jaoks süsteemi tehakse ning kelle andmete peal see peab korrektselt töötama. Soovitame järgmiseid tegevusi.

1. Määrake grupid, kelle andmetel algoritm või masinõppemudel peab korrektselt toimima.
   - Kas see peab käituma korrektselt kõige maailma rahvaste puhul, mõne riigi kodanike puhul või väiksema inimhulgaga?
   - Millised on oodatavad vanusegruppide jaotused – kui noorte ja kui vanade inimestega peab süsteem toime tulema?
   - Millist isikute soolist jaotus oodatakse?
   - Millistes loomulikes keeltes peab süsteem toime tulema?
2. Pange paika reeglid, kuidas käituda, kui on vaja teha valikuid ja kompromisse olukorras, kus ideaalset andmestikku, algoritmi või mudelit pole saadaval, tuginedes alajaotisele 4.5.2.
3. Lisage see info süsteemi hanke- või arendusnõuetesse tähtsa nõudena.

Algoritmi või masinõppemudeli loomine on kõige parem (aga mitte tingimata kõige odavam) hetk kallutatuse vältimiseks. Kui masinõppemudeli loomine on AI-süsteemi tellija kontrolli all (mudel luuakse vastavalt tema nõuetele), on võimalusi kallutatuse riskide leevendamiseks kõige rohkem. Neist anname ülevaate alajaotises 4.5.4.

Juba ehitatud või hangitud süsteemi puhul on kallutatusega tegelemine kallim ning võibolla isegi võimatu. Kui tervikuna hangitud süsteemis realiseeruvad kallutatuse riskid vastuvõetamatus mahus, tuleb süsteemi kas muuta või see välja vahetada. Kui AI-süsteemis on algoritmi või masinõppemudeli väljavahetamine võimalik (kasutatud on standardseid liideseid), siis õnnestub ehk arendada või hankida parem mudel, mis töötab sihtvalimil kvaliteetsemalt. Vahel õnnestub süsteemi ümber ehitada või lisada täiendavaid tehnilise või organisatsioonilisi meetmeid, millega õnnestub vältida terve toote, algoritmi või mudeli väljavahetamist. Võimalikke lahendusi kirjeldab alajaotis 4.5.5.

Kui süsteemi käitumist muuta ei saa, tuleb selle kasutamine lõpetada ja vajaduse korral hankida uus süsteem. Vastavaid juhtumeid käsitleme alajaotises 4.5.6.

4.5.4 Kuidas leevendada kallutatust masinõppemudeli treenimisel?

IT-tööstuse praegune majanduslik reaalsus on selline, et masinõppemudelite isearendamine on kallis ning seega tuleb tihti hakkama saada mudelitega, mida on treeninud teised. Aga kui AI-süsteemi ehitajal on kontroll selle üle, kuidas masinõppemudel luuakse, siis on võimalik sellele nõudeid seada nii andmete kogumise ja ettevalmistuse etapis kui mudeli treenimise etapis.

Andmete kogumise ja ettevalmistuse etapp

Kallutatuse allikaks võib olla:

- mingite gruppide ala- või ülesesindatus mudelis (valimi representatiivne kallutatus) [124],
- mõõtmisvead või mürarikkad andmed, mis tugevdavad loodavas mudelis ajaloolist ja/või representatiivset kallutatust [126].

Võimalikud kallutatuse leevendamise meetmed on järgnevad.

- Kui treeningandmete kogumine on süsteemi looja kontrollida (nt on projekti osa), siis soovitame korraldada täiendava andmekogumise andmete mitmekesisuse ja esinduslikkuse tagamiseks (alaesindatud gruppide jaoks andmeid juurde leida) või siis üleesindatud gruppide

andmed välja jätta (samal ajal kvaliteedimõõdikuid jälgides)[3] [124].

- Kui alaesindatud gruppide puhul on teada andmete statistiline jaotus (või see on võimalik analüütiliselt leida), soovitame kaaluda alaesindatud gruppide andmete sünteetiliselt juurde genereerimist.
- Soovitame kaaluda kallutatuse tuvastamise ja andmeauditite tööriistu, eelistades tööriistu mida jooksvalt täiendatakse ja uuendatakse [126].
- Kui andmeid ei ole võimalik juurde saada, siis soovitame rakendada mudeli treenimisel tasakaalustamise tehnikaid [126].

Mudeli treenimise etapp

Kallutatuse allikaks võib olla:

- ebaõiglus algoritmides treeningandmete kallutatuse või sobimatute optimeerimiskriteeriumide tõttu [127],
- eelarvamuslikud seosed treeningandmetes (nt nimede sidumine stereotüüpidega) [128],
- üleõppimine dominantsete gruppide andmete põhjal [127],
- liigsete isikuandmete ja tundlike andmete sisestamine ilma vajaduseta või õigusliku aluseta [128].

Võimalikud kallutatuse leevendamise meetmed on järgnevad.

- Kui on teada, et mõned tunnused ei tõsta masinõppemudeli kvaliteeti, aga võimendavad selle kallutatust, soovitame need tunnused treeningandmeid valides välja jätta.
- Soovitame kasutada väiksemaid ja paremini kontrollitavaid mudeleid, mille käitumises on vähem üllatusi [128].
- Soovitame treeningandmete ettevalmistamisel kvaliteeti kontrollida, sh eemaldada duplikaate ja treenimisandmestikku filtreerida [128].
- Kui loodava süsteemi funktsionaalsuse jaoks on olemas õigluspõhised algoritmid, soovitame kaaluda nende kasutamist [127].
- Soovitame valideerida mudeli väljundeid regulaarselt eri gruppidest võetud sisenditega, eriti kui mudelist võetakse kasutusele uus versioon [127].
- Soovitame teostada mõjuhindamisi kallutatuse kohta kogu arenduse jooksul [127].
- Kui treenimisel või peenhäälestamisel kasutatakse isikustatavaid andmeid, soovitame jälgida minimeerimise ja lõimitud andmekaitse põhimõtteidja rakendada täiendavaid lisakaitsemeetmeid nende andmete kaitseks.

Näide. Kallutatud valimiga masinõppemudelid on ka ebaefektiivsed

Mitmetes teadustöödes on üritatud korrata varasemaid eksperimente masinõppemudelite treenimisel, kuid see pole õnnestunud – kordusuuringu mudelid ei ole sama edukad kui algse uuringu mudelid. Põhjuseks on tihti olnud valimi kallutatus – efekt ilmneb vaid konkreetsete treeningandmete peal ning suuremate ja esinduslikumate andmestike peal tulemust korrata ei õnnestu. See tähendab, et kallutatud mudelid võivad ka lihtsalt mitte toimida. Kallutatud andmestikega treenitud mudelite ebaefektiivsus on levinud näiteks meditsiinis autismi [129], radioloogilisi pilte [130] ja sünnitusabi [131] käsitlevates teadustöödes.

[3] Enne selle (ja teiste) leevendusmeetmete kasutamist soovitame lugeda alajaotist 4.5.2. Lisaks meenutame, et andmete sünteesi kasutamine masinõppemudelite treenimisel võib võimendada õpetamislikku kallutatust.

4.5.5 Kuidas saab kallutatust vähendada algoritmi või mudeli liidestamisel või evituses?

Kui AI-süsteemi elutsüklis jõutakse algoritmi või masinõppemudeli liidestamiseni, on kallutatus algoritmis või mudelis juba olemas. Oluline on mõista, et turule toodud toimiva AI-süsteemi teenustajal on märkimisväärselt rohkem õiguslikke kohustusi kui arenduse käigus (vt alajaotist 3). Tasub ka arvestada, et hilises faasis võib kallutatuse leevendamine olla kallim kui süsteemi loomise alguses. Loetleme mõned võimalikud leevendusmeetmed

Masinõppemudeli AI-süsteemi tarbeks peenhäälestamise etapp

Kallutatuse allikaks võib olla:

- statistilise kallutatuse tugevnemine peenhäälestamise käigus,
- tundlike andmete kasutamine peenhäälestamisel ilma õigusliku aluseta [128],
- stiimulõpe või automaatsed tagasisidetsüklid, mis võimendavad masinõppemudeli kallutatust [128].

Võimalikud kallutatuse leevendamise meetmed on järgnevad.

- Olukorras kui baasmudelit muuta ei saa, aga seda on võimalik peenhäälestada, siis soovitame peenhäälestamise lisaeesmärgiks seada kallutatuse vähendamine.
- Soovitame korraldada peenhäälestamise protsessi läbipaistvalt, dokumenteerides kallutatuse vähendamise suunas tehtud muudatused.
- Sarnaselt treenimisega, soovitame ka peenhäälestamisel jälgida andmete minimeerimise ning lõimitud andmekaitse põhimõtteid ning rakendada täiendavaid kaitsemeetmeid [128].

Mudeli evitamise ehk mudeli AI-süsteemi integreerimise etapp

Kallutatus võib tekkida järgnevalt.

- Kõrge mõjuga kontekstides võivad diskrimineerivad või hallutsineeritud vastused viia AI-süsteemi järjepideva kallutatuseni [124, 128].
- Mudeli otsuste seletamatus muudab kallutatuse allikate tuvastamise keeruliseks ning takistab nende korrigeerimist [132].
- Jagatud kontekstis (nt vestluse kontekstiaknas) sisalduv kallutatud või profileeriv info võib mõjutada mudeli tulemeid.
- Kui puuduvad kaitsemeetmed promptisüstide vastu,on pahatahtlikel kasutajatel võimalik anda mudelile kallutatud juhiseid [128].

Võimalikud kallutatuse leevendamise meetmed on järgnevad.

- Kui mudelit parandada ei saa, siis soovitame kaaluda algoritmi või masinõppemudeli välja vahetamist vähem kallutatud variandi vastu.
- Sõltumatult mudeli omadustest, soovitame lisada inimese protsessis lõplikuks otsustajaks.
- Süsteemis tuleb tugevdada inimlikku järelevalvet [127].
- Soovitame koolitada otsustajaid ja järelevalvet mõistma AI-süsteemi tulemi kontrolli ja seletavust.
- Kui peaks juhtuma, et lõplikku otsust ei tee inimene, tuleks süsteemi sisse ehitada tulemite seletavuse ja kvaliteedi seire. Selleks soovitame lisada võimalus nõuda tulemi läbivaatust, mis edastab otsuse kontrolliks inimesele [128].

- Süsteemis tuleb kasutusele võtta vastutusskeemid ja läbipaistvusraportid [127].
- Soovitame käsitleda kallutatust ka masinõppemudeli integreerimisel andmestikega, rakendades kallutatust vähendavaid ja privaatsust säilitavad arhitektuure (nt RAG koos kaitsepiiretega).
- Regulaarselt tuleb auditeerida AI-arhitektuuri ning kontrollida seletavuse ja otsuste kvaliteeti [128].

AI-süsteemi käitamise ja seire etapp

Võimalikud kallutatuse allikad on järgnevad.

- Mudelite kallutatus triivib ajas, sest ühiskonna ootused muutuvad [125, 128].
- Masinõppemudeli tulem avaldab ekslikult isikustatavaid andmeid ning see võimaldab inimesel teha kallutatud otsuse, mida masin poleks teinud.

Võimalikud kallutatuse leevendamise meetmed on järgnevad.

- Soovitame korraldada pidev hindamine ja reaalajas järelevalve mudeli tulemite kvaliteedi üle [125].
- Regulaarselt tuleb koguda sidusrühmadelt ja kasutajatelt tagasisidet süsteemi käitumise kohta [128].

4.5.6 Millises olukorras tuleb süsteemi kasutamine peatada?

Kui riskikontrolli käigus selgub, et algoritmi või masinõppemudelit kasutav süsteem ei täida kehtivate seaduste või määruste nõudeid, ei saa seda kasutusele võtta. Kui kasutamist keelav seadus võetakse vastu süsteemi käitamise ajal, tuleb kasutamine lõpetada. Kui seadus kehtestab täiendavad tingimused, millele süsteem ei vasta, tuleb süsteemi kasutamine peatada kuni vastavuse tagamiseni. Samasuguse keelava otsuse võib mingitel juhtudel langetada ka näiteks eetikakomitee, kelle käsitlusalasse süsteem kuulub.

Eristada tuleb juhtu, kui muidu õiguspärane süsteem annab õiguseid rikkuva väljundi. Sellisel juhul ei ole põhjust süsteemi kasutamist kohe lõpetada, vaid tuleb süsteem parandada, näiteks siin juhendis antud soovitusi järgides. See on ka näide olukorrast, miks ei peaks kindlasti üritama süsteemi riske nullilähedaseks viia–see ei tarvitse olla realistlik ja nii seatud lävend muutab uute teenuste loomise väga keeruliseks.

Riskikontrolli käigus võivad mõned riskid jääda tuvastamata ja teiste riskide käsitlus ei leevenda neid täielikud. Ka riski leevendamine ei vii riski nulli. Allesjäävat riski kutsutakse jääkriskiks. Kui AI-süsteemi jääkriskid ületavad organisatsiooni riskinormi, on ratsionaalne süsteemi kasutamine lõpetada või seda üldse mitte kasutusele võtta.

Jääkrisk võib ületada riskinormi näiteks kui:

- ei ole võimalik välistada, et kallutatud AI-süsteem põhjustab isikule kehavigastusi või surma;
- ei ole võimalik välistada, et AI-süsteemi kallutatus suunab isiku põhjustama endale või teisele inimesele kehavigastusi või surma;
- süsteemi rakendamisel tekkivad rahalised kahjud ületavad organisatsiooni võimalused neid katta.

5 Tööriistad kallutatusega tegelejale

5.1 Kas mul on tegemist musta kasti või valge kasti süsteemiga?

AI-süsteemide puhul räägitakse sageli musta kasti ja valge kasti süsteemidest. Must kast on süsteem, mille sisemine toimimine on läbipaistmatu, samas kui valge kasti puhul on see läbipaistev. Praktikas ei ole see eristus aga nii selgepiiriline. Kasulikum on mõelda AI-süsteemi ja mudeli avatusest kui spektrist, mis ulatub kõige suletumast kõige avatumani.

Selle spektri defineerivad mitu peamist tegurit: kui palju teavet on avaldatud mudeli loomise kohta (selle andmed ja treeningprotsess), milline on juurdepääs mudelile endale ja kuidas seda rakendatakse. Samas lisab läbipaistmatust ka mudelite endi sisemine keerukus. Isegi süsteem, mis on oma ehituselt täielikult läbipaistev, võib olla funktsionaalselt must kast, kui me ei suuda selle otsuseid tõlgendada.

5.1.1 AI-süsteemi avatuse spekter

Avatuse dimensioonid. Loetleme peamised dimensioonid, mille kaudu süsteemi avatust hinnata.

- **Treeningandmed.** See on mudeli teadmiste, võimete ja kallutatuse vundament.
  - Kõige suletum (läbipaistmatu): andmete kohta info puudub.
  - Osaliselt läbipaistev: peamised andmeallikad (nt Common Crawl, Vikipeedia) on avalikustatud, kuid lõplikku, töödeldud andmestikku ja segamissuhteid ei jagata.
  - Kõige avatum (läbipaistev): täielik ja eeltöödeldud treeningandmestik on masintöödeldaval kujul kättesaadav.
- **Andmete töötlemine ja treeningsegu.** See, kuidas andmeid filtreeritakse ja segatakse, mõjutab mudeli võimekust sama palju kui andmed ise.
  - Kõige suletum: protsessi kohta detaile ei avaldata.
  - Osaliselt läbipaistev: metoodikat kirjeldatakse üldiselt tehnilises raportis (nt „kasutasime agressiivset deduplitseerimist“).
  - Kõige avatum: avaldatakse täielik lähtekood andmete puhastamiseks, filtreerimiseks ja segamiseks.
- **Mudeli arhitektuur ja kaalud.** See määrab, kas mudelit saab ise kasutada ja uurida.
  - Kõige suletum: arhitektuuri ja parameetrite arvu kohta info puudub.
  - Osaliselt läbipaistev (ainult arhitektuur): arhitektuuri on kirjeldatud, kuid treenitud parameetrid (kaalud) on suletud. Kasutaja teab, mis see on, aga ei saa seda ise käitada.
  - Kõige avatum (avatud kaaludega): mudeli kaalud on avalikult allalaetavad. Üldjuhul mõeldakse seda, kui mudelit nimetatakse avatud lähtekoodiga AI-mudeliks.
- **Treeningu detailid ja „retsept“.** See võimaldab teistel teadlastel tulemusi korrata.
  - Kõige suletum: info puudub.
  - Osaliselt läbipaistev: peamised hüperparameetrid (nt õpisamm) ja riistvara info on avaldatud.

  - Kõige avatum (avatud treeningretsept): avaldatakse täielik treeningkood, konfiguratsioonifailid ja logid.
- **Juurdepääsuviis ja kontroll.** See määrab, kui palju on arendajal kontrolli mudeli käitumise üle.
  - Kõige suletum (tootesse integreeritud): AI on vaid funktsionaalsus mõnes rakenduses (nt pilditöötlusprogrammis).
  - Rakendusliides ehk API-põhine juurdepääs: kommertsmudelite standard. Arendaja saadab päringu ja saab vastuse. Rakendusliidesed ise on samuti spektril: alates lihtsast päringtulem liidesest kuni täpsema kontrollini (nt juurdepääs logit-skooridele).
  - Kõige avatum (otsene juurdepääs): Kasutaja laeb alla mudeli kaalud ja käitab seda omal riistvaral, omades täielikku kontrolli.
- **Litsents.** See määrab, mida on mudeliga juriidiliselt lubatud teha.
  - Kõige suletum: kasutamine on piiratud teenusetingimustega; mudelit ei tohi (ei saa) alla laadida ega muuta.
  - Piirangutega avatud litsents: litsents lubab mõndasid kasutusviise (nt teadustöö), kuid seab piiranguid kommertskasutusele või nõuab muudatuste avalikustamist.
  - Kõige avatum (lubav litsents, nt Apache 2.0, MIT): litsents lubab mudelit kasutada, muuta ja levitada peaaegu ilma piiranguteta, kaasa arvatud kommertseesmärkidel.

Tabel 2 ilmestab seda spektrit, võrreldes mitut tuntud AI-mudelite tootjat ja teenustajat. AI-süsteemi või mudeli avatus ei ole ühene omadus, vaid arendajate ja teenustajate tehtud sõltumatute valikute kogum. Märgime, et mudelite avatust on kirjeldatud juhendi koostamise hetkel ning see võib ajas muutuda. Tabelist joonistuvad välja mitmed eristuvad strateegiad.

Suletud rakendusliideste pakkujad (OpenAI, Google Gemini, Anthropic). Nende mudelid on oma ehituselt mustad kastid, kuid nad pakuvad arendajatele võimsaid rakendusliideseid. Nende ärimudeliks on tipptasemel teenus, mida reguleerivad teenusetingimused.

Avatud kaaludega pakkujad (Meta, Mistral, Qwen, DeepSeek, Google Gemma). See on kõige levinum avatud lähtekoodiga AI-mudel. Andmed ja treeningprotsessid jäävad seejuures tihti omanduslikuks, mis teeb nad teadusliku korratavuse seisukohast osaliselt läbipaistmatuks.

Litsentside mitmekesisus. Isegi avatud mudelite seas on litsentsides suured erinevused. Kui Allen AI kasutab lubavat Apache 2.0 litsentsi, siis Meta (Llama) ja Google (Gemma) kasutavad oma eralitsentse, mis võivad seada piiranguid näiteks väga suurtele ettevõtetele. Cohere'i litsents on aga mitteäriline, piirates otsest kommertskasutust.

Ökosüsteemi-põhine avatus (NVIDIA). NVIDIA Nemotron seeria mudelid on küll avatud kaaludega, kuid nende litsents on strateegiliselt piiratud: see lubab vaba katsetamist, kuid äriliseks kasutamiseks tootmises nõuab see liitumist NVIDIA tasulise AI Enterprise platvormiga. See on proovi-enne-ostmist-strateegia, mille eesmärk on siduda kasutajad oma riistvara ja tarkvara ökosüsteemidega.

Tootesse integreeritud AI (Apple, Midjourney). Need süsteemid ei ole mõeldud arendaja tööriistadeks, vaid on integreeritud lõppkasutajatoodetesse (nt operatsioonisüsteemi või vestlusrakendusse). Kasutaja suhtleb AI-ga läbi toote, mitte otse.

Täieliku läbipaistvuse mudel (Allen AI). OLMo seeria on erandlik, kuna selle eesmärk on teaduslik korratavus. Avaldatud on kõik: andmed, kood ja mudeli parameetrid (kaalud).

Tabel 2. Levinud suurte keelemudelite avatus

Mudel / Arendaja	Treening-andmed	Treening-andmete eeltöötlus	Arhitektuur ja kaalud	Treeningu detailid ("retsept")	Juurdepääs ja kontroll	Litsents
OpenAI (Tippmudelid)	Suletud	Suletud	Osaliselt avatud	Suletud	API	Omanduslik
Google (Gemini seeria)	Suletud	Suletud	Osaliselt avatud	Osaliselt avatud	API	Omanduslik
Anthropic (Claude seeria)	Suletud	Suletud	Osaliselt avatud	Osaliselt avatud	API	Omanduslik
Meta (Llama seeria)	Osaliselt avatud	Suletud	**Avatud**	Osaliselt avatud	**Kaalud** ja partnerite API-d	Llama litsents (piirangutega)
Mistral (Avatud mudelid)	Osaliselt avatud	Suletud	**Avatud**	Osaliselt avatud	**Kaalud** ja API	**Apache 2.0 (lubav)**
Alibaba (Qwen3 seeria)	Osaliselt avatud	Suletud	**Avatud**	Osaliselt avatud	**Kaalud** ja partnerite API-d	**Apache 2.0 (lubav)**
DeepSeek (DeepSeek-R1)	Osaliselt avatud	Osaliselt avatud	**Avatud**	Osaliselt avatud	**Kaalud** ja API	**MIT (lubav)**
Google (Gemma seeria)	Osaliselt avatud	Suletud	**Avatud**	Osaliselt avatud	**Kaalud**	Gemma litsents (piirangutega)
NVIDIA (Nemotron seeria)	Osaliselt avatud	Suletud	**Avatud**	Osaliselt avatud	**Kaalud** (piiratud)	Eralitsents (NVIDIA, produktsioonis tasuline)
OpenAI (Whisper seeria)	Osaliselt avatud	Suletud	**Avatud**	Osaliselt avatud	**Kaalud** ja API	**MIT (lubav)**
OpenAI (o1-mini)	Suletud	Suletud	**Avatud**	Suletud	**Kaalud**	**Apache 2.0 (lubav)**
xAI (Grok seeria)	Osaliselt avatud	Suletud	Suletud (Grok-1)	Osaliselt avatud	**Kaalud** ja API	**Apache 2.0 (lubav, Grok-1)**
Cohere (Command R+)	Osaliselt avatud	Suletud	**Avatud**	Osaliselt avatud	**Kaalud** ja API	CC BY-NC-SA (mitteäriline)
Apple Intelligence	Suletud	Suletud	Suletud	Suletud	Tootesse integreeritud	Omanduslik
Midjourney	Suletud	Suletud	Suletud	Suletud	Tootesse integreeritud	Omanduslik
Black Forest Labs (FLUX seeria)	Suletud	Suletud	Osaliselt avatud	Suletud	**Kaalud** ja API	Flux Litsents (mitteäriline)
Allen AI (OLMo 2 seeria)	**Avatud**	**Avatud**	**Avatud**	**Avatud**	**Kaalud**	**Apache 2.0 (lubav)**

5.1.2 Mudeli seletavus

Oluline on mõista, et isegi täielikult valge kasti mudel, mille kõik komponendid (andmed, kood, kaalud) on avalikud, on funktsionaalselt siiski must kast ilma spetsiifiliste seletavus (*explainability/interpretability*) tehnikateta. Mudeli sisemine toimimine on liiga keeruline, et seda otse mõista, isegi kui meil on olemas selle kaalud ja nende aktivatsioonid kindlate sisendväärtuste korral. Seetõttu on vaja eraldi meetodeid, et uurida, miks mudel teeb mingeid otsuseid.

Küll aga on erinevus selles, kas on must kast ainult mudel või kogu süsteem. Kui kasutada välist API-teenust, saab kogu süsteemi pidada mustaks kastiks. Sellise rakendusliidese integreerija ei kontrolli mudeli taristut, võimalikke päringutele eelnevaid või vastustele järgnevaid filtreid ega saa garanteerida teenuse stabiilsust. Kui aga majutada avatud kaaludega mudelit ise, muutub süsteem läbipaistvamaks, sest juurutajal on sel juhul kontroll kogu ahela üle alates riistvarast kuni mudeli tulemite töötlemiseni. See eristus on praktikas tihti olulisemgi kui mudeli enda täielik läbipaistvus. Järgnevalt me käsitleme lihtsustavalt musta kasti süsteemidena AI-süsteeme, mille korral ei ole juurutajal juurdepääsu mudelile ja selle käitamisprotsessile, valge kasti süsteemidena aga süsteeme, kus mudelit käitab juurutaja.

5.2 Milliseid meetmeid saab kasutada musta kasti süsteemi kallutatuse puhul?

Musta kasti süsteemide puhul, näiteks välise API-teenuse kasutamise korral puudub juurdepääs mudeli treeningandmetele ja lähtekoodile. Seetõttu ei ole võimalik mudelit ennast muuta. Kallutatuse leevendamise meetmed peavad olema süsteemivälised ja keskenduma sellele, kuidas mudelit testitakse, rakendatakse ja milliseid kaitsemehhanisme selle ümber ehitatakse.

Süstemaatiline testimine ja auditeerimine. Enne leevendusmeetmete rakendamist tuleb mõista süsteemi kallutatuse olemust ja ulatust. Kuna mudeli sisse ei näe, tuleb seda testida käitumuslikult ehk päringute-tulemite analüüsi (*input-output analysis*) kaudu. See sarnaneb tarkvara läbistustestimisega (*penetration testing*).

- **Võrdlustestid.** Süsteemi saab testida standardsete kallutatuse mõõtlusalustega (nagu näiteks SafetyBench või AgentHarm), et võrrelda selle tulemusi teiste mudelitega.
- **Lööktestimine (*red teaming*).** See hõlmab süsteemi teadlikku käitamist provokatiivsete või tundlike sisendandmetega, et tuvastada potentsiaalselt kallutatud või kahjulikke vastuseid. Eesmärk on leida üles süsteemi nõrgad kohad.
- **Kontrafaktuaalne analüüs.** Süsteemi käitatakse erinevate minimaalselt muudetud sisendväärtustega, näiteks muutes ainult ühte demograafilist tunnust (nimi, sugu, vanus), ja jälgitakse, kas mudeli tulem muutub oluliselt. See võib aidata tuvastada diskrimineerivat käitumist.

Väliste kaitsepiirete rakendamine. Kui mudelit ennast muuta ei saa, tuleb luua selle ümber kaitsekiht, mis filtreerib nii päringuid kui ka tulemeid.

- **Päringu valideerimine.** Kontrollitakse kasutaja päringuid enne nende mudelile saatmist. Kui päring on olemuselt provokatiivne või võib esile kutsuda kallutatuse, saab selle blokeerida või ümber sõnastada.
- **Tulemi valideerimine.** Kontrollitakse mudeli vastust enne selle kuvamist kasutajale. Kui vastus sisaldab kallutatud keelekasutust, kahjulikke stereotüüpe või faktivigu, saab selle asendada neutraalse teatega („Ma ei saa sellele küsimusele vastata“) või suunata inimoperaatori poole.

Organisatsioonilised meetmed. Tihti on kõige tõhusamad meetmed mitte tehnilised, vaid strateegilised.

- **Kasutusjuhtude piiramine.** Kui testimise käigus selgub, et süsteem on mingis valdkonnas

(nt meditsiiniliste soovituste andmine või kandidaatide hindamine) ebausaldusväärne või kallutatud, on kõige kindlam meede süsteemi kasutamine selles kontekstis keelata.

- **Inimjärelevalve tagamine.** Kõrge riskiga otsuste puhul ei tohiks tehisintellekti tulem olla kunagi lõplik otsus, vaid ainult informatiivne otsustusabi inim-eksperdile. See tagab, et lõpliku vastutuse kannab inimene.
- **Teenustaja vahetamine.** Kui süsteemi kallutatus on liiga suur ja teenustaja ei paku piisavaid lahendusi, on viimaseks abinõuks teise, usaldusväärsema toote või teenustaja valimine. See loob turusurvet, mis motiveerib arendajaid looma õiglasemaid süsteeme.

5.3 Milliseid meetmeid saab kasutada valge kasti süsteemi kallutatuse puhul?

Valge kasti stsenaariumi korral on olemas juurdepääs mudeli sisemistele komponentidele: mudeli kaaludele ja mõnikord ka treeningandmetele. See olukord tekib tavaliselt siis, kui süsteemi arendatakse organisatsiooni sees või koostöös partneriga, kes tagab läbipaistvuse, või kui organisatsioon evitab välise süsteemi ise. See annab laiemad võimalused kallutatuse leevendamiseks, kuna on võimalik tegeleda probleemi algpõhjustega AI-süsteemi elutsükli igas etapis.

Andmestiku ja mudeli auditeerimine. Kallutatuse algpõhjuste leidmiseks saab kasutada mudeli diagnostika tööriistu, nagu Google'i What-If Tool või Microsofti Responsible AI Toolbox. Need võimaldavad interaktiivselt analüüsida treenitud mudeli käitumist erinevatel andmesegmentidel, visualiseerida tulemuste erinevusi demograafiliste gruppide vahel ja genereerida kontrafaktuaale. Selline audit annab aluse järgmiste leevendusmeetmete valikuks.

Andmetasandi meetmed. Kui auditi käigus on ilmnenud probleeme andmestikuga, saab enne mudeli treenimist rakendada järgmisi meetmeid.

- **Andmete tasakaalustamine.** Alaesindatud andmegruppide mõju saab vähendada näiteks andmete ümberkaalumisega (*reweighing*) treeningprotsessis, kus alaesindatud andmepunktidele antakse suurem kaal [133].
- **Andmete täiendamine (*data augmentation*).** Alaesindatud gruppide jaoks saab luua täiendavaid sünteetilisi andmeid, et parandada nende esindatust.

Mudelitasandi meetmed. Neid meetmeid rakendatakse mudeli treenimise käigus, et suunata seda õiglasemate tulemuste poole.

- **Õiglusele optimeeritud treening.** Algoritmidele saab lisada õigluse piiranguid, näiteks karistusfunktsiooni, mis takistab mudelil tundlike tunnuste kasutamist või soodustab võrdseid tulemusi eri gruppide vahel.
- **Mudeli peenhäälestamine (*fine-tuning*).** Eeltreenitud mudelit saab peenhäälestada kvaliteetse ja tasakaalustatud andmestiku abil, et „ümber õpetada" soovimatuid seoseid ja õpetada õiglasemat käitumist.
- **Võistlev kallutatuse eemaldamine (*adversarial debiasing*).** Tehnika, kus üks mudel püüab teha ennustust ja teine püüab selle põhjal ära arvata tundlikku tunnust. Esimest mudelit treenitakse nii, et teine oma ülesandes ebaõnnestuks [134].
- **Promptitehnika (*prompt engineering*).** Mudeli käitumist saab suunata ka hoolikalt koostatud juhiste ehk promptide abil. Lisades päringule meeldetuletusi õigluse olulisusest või

paludes mudelil enne vastamist analüüsida potentsiaalset kallutatust, on võimalik diskrimineerivat käitumist vähendada [135].

- **Tunnuste suunamine (*feature steering*).** See on tehnika, kus muudetakse otse mudeli sisemisi närvivõrgu aktivatsioone. Tuvastades mudeli sees spetsiifilised tunnused, mis on seotud kallutatusega, saab nende mõju maha suruda või muuta, vähendades seeläbi soovimatut käitumist [136].

Väljunditasandi meetmed. Neid meetmeid rakendatakse pärast mudeli ennustust, et korrigeerida selle tulemit.

- **Otsustuslävede kalibreerimine.** Mudeli otsustuslävesid (nt mis skoorist alates loetakse taotlus heakskiidetuks) saab kalibreerida erinevate gruppide jaoks eraldi, et saavutada soovitud õigluse näitaja. Tööriistad nagu Fairlearn pakuvad seda funktsionaalsust [137].

5.4 Kallutatuse leevendamise meetodite rakendamise näited

Tehisintellekti kallutatuse leevendamise edukate, reaalses elus rakendatud näidete leidmine on keerulisem kui ebaõnnestumiste leidmine. Seda seetõttu, et ebaõnnestumised on skandaalid ja seega uudisväärtusega. Süsteem, mis töötab märkamatult, ootuspäraselt ja õiglaselt, ei paku avalikkusele niivõrd suurt huvi.

Seetõttu keskendub see alajaotis mitte niivõrd reaalsete, tootmises olevate süsteemide edulugudele, kuivõrd tehniliste meetodite ja lähenemiste demonstreerimisele. Need näited pärinevad peamiselt akadeemilistest uuringutest ja tööriistakomplektide õpetustest ning illustreerivad, kuidas kallutatust on võimalik tehnilisel tasandil leevendada.

Näide. Andmete ümberkaalumine krediidiskoorimisel (AIF360).

AIF360 tööriistakomplekti õpetus illustreerib andmetasandi sekkumise põhimõtet, kasutades vanuselise kallutatuse näidet Saksamaa krediidiandmestikus.

- **Probleem.** Treenimata mudel näitas selget eelistust vanema vanusegrupi suhtes.
- **Rakendatud meetod.** Kasutati ümberkaalumise eeltöötlustehnikat (*reweighing*), mis annab treeningu käigus suurema kaalu alaesindatud gruppide andmepunktidele.
- **Tulemus.** Demonstratsioonis langes gruppidevaheline erinevus positiivsete laenuotsuste saamisel nullini. See näitab, kuidas andmete eeltöötlus võib ideaaltingimustes kallutatust vähendada [133].

Näide. Otsustuslävede optimeerimine laenuotsustel (Fairlearn).

Microsofti Fairlearni õpetus demonstreerib järeltöötluse tehnikat, kus korrigeeritakse juba treenitud mudeli otsuseid.

- **Probleem.** Laenuotsuste mudel näitas soolist kallutatust.
- **Rakendatud meetod.** Kasutati algoritmi **ThresholdOptimizer**, mis leiab iga demograafilise grupi jaoks eraldi optimaalse otsustusläve (nt skoor, millest alates laen heaks kiidetakse), et tagada õiglusnäitajate täitmine.
- **Tulemus.** See meetod vähendas oluliselt gruppidevahelisi lõhesid, tuues kaasa väikese languse mudeli üldises täpsuses, mis on tüüpiline kompromiss õigluse ja täpsuse vahel [137].

Näide. Diskrimineerimise vähendamine promptimise abil (Anthropic).

Suurte keelemudelite puhul on üheks tõhusaks meetmeks suure keelemudeli käitumise suunamine hoolikalt koostatud promptide abil.

- **Kontekst.** Leiti, et Claude-seeria keelemudel tegi kõrge riskiga stsenaariumides süstemaatiliselt erinevaid otsuseid, sõltuvalt päringus mainitud demograafilistest tunnustest.
- **Rakendatud meede.** Enne tegeliku ülesande andmist lisati mudeli sisendandmetesse spetsiaalseid promptid, näiteks paludes mudelil enne vastamist analüüsida potentsiaalset kallutatust ja keskenduda ainult kandidaadi kvalifikatsioonile.
- **Tulemus.** Sellised promptipõhised sekkumised vähendasid oluliselt mudeli kalduvust diskrimineerivatele otsustele, säilitades samal ajal mudeli üldise otsustusvõime muudes aspektides [135].

Näide. Kaitsepiirete kasutamine jaekaubanduse vestlusrakenduses (NeMo GuardRails).

NVIDIA NeMo Guardrails on avatud lähtekoodiga raamistik, mis võimaldab rakendada reegli- või eeskirjapõhiseid kaitsekihte musta kasti tüüpi LLM-ide ümber, et tagada õiglus ja ohutus. Seda on lihtne integreerida olemasolevate valmismudelitega nii levitaja süsteemis kui pilves, sest see ei eelda juurdepääsu mudeli kaaludele ega eelda uue mudeli treenimist.

- **Probleem.** Jaekaubanduse veslusrakendus (nt lemmikloomatarvete soovitaja) genereeris kallutatud vastuseid, näiteks sisaldasid tema vastused soolisi eelarvamusi tootesoovitustes, või olid selle vastused muul, teemaga mitteseotud moel kallutatud, mis halvas kasutajakogemust.
- **Rakendatud meetod.** Sellises olukorras saaks lisada päringu- ja tulemikaitseröövpad koos Colang-eeskirjadega, mis tuvastavad kallutatust (nt demograafilisel alusel) ja suunavad vastused tagasi teemale. Lisaks saaks rakendada sisumoderatsiooni ja teemakontrolli, kasutades hindamiseks LLM-kohtunikuna (*LLM-as-judge*) metoodikat. Need toimiksid kaitsekihina mudeli ümber.
- **Tulemus.** Selline lähenemine võiks vähendada kallutatud vastuste arvu ja parandada kasutajakogemust. Avaldatud juhtumiuuringuid selle kohta jaekaubanduse kontekstis seni ei ole, kuid olemas on teadusartikleid ja tehnilisi blogipostitusi, mis demonstreerivad NVIDIA NeMo Guardrails kasutamist kaitsepiirete loomiseks ja turvalisuse parendamiseks üldiselt [138, 139].

Näide. Sotsiaalse kallutatuse korrigeerimine tunnuste suunamisega (Anthropic).

See on tehniliselt keerulisem lähenemine, mis sekkub otse mudeli sisemisse toimimisse.

- **Kontekst.** Suurtes keelemudelites on võimalik tuvastada spetsiifilisi mustreid (tunnuseid, *features*) närvivõrgu aktivatsioonides, mis on seotud mingite kontseptsioonidega, sealhulgas sotsiaalse kallutatusega.
- **Rakendatud meede.** Kasutati tunnuste suunamise (*feature steering*) tehnikat. Kui mudel genereerib vastust, muudetakse reaalajas neid närvivõrgu aktivatsioone, mis on seotud kallutatusega.
- **Tulemus.** Mõõdukas sekkumine suutis vähendada sotsiaalset kallutatust mudeli vastustes ilma, et see oleks oluliselt kahjustanud mudeli üldist sooritusvõimet muudes ülesannetes [136].

Bibliograafia

[1] *Euroopa Parlamendi ja Nõukogu määrus (EL) 2024/1689, 13. juuni 2024, millega nähakse ette tehisintellekti käsitlevad ühtlustatud õigusnormid ning muudetakse määruseid (EÜ) nr 300/2008, (EL) nr 167/2013, (EL) nr 168/2013, (EL) 2018/858, (EL) 2018/1139 ja (EL) 2019/2144 ning direktiive 2014/90/EL, (EL) 2016/797 ja (EL) 2020/1828 (Tehisintellekti määrus)*. https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=OJ:L_202401689. Euroopa Liidu Teataja L 2024/1689, 12.7.2024. Kasutatud: 2025-06-03. 2024.

[2] K. Abendroth Dias *et al. Generative AI Outlook Report – Exploring the Intersection of Technology, Society and Policy*. Toim. E. Navajas Cawood *et al.* Publications Office of the European Union, Luxembourg. EUR 40337, JRC142598. 2025. DOI: 10.2760/1109679. URL: https://publications.jrc.ec.europa.eu/repository/handle/JRC142598.

[3] Euroopa Parlament ja nõukogu. *Euroopa Parlamendi ja nõukogu määrus (EL) 2024/903, 13. märts 2024, koostalitlusvõime tugevdamise kohta avalikus sektoris kogu liidus (Interoperable Europe Act)*. Euroopa Liidu Teataja L 2024/903, 11.04.2024. 2024. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=OJ:L_202400903.

[4] European Commission. *AI Continent Action Plan*. https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan. COM(2025) 165 final. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Brussels, aprill 2025.

[5] European Commission. *AI Continent Action Plan: Accelerating trustworthy AI in key sectors including public services*. European Commission – Digital Strategy Library. COM(2025)165 final; Accessed: 2025-07-09. 2025. URL: https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan.

[6] European Commission. *Commission sets course for Europe's AI leadership with ambitious AI Continent Action Plan*. European Commission – Digital Strategy. Accessed: 2025-07-09. Aprill 2025. URL: https://digital-strategy.ec.europa.eu/en/news/commission-sets-course-europes-ai-leadership-ambitious-ai-continent-action-plan.

[7] Stuart Russell, Karine Perset ja Marko Grobelnik. *Updates to the OECD's definition of an AI system explained*. 29. november 2023. URL: https://oecd.ai/en/wonk/ai-system-definition-update.

[8] Harini Suresh ja John Guttag. „A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle". Teoses: *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 2021.

[9] *Masinõppe projektiks valmistumine: kontroll küsimustik*. Tehniline raport. Juhendmaterjal masinõppe projektiks valmistumiseks, sisaldab kontrollküsimustikku. Oktoober 2020. URL: https://www.kratid.ee/_files/ugd/980182_6c40cd0822e74f89aae111d5b146d612.pdf.

[10] *Juhendmaterjal krattide hankimiseks*. Tehniline raport. Juhend annab ülevaate sagedasematest probleemidest ja võimalikest lahendustest andmeteaduse projektide puhul. November 2019. URL: https://www.kratid.ee/_files/ugd/980182_8d0fcb0be030467aa55a4f2839560cbe.pdf.

[11] *Krati tehnoloogia valdkonna ülevaade projekti juhtidele*. Tehniline raport. Ülevaade krati tehnoloogia valdkonnast projekti juhtidele. URL: https://www.kratid.ee/_files/ugd/980182_9b3ebd2cec704bc98b2727007bdca580.pdf.

[12] Joanna Redden *et al. Automating Public Services: Learning from Cancelled Systems*. Supported publication by the Data Justice Lab. September 2022. URL: https://www.carnegieuktrust.org.uk/publications/automating-public-services-learning-from-cancelled-systems/.

[13] *Tehisintellekti ja masinõppe tehnoloogia riskide ja nende leevendamise võimaluste uuring*. Tehniline raport. Riigi Infosüsteemi Amet, Cybernetica AS, 2024. URL: https://www.ria.ee/kuberturvalisus/riiklik-koordinatsioonikeskus-ncc-ee/tulevikutehnoloogiate-kuberturvalisus.

[14] High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[15] „Eesti Vabariigi põhiseadus“ (). URL: https://www.riigiteataja.ee/akt/115052015002?leiaKehtiv.

[16] *Euroopa Liidu lepingu ja Euroopa Liidu toimimise lepingu konsolideeritud versioonid Euroopa Liidu lepingu konsolideeritud versioon Euroopa Liidu toimimise lepingu konsolideeritud versioon Protokollid Euroopa Liidu toimimise lepingu lisad Deklaratsioonid, mis on lisatud 13. detsembril 2007 alla kirjutatud Lissaboni lepingu vastu võtnud valitsustevahelise konverentsi lõppaktile.* ELT C 202, 7.6.2016, p. 1–388. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/?uri=celex%3A12016ME%2FTXT.

[17] *Euroopa Liidu toimimise lepingu (ELi toimimise leping) konsolideeritud versioon*. C 326/49. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/PDF/?uri=CELEX:12012E/TXT.

[18] European Union. *Summary: Treaty on the Functioning of the European Union*. URL: https://eur-lex.europa.eu/ET/legal-content/summary/treaty-on-the-functioning-of-the-european-union.html.

[19] Equality ja UK Human Rights Commission. *Met Police's use of facial recognition tech must comply with human rights law, says regulator*. Permission granted to intervene in judicial review regarding Met Police live facial recognition policy. August 2025. URL: https://www.equalityhumanrights.com/met-polices-use-facial-recognition-tech-must-comply-human-rights-law-says-regulator.

[20] Europol Innovation Lab. *AI Bias in Law Enforcement: A Practical Guide*. Observatory Report from the Europol Innovation Lab. Luxembourg, 2025. DOI: 10.2813/8081090. URL: https://www.europol.europa.eu/publications-events/publications/ai-bias-in-law-enforcement.

[21] Hadas Kotek, Rikker Dockum ja David Q. Sun. *Gender Bias and Stereotypes in Large Language Models*. August 2023. URL: https://arxiv.org/pdf/2308.14921.

[22] USA District Court for the Northern District of California. *Order Granting Preliminary Collective Certification – Mobley v. Workday, Inc.* Case No. 23-cv-00770-RFL. Mai 2025. URL: https://www.govinfo.gov/content/pkg/USCOURTS-cand-3_23-cv-00770/pdf/USCOURTS-cand-3_23-cv-00770-1.pdf.

[23] Various legal & news outlets. *Mobley v. Workday, Inc. – AI hiring discrimination litigation (claims proceed)*. Federal case alleging disparate impact by Workday's applicant-screening AI; claims allowed to proceed and conditional,certification steps in 2025. Juuni 2025. URL: https://www.jdsupra.com/legalnews/ai-bias-lawsuit-against-workday-reaches-5119500/.

[24] Arshon Harper. *Complaint in Harper v. Sirius XM Radio, LLC*. Case No. 2:25-cv-12403-TGB-APP, U.S. District Court for the Eastern District of Michigan. August 2025. URL: https://www.fisherphillips.com/a/web/9MgBFhRPkToes9HKGYtqKF/aAkZr3/harper-v-sirius-xm-radio-ed-mi-225cv12403.pdf.

[25] Fisher Phillips. *Another Employer Faces AI Hiring Bias Lawsuit: 10 Actions You Can Take to Prevent AI Litigation*. Complaint filed 8/2025 alleging AI hiring tool produced racially biased results; case at complaint stage. August 2025. URL: https://www.fisherphillips.com/en/news-insights/another-employer-faces-ai-hiring-bias-lawsuit.html.

[26] Equality ja Human Rights Commission. *Uber Eats courier wins payout with help of equality watchdog, after facing problematic AI checks*. Märts 2024. URL: https://www.equalityhumanrights.com/media-centre/news/uber-eats-courier-wins-payout-help-equality-watchdog-after-facing-problematic-ai.

[27] Ülle Madise *et al.*, toim. *Eesti Vabariigi Põhiseadus – Kommenteeritud väljaanne*. https://pohiseadus.ee. Accessed: 2025-06-03. 2020.

[28] Ülle Madise *et al.*, toim. *Eesti Vabariigi Põhiseadus – Kommenteeritud väljaanne: § 12*. https://pohiseadus.ee/sisu/3483/paragrahv_12. Accessed: 2025-06-03. 2020.

[29] *Euroopa Liidu põhiõiguste harta*. 2012/C 326/02. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:12012P/TXT.

[30] Publications Office of the European Union. *Summary: Charter of Fundamental Rights of the European Union*. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/?uri=legissum:l33501.

[31] *Nõukogu direktiiv 2000/43/EÜ, 29. juuni 2000, millega rakendatakse võrdse kohtlemise põhimõte sõltumata isikute rassilisest või etnilisest päritolust*. EÜT L 180, 19/07/2000, p. 22–26. URL: http://data.europa.eu/eli/dir/2000/43/oj.

[32] *Nõukogu direktiiv 2000/78/EÜ, 27. november 2000, millega kehtestatakse üldine raamistik võrdseks kohtlemiseks töö saamisel ja kutsealale pääsemisel*. EÜT L 303, 2.12.2000, p. 16–22. URL: http://data.europa.eu/eli/dir/2000/78/oj.

[33] Ühinenud Rahvaste Organisatsiooni Peaassamblee. *Inimõiguste ülddeklaratsioon*. ÜRO peaassamblee resolutsioon 217 A. Detsember 1948. URL: https://inimoigusedeestis.ee/hartadpaktid-deklaratsioonid-kohtud-jne/inimoiguste-ulddeklaratsioon/.

[34] ÜKP fondide võrdõiguslikkuse kompetentsikeskus. *Inimõiguste ülddeklaratsioon*. URL: https://kompetentsikeskus.sm.ee/et/vordsed-voimalused/vordne-kohtlemine/oigusaktid/uro/inimoiguste-ulddeklaratsioon.

[35] Ühinenud Rahvaste Organisatsioon. *Konventsioon naiste diskrimineerimise kõigi vormide likvideerimise kohta*. URL: https://www.riigiteataja.ee/akt/23988.

[36] Ühinenud Rahvaste Organisatsioon. *Rahvusvaheline konventsioon rassilise diskrimineerimise kõigi vormide likvideerimise kohta*. URL: https://www.riigiteataja.ee/akt/23980.

[37] Ühinenud Rahvaste Organisatsioon. *Kodaniku- ja poliitiliste õiguste rahvusvaheline pakt. Mitteametlik tõlge*. URL: https://www.riigiteataja.ee/akt/23982.

[38] Ühinenud Rahvaste Organisatsioon. *Majanduslike, sotsiaalsete ja kultuurialaste õiguste rahvusvaheline pakt. Mitteametlik tõlge*. URL: https://www.riigiteataja.ee/akt/23981.

[39] Council of Europe. *Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*. Council of Europe Treaty Series 225. Vilnius: Council of Europe, september 2024. URL: https://rm.coe.int/1680afae3c.

[40] *Eesti seisukohad Euroopa Komisjoni soovituse „Nõukogu otsus, millega antakse luba allkirjastada Euroopa Liidu nimel tehisintellekti, inimõigusi, demokraatiat ja õigusriiki käsitlev Euroopa Nõukogu raamkonventsioon", kohta. Seletuskiri*. URL: https://eelnoud.valitsus.ee/main#aZWBecQZ.

[41] Ceyhun Necati Pehlivan, Nikolaus Forgó ja Peggy Valcke, toim. *The EU Artificial Intelligence (AI) Act: A Commentary*. Netherlands: Kluwer Law International, 2024. ISBN: 9789403532271. URL: https://law-store.wolterskluwer.com/s/product/the-eu-artificial-intelligence-ai-act-a-commentary/01tPg000007gkK9IAI.

[42] CEDPO AI and Data Working Group. *Fundamental Rights Impact Assessments: What are they? How do they work?* https://cedpo.eu/wp-content/uploads/CEDPO-micro-insight-paper-fundamental-rights-impact-assessments.pdf. Micro-Insights Series; Authors: Thomas Ajoodha and Jared Browne; published [10/1/2025]. Jaanuar 2025.

[43] ISO/IEC. *ISO/IEC 42005:2025 Artificial Intelligence — Management system for artificial intelligence — Requirements and guidance*. 2025. URL: https://www.iso.org/standard/42005 (vaadatud 08.07.2025).

[44] Council of the European Union. *Outcomes of the discussions on simplification activities in the digital field*. Document ST-9383-2025-INIT. https://data.consilium.europa.eu/doc/document/ST-9383-2025-INIT/en/pdf. Juuni 2025.

[45] Commission Nationale de l'Informatique et des Libertés (CNIL). *Artificial Intelligence and Public Services: CNIL Publishes Results of Its Sandbox*. 2025. URL: https://www.cnil.fr/en/artificial-intelligence-and-public-services-cnil-publishes-results-its-sandbox (vaadatud 08.07.2025).

[46] Justiits- ja Digiministeerium. *Valitsus avaldas toetust Eesti liitumisele Balti tehisaru gigatehase projektiga*. JustDigi uudis. August 2025. URL: https://www.justdigi.ee/uudised/valitsus-avaldas-toetust-eesti-liitumisele-balti-tehisaru-gigatehase-projektiga.

[47] *Andmete ja tehisintellekti valge raamat 2024-2030*. URL: https://www.kratid.ee/_files/ugd/7df26f_9a6d060409214b9da2774e5b5eabf717.pdf.

[48] *Eesti digiühiskonna arengukava 2030*. URL: https://www.mkm.ee/digiriik-ja-uhenduvus/digiuhiskonna-arengukava-2030.

[49] High-Level Expert Group on Artificial Intelligence. *OECD AI Principles overview*. 2019. URL: https://oecd.ai/en/ai-principles.

[50] *Euroopa deklaratsioon digiõiguste ja -põhimõtete kohta digikümnendiks*. OJ C 23, 23.1.2023, p. 1–7. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:32023C0123(01).

[51] European Parliament: Directorate-General for External Policies of the Union ja A. Ünver. *Artificial intelligence (AI) and human rights – Using AI as a weapon of repression and its impact on human rights – In-depth analysis*. Publications Office of the European Union, 2024. URL: https://data.europa.eu/doi/10.2861/52162.

[52] European Commission, Directorate-General for Digital Transformation. *Code of Practice for General-Purpose AI Models: Safety and Security Chapter*. Published on 10 July 2025; endorsed by the AI Board and the Commission as an adequate voluntary tool for compliance with the AI Act. 2025. URL: https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

[53] European Parliament and Council of the European Union. *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on European Critical Raw Materials*. EUR-Lex. Accessed: 2025-07-09. 2022. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:32022R2065.

[54] *CCommission launches public consultation and call for evidence on the Apply AI Strategy*. European Commission – Digital Strategy. Aprill 2025. URL: https://digital-strategy.ec.europa.eu/en/consultations/commission-launches-public-consultation-and-call-evidence-apply-ai-strategy.

[55] *Commission seeks feedback on the future Strategy for Artificial Intelligence in Science*. European Commission – Research & Innovation. Aprill 2025. URL: https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/commission-seeks-feedback-future-strategy-artificial-intelligence-science-2025-04-10_en.

[56] *AI Continent – new cloud and AI development act*. European Commission. 2025. URL: https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/14628-AI-Continent-new-cloud-and-AI-development-act_en.

[57] Arcangelo Leone de Castris. *AI governance around the world: European Union*. Report (version v3). Published August 8, 2025; accessed 2025-09-23. August 2025. DOI: 10.5281/zenodo.17054419. URL: https://zenodo.org/records/17054419.

[58] Euroopa Liit. „Euroopa Parlamendi ja nõukogu määrus (EL) 2016/679, 27. aprill 2016, füüsiliste isikute kaitse kohta isikuandmete töötlemisel ja selliste andmete vaba liikumise ning direktiivi 95/46/EÜ kehtetuks tunnistamise kohta (isikuandmete kaitse üldmäärus) (EMPs kohaldatav tekst)“. *Euroopa Liidu TEataja L119* 59 (4. mai 2016), lk. 1–88.

[59] *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*. ELT L 119, 4.5.2016, p. 89–131. URL: http://data.europa.eu/eli/dir/2016/680/oj.

[60] *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 (EMPs kohaldatav tekst.)* ELT L 295, 21.11.2018, p. 39–98. URL: http://data.europa.eu/eli/reg/2018/1725/oj.

[61] European Data Protection Board. *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models*. Tehniline raport. Adopted on 17 December 2024. European Data Protection Board, detsember 2024. URL: https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf.

[62] Kris Shrishak. *AI-Complex Algorithms and Effective Data Protection Supervision - Bias Evaluation*. Tehniline raport. European Data Protection Board, Support Pool of Experts Programme, jaanuar 2025. URL: https://www.edpb.europa.eu/system/files/2025-01/d1-ai-bias-evaluation_en.pdf.

[63] P. Dewitte. „AI Meets the GDPR: Navigating the Impact of Data Protection on AI Systems". Teoses: *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*. Toim. N. A. Smuha. Cambridge Law Handbooks. Cambridge University Press, 2025, lk. 133–157.

[64] *Isikuandmete kaitse seadus*. RT I, 04.01.2019, 11. URL: https://www.riigiteataja.ee/akt/104012019011?leiaKehtiv.

[65] German Federal Constitutional Court. *Judgment of 16 February 2023 – Automated data analysis*. Veebruar 2023. URL: https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/EN/2023/02/rs20230216_1bvr154719en.html.

[66] German Federal Constitutional Court (Bundesverfassungsgericht). *Legislation in Hesse and Hamburg regarding automated data analysis for the prevention of criminal acts is unconstitutional*. Press release (English) summarising 1 BvR 1547/19 and 1 BvR 2634/20 judgments. Veebruar 2023. URL: https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/EN/2023/bvg23-018.html?nn=148454.

[67] Euroopa Parlamendi mõttekoda. *Algorithmic discrimination under the AI Act and the GDPR*. https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2025)769509. 2025.

[68] Giovanni Sartor, Francesca Lagioia *et al.* „The impact of the General Data Protection Regulation (GDPR) on artificial intelligence" (2020).

[69] European Data Protection Board. *EDPB Opinion 2024/28 on certain data protection aspects related to the processing of personal data in the context of AI models*. 2024. URL: https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf.

[70] AKI. *Inspektsioonile esitatud rikkumisteadetest*. 2024. URL: https://www.aki.ee/uudised/inspektsioonile-esitatud-rikkumisteadetest (vaadatud 08.07.2025).

[71] *Avaliku teabe seadus*. RT I 2000, 92, 597. URL: https://www.riigiteataja.ee/akt/130122024005?leiaKehtiv.

[72] American Civil Liberties Union. *ACLU v. Clearview AI*. 2022. URL: https://www.aclu.org/cases/aclu-v-clearview-ai.

[73] American Civil Liberties Union ja ACLU of Illinois. *Big Win, Settlement Ensures Clearview AI Complies With Groundbreaking Illinois Biometric Privacy Law*. Mai 2022. URL: https://www.aclu.org/press-releases/big-win-settlement-ensures-clearview-ai-complies-with-groundbreaking-illinois.

[74] Jane Bambauer. „Cambridge Analytica and the Meaning of Privacy Harm" (). URL: https://pep.gmu.edu/wp-content/uploads/sites/28/2019/01/Bambauer_PEP_White_Paper_Cambridge_Analytica.pdf.

[75] *Facebook sai Cambridge Analytica skandaali eest pool miljonit naela trahvi*. ERR. Oktoober 2018. URL: https://www.err.ee/871909/facebook-sai-cambridge-analytica-skandaali-eest-pool-miljonit-naela-trahvi.

[76] Michael Veale ja Frederik Zuiderveen Borgesius. „AI Meets the GDPR“. Teoses: *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*. Toim. Markus Dubber, Frank Pasquale ja Sunit Das. Cambridge University Press, 2022, lk. 576–596. URL: https://www.cambridge.org/core/books/cambridge-handbook-of-the-law-ethics-and-policy-of-artificial-intelligence/ai-meets-the-gdpr/94476F95CE264B80C00B46BA8506F474.

[77] European Parliament and Council of the European Union. *Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 concerning measures for a high common level of cybersecurity across the Union, and amending Regulation (EU) 2019/881*. EUR-Lex. Detsember 2022. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:32022L2555.

[78] Hajo Michael Holtz ja Jonas Ledendal. *AI Data Governance – Overlaps Between the AI Act and the GDPR*. Preprint, Uppsala University / Lund University. Forthcoming in "Law, Innovation and Technology"(Spring 2026). September 2025. URL: https://uu.diva-portal.org/smash/get/diva2:1996843/FULLTEXT01.pdf.

[79] GOV.UK. *Auditing algorithms: the existing landscape, role of regulators and future outlook*. Digital Regulation Cooperation Forum. Findings from the DRCF Algorithmic Processing workstream – Spring 2022. September 2022. URL: https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook.

[80] *Euroopa Parlamendi ja nõukogu määrus (EL) 2023/2854, ühtlustatud õigusnormide kohta, millega reguleeritakse õiglast juurdepääsu andmetele ja andmete kasutamist, millega muudetakse määrust (EL) 2017/2394 ja direktiivi (EL) 2020/1828 (andmemäärus)*. OJ L, 2023/2854, 22.12.2023. Detsember 2023. URL: http://data.europa.eu/eli/reg/2023/2854/oj.

[81] *Euroopa Parlamendi ja nõukogu määrus (EL) 2022/868, Euroopa andmehalduse kohta ning millega muudetakse määrust (EL) 2018/1724 (andmehalduse määrus)*. OJ L 152, 3.6.2022, p. 1–44. Mai 2022. URL: http://data.europa.eu/eli/reg/2022/868/oj.

[82] European Union. *European data governance*. EUR-Lex: European Union legal summary. Summary of Regulation (EU) 2022/868 (Data Governance Act). URL: https://eur-lex.europa.eu/ET/legal-content/summary/european-data-governance.html.

[83] Ernst & Young Baltic AS. *Eesti andmehalduse metoodikaprojekt: Eesti andmehalduse raamistik*. Tehniline raport. Funded by the European Commission’s Structural Reform Support Programme and implemented by Ernst & Young Baltic AS in cooperation with the European Commission’s Directorate-General for Structural Reform Support. August 2020. URL: https://www.kratid.ee/_files/ugd/980182_8a6e296751984c0c975f61ea8358770e.pdf.

[84] Eesti Statistikaamet. *Andmekvaliteedi haldus: Asutuse ülesanded andmekvaliteedi tagamisel*. Tehniline raport. Andmehalduse tehniline dokument. Statistikaamet, Majandus- ja Kommunikatsiooniministeerium, mai 2023. URL: https://www.kratid.ee/_files/ugd/980182_346bbb77c4eb4118b4fdb40ab5ee1b1a.pdf.

[85] Eesti Statistikaamet. *Andmehalduse juhised. Andmekvaliteedi juhis*. Tehniline raport. Juhis annab ülevaate andmekvaliteedi mõistest ja olemusest, andmekvaliteedi juhtimise tegevustest, andmekvaliteedi mudelist ja kvaliteediprobleemide põhjuste analüüsist ja mõjude hindamisest. Statistikaamet, Majandus- ja Kommunikatsiooniministeerium, mai 2023. URL: https://www.kratid.ee/_files/ugd/980182_f04ff5003d2b408faff965c81ab7e974.pdf.

[86] *Andmekvaliteedi haldus: kasutuslood ja funktsionaalsused*. Tehniline raport. Andmehalduse tehniline dokument. Statistikaamet, Majandus- ja Kommunikatsiooniministeerium, märts 2023. URL: https://www.kratid.ee/_files/ugd/980182_6f260e68d00742ffbc660e9e661ededf.pdf.

[87] Eesti Statistikaamet. *Andmekirjelduse juhis*. Tehniline raport. Versioon 2.0. Juhis annab ülevaate andmekirjelduse mõistest, koostamise protsessidest ning praktilisi soovitusi andmekirjelduse koostamisel ja haldamisel. August 2023. URL: https://www.kratid.ee/_files/ugd/7df26f_aa72d2833b514f20a07531d3b5bf3d9c.pdf.

[88] Eesti Statistikaamet. *Andmekirjelduse juhis, Lisa 1: Andmekirjelduse standard*. Tehniline raport. Versioon 3.0. Andmekirjelduse standard esitab andmekirjelduse elementide (metaandmete) täpse loetelu ja semantika ning on juurutatud RIHAKEse ja RIHA rakendustes. Juuli 2024. URL: https://www.kratid.ee/_files/ugd/980182_87343e9f929143b1a566d1c96791ec09.pdf.

[89] *Andmekirjelduse juhis, Lisa 2: Sõnastike koostamine andmekirjeldustes. Praktiline juhis*. Tehniline raport. Versioon 0.9. Juhis esitab nõuded ja soovitused ühtsete sõnastike loomisele andmekirjelduste koostamise kontekstis ning koondab Statistikaameti praktilised kogemused valdkonna terminoloogia korrastamisel ja sõnastike koostamisel. Eesti Statistikaamet, juuni 2024. URL: https://www.kratid.ee/_files/ugd/980182_98310b0a04fc4cdbb09fdd47394b6aa4.pdf.

[90] Eesti Statistikaamet. *Ärireeglite kirjeldamise praktilisi näiteid*. Tehniline raport. Versioon 1.0. Dokument sisaldab asutuste praktilisi näiteid ärireeglite kirjeldamisel. Statistikaamet, mai 2024. URL: https://www.kratid.ee/_files/ugd/980182_d412c26dce67450ab6dd6be439fc92f1.pdf.

[91] Majandus- ja Kommunikatsiooniministeerium ning Statistikaamet. *Lühijuhis: Andmekogude andmekoosseisude uuendamine RIHAs*. Tehniline raport. Juhis selgitab riigiasutustele RIHA andmekirjelduste uuendamise eesmärki ja tegevusi vastavalt Vabariigi Valitsuse ülesandele. Detsember 2024. URL: https://www.kratid.ee/_files/ugd/980182_d68b16df30ff4e67ab747ae971ea5318.pdf.

[92] Majandus- ja Kommunikatsiooniministeerium. *Andmeladude olemus ja selle funktsioonid. Selgitav juhis*. Tehniline raport. Juhis selgitab andmeladude olemust ja kasutusjuhtude arendamist ning ühtlustab andmeladudega seotud tehnilise ja õigusliku raamistikku. Aprill 2023. URL: https://www.kratid.ee/_files/ugd/980182_5b9c740580a543c0ab48480b07c7ce5f.pdf.

[93] *Juhendmaterjal andmete annoteerimiseks*. Tehniline raport. Juhendi eesmärk on aidata pildi- või videofailide omanikel annoteerida oma andmeid süsteemide treenimiseks. URL: https://www.kratid.ee/_files/ugd/980182_4140abbee26c4bda80d31d895a15455a.pdf.

[94] Euroopa Parlament ja Euroopa Liidu Nõukogu. *Määrus (EL) 2023/988 Euroopa Parlamendi ja nõukogu 10. mai 2023. aasta määrus üldise tooteohutuse kohta, millega muudetakse Euroopa Parlamendi ja nõukogu määrust (EL) nr 1025/2012 ja direktiivi (EL) 2020/1828 ning tühistatakse Euroopa Parlamendi ja nõukogu direktiiv 2001/95/EÜ ja nõukogu direktiiv 87/357/EMÜ*. EUR-Lex. 2023. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:32023R0988.

[95] *Toote nõuetele vastavuse seadus*. URL: https://www.riigiteataja.ee/akt/111032025003?leiaKehtiv.

[96] Euroopa Parlament ja Euroopa Liidu Nõukogu. *Määrus (EL) 2024/2847, mis käsitleb digielemente sisaldavate toodete küberturvalisuse horisontaalseid nõudeid ja millega muudetakse määrusi (EL) nr 168/2013 ja (EL) 2019/1020 ning direktiivi (EL) 2020/1828*. EUR-Lex. 2024. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=OJ:L_202402847.

[97] *Küberturvalisuse seadus*. URL: https://www.riigiteataja.ee/akt/121062024015?leiaKehtiv.

[98] International Organization for Standardization. *Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making*. ISO/IEC TR 24027:2021, Edition 1. November 2021. URL: https://www.iso.org/standard/77607.html.

[99] IEEE Standards Association. *IEEE P7003 - Algorithmic Bias Considerations*. Tehniline raport. Published: 2025-01-24. 2023. DOI: 10.1109/IEEESTD.2025.10851955. URL: https://ieeexplore.ieee.org/servlet/opac?punumber=10851953.

[100] National Institute of Standards ja Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Tehniline raport NIST AI 100-1. Available at: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf. U.S. Department of Commerce, jaanuar 2023. URL: https://doi.org/10.6028/NIST.AI.100-1.

[101] R. Schwartz *et al. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. Available online at https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence. Special Publication (NIST SP) - 1270, National Institute of Standards and Technology. 2022. URL: https://doi.org/10.6028/NIST.SP.1270.

[102] National Institute of Standards ja Laurie E. Locascio Technology Gina M. Raimondo. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1, U.S. Department of Commerce. Juuli 2024. URL: https://doi.org/10.6028/NIST.AI.600-1.

[103] Riigi Infosüsteemi Amet. *Eesti Infoturbestandard. Riskihaldusjuhend*. 2024. URL: https://eits.ria.ee/et/abimaterjalid/riskihaldusjuhend.

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

[105] *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.

[106] *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.

[107] *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.

[108] *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.

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

[110] Heidi Ledford. „Millions affected by racial bias in health-care algorithm“. *Nature* 574.31 (2019), lk. 2.

[111] Jeffrey Dastin. „Amazon scraps secret AI recruiting tool that showed bias against women“. *reuters.com* (2018). URL: https://www.reuters.com/article/uk-amazon-com-jobs-automation-insight-idUKKCN1MK08G.

[112] Weihao Xuan *et al. MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation*. 2025. arXiv: 2503.10497 [cs.CL]. URL: https://arxiv.org/abs/2503.10497.

[113] Wenxuan Wang *et al.* „All Languages Matter: On the Multilingual Safety of LLMs“. Teoses: *Findings of the Association for Computational Linguistics: ACL 2024*. Toim. Lun-Wei Ku, Andre Martins ja Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, august 2024, lk. 5865–5877. DOI: 10.18653/v1/2024.findings-acl.349. URL: https://aclanthology.org/2024.findings-acl.349/.

[114] Nikhil Sharma, Kenton Murray ja Ziang Xiao. „Faux Polyglot: A Study on Information Disparity in Multilingual Large Language Models“. Teoses: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Toim. Luis Chiruzzo, Alan Ritter ja Lu Wang. Albuquerque, New Mexico: Association for Computational Linguistics, aprill 2025, lk. 8090–8107. ISBN: 979-8-89176-189-6. DOI: 10.18653/v1/2025.naacl-long.411. URL: https://aclanthology.org/2025.naacl-long.411/.

[115] Julia Unwin. *Kindness, Emotions and Human Relationships: The Blind Spot in Public Policy*. Report. Carnegie UK Trust, 2018. URL: https://www.carnegieuktrust.org.uk/publications/kindness-emotions-and-human-relationships-the-blind-spot-in-public-policy/.

[116] Karen Yeung. *A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility Within a Human Rights Framework*. MSI-AUT Report MSI-AUT (2018) 05. Posted: 10 Dec 2018. The University of Birmingham, november 2018, lk. 94. URL: https://ssrn.com/abstract=3286027.

[117] Auli Viidalepp. „The expected AI as a sociocultural construct and its impact on the discourse on technology“. Doktoritöö. University of Tartu, 2023.

[118] Johann Laux ja Hannah Ruschemeier. „Automation Bias in the AI Act: On the Legal Implications of Attempting to De-Bias Human Oversight of AI“. *arXiv preprint arXiv:2502.10036* (2025).

[119] Pierre Le Jeune *et al. RealHarm: A Collection of Real-World Language Model Application Failures*. 2025. arXiv: 2504.10277 [cs.CY]. URL: https://arxiv.org/abs/2504.10277.

[120] Gerwin van Schie, Laura Candidatu ja Diletta Huyskes. „Motherhood in the Datafied Welfare State:: Investigating the Gendered and Racialized Enactment of Citizenship in Dutch Algorithmic Governance“. Teoses: mai 2025, lk. 209–226. ISBN: 9789048562718. DOI: 10.2307/jj.28874939.18.

[121] Yuxuan Dai ja Zhaohui Wang. „Predictive Policing and Algorithmic Fairness“. *Synthese* 201.3 (2023), lk. 1–29.

[122] EEOC. *Press Release: iTutorGroup to Pay $365,000 to Settle EEOC Discriminatory Hiring Suit.* 2023. URL: https://www.eeoc.gov/newsroom/itutorgroup-pay-365000-settle-eeoc-discriminatory-hiring-suit.

[123] David Weinberger. *Playing with AI Fairness*. https://pair-code.github.io/what-if-tool/ai-fairness.html. Google PAIR blog post. 2018.

[124] Ninareh Mehrabi *et al.* „A Survey on Bias and Fairness in Machine Learning“. *arXiv preprint arXiv:1908.09635* (2019). URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11880872/.

[125] S. Ghosh, S. Sanyal ja S. Mukhopadhyay. „A Classification Framework for AI Bias Impacts and Effective Mitigation Strategies“. *Social Sciences* 6.1 (2024), lk. 3. DOI: 10.3390/socsci6010003. URL: https://www.mdpi.com/2413-4155/6/1/3.

[126] S. Ghosh, S. Sanyal ja S. Mukhopadhyay. *Artificial Intelligence (AI) Bias Impacts Classification Framework for Effective Mitigation*. 2023. URL: https://pure.jgu.edu.in/id/eprint/6745/1/Artificial%5C%20intelligence%5C%20%5C%28AI%5C%29%5C%20bias%5C%20impacts%5C%20classification%5C%20framework%5C%20for%5C%20effective%5C%20mitigation.pdf.

[127] X. H. Phan, T. H. Nguyen, H. M. Le *et al.* „Fairness-aware Artificial Intelligence: A Systematic Review“. *Neural Networks* (2024). DOI: 10.1016/j.neunet.2024.03.001. URL: https://www.sciencedirect.com/science/article/pii/S0893395224002667.

[128] Isabel Barberá ja Murielle Popa-Fabre. *Expert Report on Privacy and Data Protection Risks in Large Language Models (LLMs)*. Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, Convention 108. Presented at the 48th Plenary Meeting, 17 June 2025. 2025. URL: https://media.licdn.com/dms/document/media/v2/D4E1FAQFUWjwoEARZpQ/feedshare-document-pdf-analyzed/B4EZdAFWr0HgAY-/0/1749126851206.

[129] Daniel Bone *et al.* „Applying machine learning to facilitate autism diagnostics: pitfalls and promises.“ eng. *J Autism Dev Disord* 45.5 (mai 2015), lk. 1121–1136. ISSN: 1573-3432 (Electronic); 0162-3257 (Print); 0162-3257 (Linking). DOI: 10.1007/s10803-014-2268-6.

[130] Michael Roberts *et al.* „Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans“. *Nature Machine Intelligence* 3 (märts 2021).

[131] Gilles Vandewiele *et al.* „Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling“. *Artificial Intelligence in Medicine* 111 (jaanuar 2021), lk. 101987. ISSN: 0933-3657. DOI: 10.1016/j.artmed.2020.101987. URL: http://dx.doi.org/10.1016/j.artmed.2020.101987.

[132] Ching-Hua Chuan *et al.* „EXplainable Artificial Intelligence (XAI) for facilitating recognition of algorithmic bias: An experiment from imposed users' perspectives“. *Telematics and Informatics* 91 (2024), lk. 102135. ISSN: 0736-5853. DOI: https://doi.org/10.1016/j.tele.2024.102135. URL: https://www.sciencedirect.com/science/article/pii/S073658532400039X.

[133] Rachel K. E. Bellamy *et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. Oktoober 2018. URL: https://arxiv.org/abs/1810.01943.

[134] Jenny Yang *et al.* „An adversarial training framework for mitigating algorithmic biases in clinical machine learning“. *NPJ Digital Medicine* 6.1 (2023). Published online 2023-03-29, lk. 55. DOI: 10.1038/s41746-023-00805-y. URL: https://www.nature.com/articles/s41746-023-00805-y.

[135] Alex Tamkin *et al. Evaluating and Mitigating Discrimination in Language Model Decisions*. 2023. arXiv: 2312.03689 [cs.CL]. URL: https://arxiv.org/abs/2312.03689.

[136] Esin Durmus *et al. Evaluating Feature Steering: A Case Study in Mitigating Social Biases*. 25. oktoober 2024. URL: https://anthropic.com/research/evaluating-feature-steering.

[137] Hilde Weerts *et al.* „Fairlearn: Assessing and Improving Fairness of AI Systems“. *Journal of Machine Learning Research* 24 (2023). URL: http://jmlr.org/papers/v24/23-0389.html.

[138] Traian Rebedea *et al. NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails*. 2023. arXiv: 2310.10501 [cs.CL]. URL: https://arxiv.org/abs/2310.10501.

[139] Georgi Botsihhin ja Lorenzo Boccaccia. *Enhancing LLM Capabilities with NeMo Guardrails on Amazon SageMaker JumpStart*. AWS Machine Learning Blog. Amazon Web Services. 5. veebruar 2025. URL: https://aws.amazon.com/blogs/machine-learning/enhancing-llm-capabilities-with-nemo-guardrails-on-amazon-sagemaker-jumpstart.

[140] *ISO/IEC 11179-1:2023. Information technology — Metadata registries (MDR)*. URL: https://www.iso.org/standard/78914.html.

[141] *ISO/IEC/IEEE 15288:2023.Systems and software engineering — System life cycle processes*. URL: https://www.iso.org/standard/81702.html.

[142] *ISO/IEC/IEEE 15939:2017. Systems and software engineering — Measurement process*. URL: https://www.iso.org/standard/71197.html.

[143] *ISO/IEC 19501:2005. Information technology — Open Distributed Processing — Unified Modeling Language (UML) Version 1.4.2*. URL: https://www.iso.org/standard/32620.html.

[144] *ISO/IEC 25002:2024. Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model overview and usage*. URL: https://www.iso.org/standard/78175.html.

[145] *ISO/IEC 25012:2008. Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model*. URL: https://www.iso.org/standard/35736.html.

[146] *ISO/IEC 25024:2015. Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of data quality*. URL: https://www.iso.org/standard/35749.html.

[147] *ISO/IEC 25023:2016. Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Measurement of system and software product quality*. URL: https://www.iso.org/standard/35747.html.

Lisa A AI-süsteemi kallutatuse leevendamisega kaudselt seotud standardid

AI-süsteemi kallutatusega otseselt seotud rahvusvahelised standardid on esitatud alajaotises 3.3. Siin toome ära standardid, mis aitavad leevendada AI-süsteemi kallutatust kaudsemalt.

Tabel 3. AI-süsteemi kallutatuse leevendamisega kaudselt seotud standardid

Standardi number	Standardi nimi	Selgitus
ISO/IEC 11179-1:2023 [140]	Metadata registries (MDR)	Standard loob aluse metaandmete ja metaandmeregistrite mõistmiseks, käsitledes metaandmete haldamist ja andmete kirjeldusi.
ISO/IEC 15288:2023 [141]	Systems and software engineering – System life cycle processes	Standard kirjeldab süsteemi elutsükli protsesse, mis toetavad süsteemide arendust, hankimist ja koostööd erinevate poolte vahel.
ISO/IEC 15939:2017 [142]	Systems and software engineering — Measurement process	Standard kirjeldab mõõtmisprotsessi süsteemi, aidates määratleda, rakendada ja täiustada näitajaid vastavalt konkreetsetele vajadustele.
ISO/IEC 19501:2005 [143]	Information technology – Open Distributed Processing – Unified Modeling Language (UML)	Standard kirjeldab UML-keelt, mis võimaldab tarkvarasüsteemide visuaalset modelleerimist ja dokumenteerimist ühtsel ja standardiseeritud viisil.
ISO/IEC 25002:2024 [144]	Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality model overview and usage	Standard määratleb kvaliteedimudelite raamistiku, kirjeldades nende struktuuri, tähendust ning seoseid mõõtmise, nõuete ja hindamisega.
ISO/IEC 25012:2008 [145]	Software product Quality Requirements and Evaluation (SQuaRE) – Data quality model	Standard kirjeldab andmekvaliteedi üldmudelit, mida saab kasutada andmekvaliteedi nõuete määratlemiseks, mõõtmiseks ja hindamiseks struktureeritud andmete puhul.

Standardi number	Standardi nimi	Selgitus
ISO/IEC 25024:2015 [146]	Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data quality	Standard määratleb andmekvaliteedi näitajad, mis võimaldavad kvantitatiivselt hinnata andmekvaliteeti vastavalt ISO/IEC 25012 standardis toodud omadustele. Standard pakub juhiseid nende näitajate rakendamiseks kogu andme-elutsükli vältel ning on mõeldud kasutamiseks erinevates infosüsteemides ja organisatsioonilistes rollides alates arendajatest kuni kvaliteedijuhtideni.
ISO/IEC 25023:2016 [147]	Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of system and software product quality	Standard määratleb süsteemi ja tarkvara kvaliteedi hindamiseks kvantitatiivsed näitajad, mida kasutatakse koos standardiga ISO/IEC 25010. Standard toetab kvaliteedinõuete määratlemist ja hindamist kogu arendustsükli vältel ning sobib kasutamiseks kvaliteedi tagamisel, juhtimisel, tarnimisel, hankimisel ja hooldusel.