TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Kaupo Lepasepp 183820IAAM

AUTONOOMSETE JA INTELLIGENTSETE SÜSTEEMIDEGA SEOTUD EETILISED JA SOTSIAALSED NÕUDED

Magistritöö

Juhendaja: Peeter Lorents
PhD

Tallinn 2020

Autorideklaratsioon

Kinnitan, et olen koostanud antud magistritöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Kaupo Lepasepp

17. mai 2020

Annotatsioon

Magistritöö käsitleb nõrka tehisintellekti kasutavate autonoomsete ja intelligentsete infosüsteemide (AIS) loomise ja käitamisega seonduvaid eetilisi ja sotsiaalseid probleeme, nõudeid eetilisele AIS-ile ja nende nõuete täitmise meetodeid.

Eetilisele AIS-ile esitatud eetilised ja sotsiaalsed nõuded rakenduvad mitte AIS-ile, vaid sellise süsteemi arendaja ja käitaja tegevusele. Eetiline on AIS, (1) mille käitaja on tuvastanud konkreetses käitamise kontekstis asjakohased eetilised ja sotsiaalsed nõuded, (2) määranud nende alusel nõuded süsteemi toimimisele ja (3) mis täidab neid nõuetekohaselt. Magistritöö kirjeldab üldnõudeid eetilisele AIS-i. Iga konkreetse AIS-i käitamisel tuleb üldnõudeid konkretiseerida vastavalt süsteemi kasutamise eetilisele ja sotsiaalsele kontekstile, arvestades kontekstide erinevust ja võimalikku muutumist ajas. Eetilise AIS-i nõudeid tuleb AIS-ide kavandamisel ja käitamisel järgida juba täna.

AIS-iga seonduvad eetilised ja sotsiaalsed probleemid, riskid ja nende põhjused on mitmekesised ning need avalduvad AIS-i elukaare erinevates etappides. Magistritöös pakub autor välja eetilise AIS-i elukaare halduse kontseptsiooni. Eetilise AIS-i elukaare halduse, seda toetavate meetodite ja vahendite käsitlemisega annab töö juhised, mille abil saab AIS-i analüütik, kavandaja ja käitaja analüüsida konkreetsele süsteemile rakenduvaid eetilisi ja sotsiaalseid nõudeid ning hallata nendest kõrvalekaldumise riski.

Magistritöö on kirjutatud eesti keeles ning sisaldab teksti 72 leheküljel, 5 peatükki, 14 joonist, 5 tabelit.

Abstract

Ethical and Social Requirements for Autonomous and Intelligent Systems

This Masters' thesis analyses the ethical and social problems accompanying development and operation of weak artificial intelligence based autonomous and intelligent systems (AIS), requirements for ethical AIS and methods for achieving these requirements.

The Master's thesis concludes that the ethical and social requirements for an ethical AIS are requirements applicable towards a developer or an operator of the system. Only such AIS can be qualified as an ethical AIS for which the following criteria are met: (1) the operator has analysed the ethical and social norms relevant in the social context where AIS will be used, (2) the operator has set the ethical and social requirements to AIS based on the analysis and (3) the system meets the set ethical and social requirements in compliance with the specification. The general requirements for an ethical AIS are the following: (1) respect to human rights and human dignity, (2) fairness of the decisions and impact of AIS, (3) transparency and explainability of the decisions of AIS, (4) reliability and risk management and (5) accountability and liability for harmful decisions and impact of AIS.

The current development of technology has no practical way to assure universal compliance with the ethical and social requirements for an ethical AIS. Instead the analysts must determine the ethical and social requirements arising from the context of using the AIS and rules for balancing the conflicting requirements, describe the ethical and social requirements adequate and appropriate for the AIS being developed and design measures implementing these requirements separately in the analysis and development phase of a specific AIS. The requirements for ethical AIS are applicable already today, and thus must be considered in analysis, development, and operation of a future-proof AIS.

Ethical and social problems and risks associated with AIS arise from various sources and are exposed in different stages of AIS life cycle. The Master's thesis analyses various types of problems and risk with aim to help the analysist to identify risks and problems relevant for a specific AIS being analysed, designed, or operated. The Master's thesis also envisages a life-cycle management for ethical AIS with description of the relevant phases. Together with the systematic description of various measures, methods, and tools for supporting the life cycle of ethical AIS, this thesis aims to serve a as practical guidance for practitioners in defining and meeting the ethical and social requirements for a specific AIS.

The Master's thesis further describes more common methods and tools available for developers and operators of an ethical AIS. Although the maturity level of such methods and tools is relatively low, using a combination of the available methods and tools facilitates compliance with the ethical and social requirements.

The thesis is in Estonian and contains 72 pages of text, 5 chapters, 14 figures, 5 tables.

Lühendite ja mõistete sõnastik

Arendaja	Arendaja tähistab käesolevas töös isikut või organisatsiooni, mis osaleb AIS-i väljatöötamises. Arendaja ise ei kasuta AIS-i oma organisatsiooni protsessides, mistõttu ei mõjuta arendaja tegevus AIS-i arendamisel vahetult inimesi või ühiskonda. AIS-i arendamisel võib osaleda üks või mitu arendajat.
AI HLEG	Euroopa Komisjoni kõrgetasemeline tehisintellekti eksperdirühm
AIS	Autonoomne ja intelligentne süsteem tähistab antud töös nõrka tehisintellekti (sageli masinõppekomponent) sisaldavaid inimese suutlikkust toetavaid riist- ja tarkvarasüsteeme [1, lk 12], mis teadmistega opereerides toetavad või teostavad otsuse langetamise protsessi. Erinevad organisatsioonid tähistavad selliseid süsteeme erinevate mõistetega[1]. Autor kasutab töös läbivalt IEEE eelistatud mõistet AIS („*AIS*“).
Eetiline AIS	Eetiline AIS on AIS, (1) mille käitaja on tuvastanud konkreetses käitamise kontekstis asjakohased eetilised ja sotsiaalsed nõuded, (2) määranud nende alusel nõuded süsteemi toimimisele ja (3) mis täidab neid nõuetekohaselt.
EL	Euroopa Liit
EK	Euroopa Komisjon
EN	Euroopa Nõukogu
IEEE	Elektri- ja Elektroonikainseneride Instituut
Kallutatus	Infosüsteemi kallutatus („*bias*“) on süsteemi omadus diskrimineerida teatud isikuid või gruppe võrreldes teistega ebaõiglaselt ja süstemaatiliselt. Ebaõiglase diskrimineerimisega on tegemist juhul, kui süsteem jätab isiku või grupi ilma hüvest või soodsast otsusest või teeb nende suhtes negatiivse mõjuga otsuse põhjusel, mis on ebamõistlik ja sobimatu. Süsteem on kallutatud, kui ebaõiglased otsused on korduvad ja süsteemi toimimist iseloomustavad; ühekordne süsteemi viga ei ole käsitletav kallutatusena [2, lk 332–333], [3, lk 5].

[1] Nt IEEE kasutab mõistet “intelligentsed ja autonoomsed süsteemid” [140], EL mõistet “tehisintellekt” [36] ning OECD “tehisintellekti süsteem” [14].

Käitaja	Käitaja tähistab käesolevas töös isikut või organisatsiooni, kes juurutab ja kasutab AIS-i oma organisatsiooni eesmärkide saavutamiseks. AIS avaldab mõju inimestele või ühiskonnale vahetult käitaja eesmärkide saavutamise raames. Reeglina on AIS-il üks käitaja.
Nõrk tehisintellekt	Nõrk tehisintellekt tähistab antud töös erinevatel vahenditel (nt nagu masinõppes kasutatavad tehisnärvivõrgud, geneetilised algoritmid, evolutsioonilised meetodid) põhinevat tehisintellekti, mis küll opereerib teadmistega, kuid ei oma nende teadmistega seonduvat arusaamist ega vaba tahet [3, lk 5], [4, lk 1020, 1026].
OECD	Majanduskoostöö ja Arengu Organisatsioon [5].
STS	Sotsiotehniline süsteem on süsteem, mis hõlmab endas nii sotsiaalseid kui ka tehnilisi komponente [6, lk 120]. STS-is toimivad inimesed, tehnoloogilised komponendid ja neid ümbritsev keskkond samaaegsetes ja kompleksetes vastastikkustes mõjudes, millega tuleb arvestada STS või selle komponentide kavandamisel [7, lk 5].
Tundlik tunnus	Tundlikud tunnused on selline informatsioon isiku kohta, mille alusel on isikute eristamine, grupeerimine või isiku suhtes otsuste tegemine reeglina keelatud[1] (nt reeglina on keelatud töötaja palkamisel lähtuda inimese soost) või lubatud üksnes erandlikes oludes (nt erandina on soost lähtumine lubatud günekoloogiliste või androloogiliste haiguste diagnoosimisel).
Valdavalt inimese loodud algoritm	Valdavalt inimese loodud algoritm tähistab käesolevas töös infosüsteemis rakendavat algoritmi, mille on primaarselt loonud inimene. Infosüsteemi tegevus piirdub valdavalt inimese loodud algoritmi rakendamisel algoritmis sisalduvat juhiste täitmisega.
Valdavalt masina loodud algoritm	Valdavalt masina loodud algoritm tähistab käesolevas töös infosüsteemis rakendatavat algoritmi, mille on primaarselt loonud tarkvara või riistvara. Inimene on küll loonud vastava algoritmi loomiseks kasutatava masina (tarkvara või riistvara) või isegi osalenud algoritmi loomises, kuid inimene ei ole vastava algoritmi vahetuks autoriks. Valdavalt masina loodud algoritmi näiteks on programmide automaatse sünteesi [8, lk 119–120] või masinõppe vahenditega loodud algoritmid.

[1] Näiteks on isikud Eestis seaduse alusel kaitstud diskrimineerimise eest rahvuse, rassi, nahavärvuse, usutunnistuse või veendumuste, vanuse, puude või seksuaalse sättumuse alusel [98, Lõik 1]

Sisukord

Jooniste loetelu

Tabelite loetelu

Sissejuhatus

Käesolevas magistritöös uurib autor, kas autonoomsete ja intelligentsete süsteemide (AIS) analüüsil ja kavandamisel tuleb järgida eetilisi ja sotsiaalseid nõudeid, milline on nende nõuete sisu ja kuidas neid nõudeid täita.

Infosüsteemide eetika on peaaegu sama vana kui kaasaegne infotehnoloogia. AIS-ide, tehisintellekti ja masinõppe levik on aga viimastel aastatel kaasa toonud vastavate süsteemide levikust ja eripäradest tulenevate eetiliste ja sotsiaalsete probleemide mõistmise ja lahendamisega seonduva teadustegevuse kiire kasvu [9, lk 4], [10, lk 3]. Lisaks teadlastele tegelevad AIS-ide eetiliste ja sotsiaalsete nõuetega ka riigid [11, lk 2] [12, lk 1, 8–18] [13], rahvusvahelised organisatsioonid [14], äriühingud [15] ja standardimisasutused [1] [16]. Teadus- ja arendustegevuse kasvu taga on tõdemus, et ehkki AIS-ide, tehisintellekti ja masinõppe levik pakub ühiskonnale palju uusi võimalusi, siis tuleb ühiskonna tasakaalustatud arengu tagamiseks vältida või lahendada selliste süsteemide rakendamisega kaasneda võivaid eetilisi ja sotsiaalseid probleeme.

AIS-i eetiliste ja sotsiaalsete nõuetega seonduv kirjandus on mahukas ja lähtub erinevatest vaatenurkadest. Enamasti on käsitluste lähtepunktiks AIS-idele iseloomulik teatav autonoomsus ja intelligentsus ning nendest omadustest tulenevad piirangud AIS-i otsuste arusaadavusele ja kontrollitavusele inimese poolt. Ka ei pruugi inimesele võimetekohane olla arusaamine AIS-ide tööks kasutatavatest (suur)andmetest. Samas pole selge, kas vastavad käsitlused on teoreetilised või on eetilise AIS-i nõuded piisavalt konkreetsed, et neid oleks võimalik AIS-ide kavandamisel ja käitamisel rakendada. Ebaselgust on ka küsimuses, kas vastavad nõuded on soovituslikud või kohustuslikud.

Töö esimeses peatükis selgitab autor AIS-i eetiliste ja sotsiaalsete riskide ja nõuete seisukohast olulisi põhimõisteid. Vaadeldakse eetiliste ja sotsiaalsete nõuete olemust ning vastavate hinnangute omistamise süsteemi. Edasi analüüsitakse, kas tänase teadaoleva tehnoloogia taseme juures rakenduvad eetilised ja sotsiaalsed nõuded AIS-ile ja kellele saab omistada nõuetega seonduvaid eetilisi ja sotsiaalseid hinnanguid.

Töö teises peatükis avatakse AIS-iga kaasneda võivate eetilise ja sotsiaalse mõjuga probleemide olemust. Nii valdavalt inimese loodud algoritmid kui valdavalt masina loodud algoritmid võivad põhjustada eetilisi ja sotsiaalseid riske. Selles kontekstis uurib autor, kas valdavalt masina loodud algoritmid põhjustavad spetsiifilisi riske võrreldes valdavalt inimese loodud algoritmidega. Käsitluse eesmärgiks on analüütikute abistamine vastavate riskide haldamiseks sobivate meetodite ja vahendite valikul.

Töö kolmandas peatükis annab autor ülevaate AIS-ile esitatud eetilistest ja sotsiaalsetest nõuetest. Peatükis kontrollib autor hüpoteesi, et need nõuded on praktikas rakendatavad ning neist tuleb lähtuda AIS-ide arendamisel ja käitamisel. Hüpoteesi kontrollimiseks käsitleb autor Eesti vaatepunktist olulisemaid „algoritmi eetikaga“ seonduvaid avaliku ja erasektori algatusi ja raamistikke. Eesmärgiks on luua AIS-i arendajaid ja käitajaid abistav ülevaade eetilistest ja sotsiaalsetest nõuetest, millele eetiline AIS peab vastama.

Töö neljandas peatükis kontrollib autor hüpoteesi, et AIS-i eetiliste ja sotsiaalsete riskide haldamiseks on vajalik terviklik AIS-i elukaare haldus, mille väljatöötamisel ja rakendamisel tuleb tagada teadlikkus süsteemi eetilistest ja sotsiaalsetest mõjudest alates süsteemi kavandamisest kuni selle kasutamise lõpetamiseni. Peatüki eesmärgiks on eetilise AIS-i elukaare halduse kirjeldamine arendajaid ja käitajaid abistaval viisil.

Töö viiendas peatükis annab autor ülevaate AIS-i eetiliste ja sotsiaalsete riskide haldamiseks kasutatavatest metoodikatest, vahenditest ja nende liikidest.

Töö on teostatud süstemaatilise ülevaate meetodil [17] [18, lk 27]. Töö põhineb varem avaldatud teadustööde, kirjanduse ja poliitikadokumentide kvalitatiivsel sisuanalüüsi [18, lk 192–197], [19].

Töö tegeleb nõrka tehisintellekti rakendava AIS-i probleemide ja nõuetega, millel on eetiline ja sotsiaalne mõju. Tugeva tehisintellekti probleemistikuga töö ei tegele.

1 AIS-i seos eetiliste ja sotsiaalsete nõuetega

Käesolev töö käsitleb AIS-ide kasutamisega seonduvaid eetilisi ja sotsiaalseid riske ja nende haldamise meetodeid. Teema käsitlemiseks on esmalt vajalik selgitada vastavaid põhimõisteid – mis on eetilised ja sotsiaalsed riskid, mis on AIS-id ning milles seisnevad just nende põhjustatud eetilise ja sotsiaalsete riskid.

1.1 Eetilised ja sotsiaalsed riskid

ISO 31000:2018 defineerib riski kui ebakindluse mõju eesmärkidele. Mõju on kõrvalekaldumine oodatust, kusjuures mõju võib olla nii positiivne, negatiivne kui mõlemat [20]. Mõju võib luua või põhjustada nii võimalusi kui ka ohtusid. Seega peegeldavad eetilised ja sotsiaalsed riskid süsteemi oleku kõrvalekaldumist süsteemile seatud eetilistest ja sotsiaalsetest eesmärkidest.

Autor käsitleb eetiliste eesmärkidena selliseid eesmärke, mille saavutamisele või mittesaavutamisele on võimalik omistada eetilist hinnangut. Laiemalt peavad eetika olemasoluks eksisteerima (a) eetilised hinnangud, (b) asjad, millele eetilisi hinnanguid omistatakse, ning (c) nõuded ja eeskirjad, millele tuginedes eetilisi hinnanguid omistatakse [21, lk 287]. Eetilised nõuded on ühiskonnas eksisteerivad (õige) käitumise normid ja nõuded [21, lk 303], millega võrreldes toimub eetilise hinnangu omistamine. Eetilised nõuded võivad olla erinevates eetikates, ühiskondades, professioonides jne erinevad [1, lk 164–166], [22, lk 311]. Reeglina on eetilised nõuded väljendatud läbi süsteemi lubatud või keelatud käitumise või olekute kirjeldamise [21, lk 298]. Eetilisi hinnanguid väljendatakse selliste väärtustega nagu „hea“ või „halb“, „eetiline“ või „ebaeetiline“ jne. Erinevad eetikad võivad eetiliste hinnangutena kasutada ka muid väärtusi [21, lk 287].

Eetiliselt hinnatavad asjad peavad seonduma põhjustamise või esilekutsumisega. Eetiliselt on võimalik hinnata üksnes niisugust põhjustamist, mille käigus mingi süsteem (S_1) põhjustab või põhjustas – teo või tegevusetusega – teise süsteemi (S_2) siirdumise teatavasse seisundisse [21, lk 287–289]. Sealjuures on oluline, et põhjustamine toimub põhjustaja (S_1) vaba tahte alusel – võimalik on tuvastada põhjustaja, põhjustaja on intelligentne süsteem, mis saab aru, mida ta põhjustab ja on sellise põhjustamise endale eesmärgiks seadnud [21, lk 291–294]. Reeglina ei ole

seetõttu võimalik eetilist hinnangut omistada ebaintelligentse süsteemi või rikki läinud süsteemi poolsele põhjustamisele [21, lk 295]. Samas võib autori hinnangul eetilist hinnangut omistada sellisele põhjustamisele, kus ühe süsteemi (S_1) kontrollile allutatud vaba tahteta süsteem (S_{vts}) põhjustab teise süsteemi (S_2) seisundi muutuse. Sellisel juhul omistatakse eetiline hinnang S_1 poolsele põhjustamisele. Selliselt on võimalik eetilise hinnangu omistamine teadlikult ohtliku või kahju tekitamisele suunatud tarkvara loomisele (nt lunavara[1]).

Eetiliste hinnangute omistamine toimub läbi süsteemi S_1 poolt süsteemis S_2 esile kutsutud ehk põhjustatud oleku võrdlemise asjakohastes eetilistes nõuetes esitatud ja süsteemi S_2 jaoks lubatud või keelatud oleku kirjeldusega. Kui eetiliselt hinnataval viisil põhjustatud olek vastab eetilistes nõuetes esitatud lubatud olekule, siis saab anda hinnangu „hea“ või „eetiline“. Vastupidisel juhul on hinnanguks „halb“ või „ebaeetiline“. Eetilise hinnangu omistamisel tuleb hinnang anda hinnatavale põhjustamisele tervikuna, mitte aga selle fragmentidele [21, lk 298–303].

Lisaks eetilistele nõuetele peavad ühiskonnaelus osalejad oma eesmärkide saavutamiseks järgima ka muid, eetiliselt mittehinnatavaid nõudeid, mida autor nimetab sotsiaalseteks nõueteks. Sellised nõuded võivad olla esteetilised (ilus või inetu), majanduslikud (kasumlik või kahjumlik), õiguslikud (õiguslikult lubatud või keelatud), tehnilised (nt standarditele vastavus või mittevastavus) jne. Osaliselt võivad sellised nõuded kattuda ka eetiliste nõuetega, kuid ei pea seda tingimata tegema. Nii näiteks ei ole süsteemile esitatud tehnilised nõuded (nt töökindluse nõuded) reeglina eetiliselt hinnatavad. Samas kui aga süsteemi arendaja kavandab töökindluse nõudeid eirava süsteemi ja põhjustab sellega eetiliselt hinnatava tagajärje, siis muutub ka töökindlusega seonduvate nõuete eiramine eetiliselt hinnatavaks.

1.2 Intellekti ja tehisintellekti seos eetikanõuetega

Intellekt on süsteemi suutlikkus opereerida teadmistega. Opereerimise mõiste on siinkohal lai, hõlmates näiteks nii uute teadmiste loomist kui ka olemasolevate

[1] Lunavara on pahavara tüüp, mis piirab kasutaja ligipääsu arvutile, krüpteerides failid või lukustades kogu süsteemi. Krüpteeritud failidele või lukustatud süsteemile ligipääsu taastamiseks peab kasutaja maksma lunaraha [141, lk 13].

teadmiste süstematiseerimist või leidmist. Tehisintellekt on inimeste poolt loodud tehniliste süsteemide suutlikkus opereerida teadmistega [21, lk 60–61].

Mitte igasugune intellekti ilming ei ole eetiliselt hinnatav. Eetiliselt hinnatava põhjustamise jaoks ei piisa üksnes intellekti rakendamisest. Eetiliselt hinnatav põhjustamine nõuab lisaks veel ka intellekti evija vaba tahet. Seega saab eetilisi ja valdavat osa sotsiaalseid nõudeid esitada ja vastavaid hinnanguid anda ainult sellisele intellekti rakendamisele, mis on käsitletav vaba tahte avaldusena.

Vaba tahte olemasolu alusel eristatakse tugevat ja nõrka tehisintellekti [23, lk 10]. Tugev tehisintellekt põhineb loogilistel alustel ja omab vaba tahet. Nõrk tehisintellekt on erinevatel vahenditel (nt nagu masinõppes kasutatavad tehisnärvivõrgud, geneetilised algoritmid, evolutsioonilised meetodid) põhinev tehisintellekt, mis küll opereerib teadmistega, kuid ei oma nende teadmistega seonduvat arusaamist ega vaba tahet.

Nõrgal tehisintellektil on küll teatav autonoomia[1], kuna selle poolt teadmistega tehtavad operatsioonid ja otsused ei ole inimese poolt ette kirjutatud (vt alapeatükk 1.3.2). Nõrga tehisintellekti aluseks on valdavalt statistilistel meetoditel toimivad andmetöötluse tehnoloogiad, millel puudub tõsiseltvõetav püüdlus luua uut, iseteadlikku mõistust või mõtlemisvõimet. Seetõttu ei saa nõrgale tehisintellektile vaba tahet omistada.

Käesolev töö tegeleb nõrga tehisintellekti autonoomiast tulenevate eetiliste ja sotsiaalsete riskidega, kuna nõrk tehisintellekt on leidnud laia kasutamist ning sellel põhinevate infosüsteemide käitamine on hakanud inimestele ning ühiskonnale avaldama laiaulatuslikku mõju. Tugevat tehisintellekti ei ole praeguse tehnika taseme juures suudetud tõenäoliselt veel luua [24, lk 9].

Vaba tahte puudumise tõttu ei ole nõrga tehisintellekti põhjustatud mõju võimalik eetiliselt hinnata. Küll aga on võimalik eetiliselt hinnata nõrga tehisintellekti süsteemi arendaja või käitaja poolset põhjustamist, milles nõrk tehisintellekt osaleb põhjustamise

[1] AIS-i autonoomia definitsiooni osas puudub kirjanduses üksmeel. IEEE soovitab autonoomiat määratleda läbi vaba tahte vaatenurga [1, lk 195]. Autori ise eelistab autonoomia mõiste sisustamisel aga antropomorfust kõrvalejätvaid käsitlusi, mis loevad autonoomseteks selliseid süsteeme, mis suudavad süsteemi toimimise ajal toimuvate ootamatute sündmuste mõjul oma toimimist ise muuta [142, lk 368].

vahendina. Seetõttu on nõrgale tehisintellektile esitatud eetilised ja sotsiaalsed nõuded sisuliselt nõuded süsteemi arendaja või käitaja tegevusele.

1.3 Autonoomsed ja intelligentsed süsteemid

AIS tähistab antud töös nõrka tehisintellekti (sageli masinõppekomponent) sisaldavaid inimese suutlikkust toetavaid riist- ja tarkvarasüsteeme [1, lk 12], mis teadmistega opereerides toetavad või teostavad otsuse langetamise protsessi[1].

AIS-ile on omane teatav autonoomsus ja intelligentsus, mis tuleneb enamasti süsteemis kasutatavast valdavalt masina loodud algoritmist (nt masinõppealgoritm). Kuna AIS-ide spetsiifilised riskid tulenevad just sellistest algoritmidest, siis on vajalik mõista valdavalt inimese loodud algoritmide ja valdavalt masina loodud algoritmide omadusi ja nende erinevusi inimesele arusaadavuse ja kontrollitavuse osas.

1.3.1 Valdavalt inimese loodud algoritmid

Algoritm on sammupõhine tegevusjuhis või eeskiri, mis määrab teatavad liiki ülesannete lahendamiseks, tegevuse sooritamiseks või eesmärgi saavutamiseks vajalikud operatsioonid ja nende järjekorra [25, lk 112], [26].

Algoritmide definitsioone on mitmeid (Turingi masin, Markovi normaalalgoritm, Churchi operaatorterm jt [27, lk 169–200]), kuid põhisisult loetakse neid definitsioone samaväärseteks [25, lk 112], [28, lk 44], [29, lk 16]. Käsitletud kirjanduse alusel [21, lk

[1]Autori hinnangul võtab see mõiste kõige paremini kokku nõrka tehisintellekti praktikas kasutavate süsteemide iseloomulikud omadused (mõningane autonoomsus teadmistega opereerimisel) ilma tehisintellekti mõistega sageli kaasneva antropomorfistlike konnotatsioonita. Autor möönab, et AIS-ide seostamisel mütoloogiast (Golem, kratid jt) või *sci-fi* kirjandusest (Frankenstein, Terminator, R2D2 jt) tuttavate, iseseisva mõttemaailma, tunnete ja vaba tahtega tehislike subjektiga on kahtlemata valdkonnale tähelepanu tõmbav ja turunduslik väärtus. Samas puudub AIS-ide praktilise rakendamise vahenditel täna veel tõsiseltvõetav võimekus luua sellist vaba tahet omavat subjekti, keda tavakeeles intellekti ja selle evijatega seostatakse. Seetõttu võib tehnoloogia käsitlemine antropomorfistlike võtetega takistada tehnoloogia tegeliku olemuse mõistmist ning tehnoloogiliste riskide adekvaatset haldamist ühiskonnas. Nii on juba mõningates regulatiivse iseloomuga algatustes kõlanud eetilistel dilemmadel rajanevad üleskutsed kehtestada masinõppe-põhise „tehisintellekti“ arendamisele moratoorium [143]. Teisalt on erialakirjanduses osundatud, et lähituleviku seisukohalt ebarealistlike ohtude kirjeldamisest tulenev ühiskondlik surve AIS-e õiguslikult reguleerida võib läbi seadusandlike piirangute või keeldude muutuda valdkonna arengu suurimaks takistuseks [40].

228], [25], [29], [30, lk 79], [31], [32] toob autor välja järgmised algoritmi mitteformaalsed põhitunnused, mis on piisavad algoritmi olemuse mõistmiseks:

1. Diskreetsus. Algoritmi üksikud tegevussammud, tegevuste objektid (sisendid, väljundid) ja tegevuste sooritamise ajad peavad olema võimalikult selgelt määratletud ja eristatavad nende tegevuste järjestuses [33, lk 3];

2. Elementaarsus. Algoritmi iga samm ja selle kirjeldus peab olema võimalikult lihtne [34, lk 5–6];

3. Direktiivsus. Algoritmi iga sammu puhul peab olema selge, kas see lõpetab algoritmi tegevuse või kui algoritm jätkab, siis milline on järgmine tegevussamm [21, lk 228];

4. Determineeritus. Algoritmi iga sammu puhul peab tulemus olema võimalikult üheselt määratud sellega, millest vastava sammu sooritamisel lähtutakse [34, lk 5–6];

5. Reprodutseeritavus. Algoritmi sammud ja nende kogumid peavad olema korratavad [33, lk 3].

Traditsiooniliselt on algoritm valdavalt infosüsteemi arendavate inimeste looming – inimesed defineerivad algoritmi eesmärgi, algoritmi sammud ja nende sisu, algoritmi sisendid ja väljundid. Selline valdavalt inimese loodud algoritm on allutatud inimese kontrollile ja sellel puudub autonoomia. Valdavalt inimese loodud algoritm on (teisele) inimesele arusaadav ja selle toimimise reeglid on (teiste) inimeste poolt kontrollitavad.

Samas ei saa kõiki algoritmide rakendamisi pidada automaatselt eetiliselt või sotsiaalselt neutraalseteks vaid põhjusel, et nad tegutsevad andmete ja loogiliselt ja matemaatiliselt määratletud reeglite alusel. Esiteks võivad algoritmid olla nende arendajate ja käitlejate poolt teadlikult kallutatud eelistama neile sobivaid otsuseid või soovitusi[1]. Teiseks

[1] Näiteks laskis American Airlines (AA) juba 1960. aastal koos IBM-iga käiku lennupiletite reserveerimise süsteemi SABRE, mis oli teadlikult kallutatud – süsteem pakkus pileti ostjale esimesena AA enda kallimaid lennupileteid, sest AA oli märganud, et kasutajad tegid valiku esimeste valikute seast. Järgnenud uurimises AA isegi ei varjanud, et süsteem oli üles ehitatud eelistama AA kallimaid pakkumisi [[49, p. 2] kaudu [145]]. See näitab, et algoritmide kallutamine nende käitajate huvides on peaaegu sama

võivad algoritmide autoriks olevad inimesed algoritme kallutada ka tahtmatult[1]. Kolmandaks võib kallutatus esineda ka algoritmi töödeldavates andmetes. Algoritmi toimimiseks tuleb algoritmidega töödeldavad füüsikalised (nt vihm, päiksepaiste, värvus) ja sotsiaalsed fenomenid (nt sugu, rass, rahvus) teisendada töötlemiseks sobivasse vormi (sisendid) ja töötluse tulemuste kasutajatele sobivasse vormi (väljundid). Selline teisendamine ei ole autori hinnangul eetiliselt ja sotsiaalsete neutraalne, kuna teisendamine on mõjutatud seda läbiviiva inimese sotsiaalsest ja kultuurilisest hinnangutest. Sellised otsused ei ole faktiotsused („on“ vs „ei ole“), vaid tegemist on hinnangu omistamise ehk väärtusotsusega[2]. Vastupidine seisukoht on klassikaline naturalistlik eksitus[3].

Keerulisemate inimese loodud algoritmide puhul võib algoritmi mõistmist ja kontrollimist piirata algoritmi keerukus (mis ei seisne ainult sammude ja nendevaheliste suhete rohkuses), kuid kuna ka sellised algoritmid on siiski veel loodud inimese poolt, siis on ka nende kontrollimine (vähemalt teoreetiliselt) ikkagi inimesele jõukohane – vajadusel tuleb suurendada algoritmi mõistmises ja kontrollimises osalevate inimeste arvu (praktikas ei pruugi aga nõutava arvu kvalifitseeritud inimeste ning tööaja leidmine olla võimalik).

---

vana kui infosüsteemide kommertskasutus. 1984. aastal otsustas Ameerika Ühendriikide lennuamet siiski, et AA peab tegema SABRE otsingualgoritmi avalikuks, so avaldama kasutajatele piletite soovitamise ja sorteerimise aluseks olevad kriteeriumid ning neile omistatud kaalud [144, lk 3]. Seda otsust loetakse algoritmi läbipaistvuse õigusliku reguleerimise alguseks.

[1] Mõnede autorite arvates on valdavalt inimese loodud algoritmid vältimatult laetud nende autoriks olevate inimeste väärtustega ja tahestahtmata kallutatud – autoriks olevad inimesed määravad algoritmid toimimise reeglid, pidades silmas nende poolt soovitud eesmärke, mis vältimatult lähtuvad teatud huvidest ja seega paratamatult eelistavad teatud huvisid teistele huvidele ja teatud inimgruppe teistele gruppidele [26], [54], [69].

[2] Näiteks võib tuua inimese soo määratlemise. Traditsiooniliselt on küsimus lihtne – inimene on naissoost (1) või meessoost (0). Seega on reeglina soo teisendamine arvutatavasse vormi lihtne – saab omistada väärtuse 1 või 0. Kui aga asuda hindama, kas ja milliste kriteeriumite alusel sellesse süsteemi sobitada erinevaid sooga seonduvaid fenomene (transeksuaalsus, hermafrodiitsus jms), siis on tegemist keeruliste ja hinnanguliste otsustega, mis sõltuvad nii otsusest mõjutatud valdkonnast (nt e-valimiste puhul ei ole valija sugu oluline, kuid sportlase soo selge määratlemine on võistluse aususe huvides väga vajalik) kui ka otsustaja väärtussüsteemist (nt kristliku eetika jaoks on sugu rangelt binaarne fenomen, samas kui liberaalne eetika võib aktsepteerid “kolmandat”, määratlemata vms sugu). Probleemi keerukust ja praktilisust illustreerivad IAAF-i pingutused kergejõustiklase Caster Semenya soo määratlemisel [146].

[3] Naturalistlik eksitus on väärtusotsuse samastamises faktiotsusega, s.o “viga, mis seisneb eetilise e[hk] väärtuselise mõiste samastamises naturaalse mõiste e[hk] faktilise kirjeldusega” [147, lk 53].

Siiski realiseerib inimese loodud reeglite alusel toimiv algoritm selle loonud inimese vaba tahet, looja tahte realisatsioon on algoritmi reeglite kaudu kontrollitav ning algoritmi kaudu toimuvale põhjustamisele on võimalik anda eetilist hinnangut. Ka saab jaatada algoritmi loonud inimese vastutust muude sotsiaalsete nõuete täitmise eest.

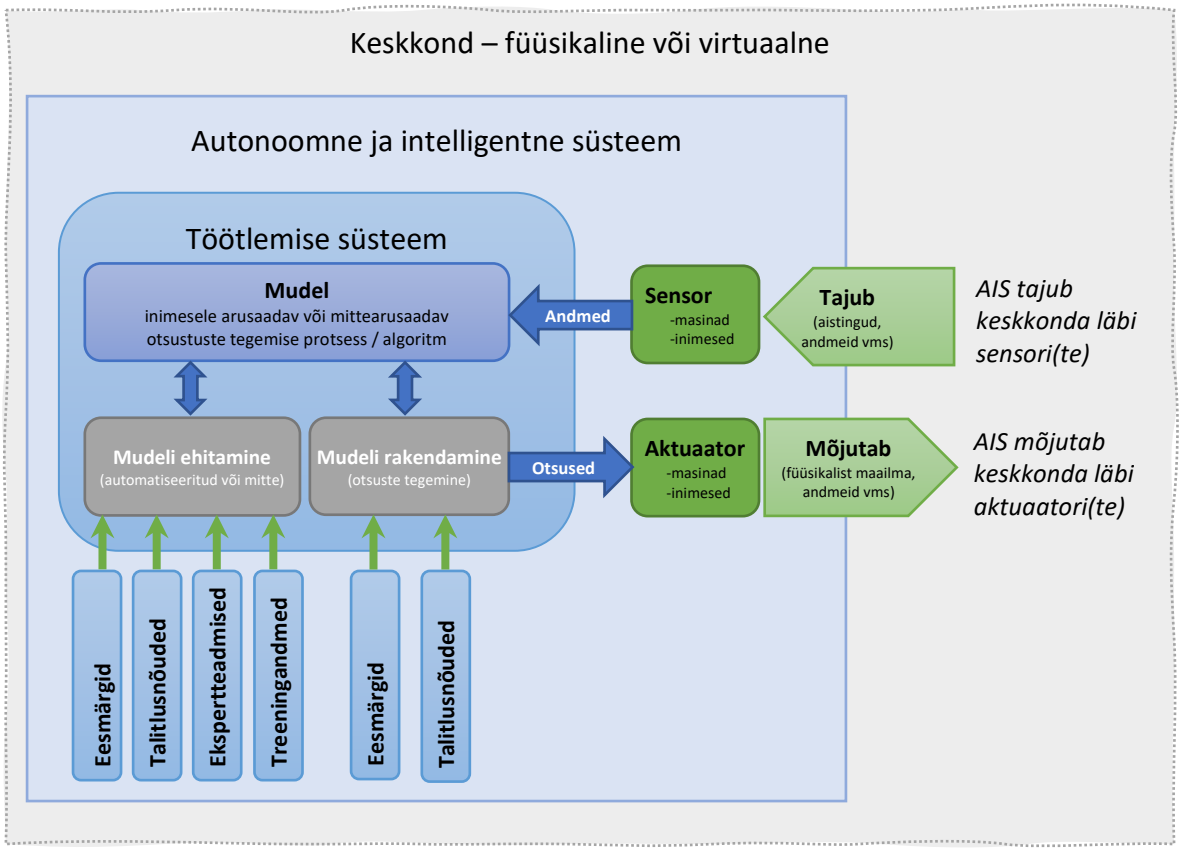
1.3.2 Autonoomsed ja intelligentsed süsteemid

AIS-il puudub ühtne ja üldtunnustatud definitsioon [35, lk 7]. Kirjanduses hõlmab vastav fenomen mitmekesiseid tehnoloogiaid, mis omavad erineval tasemel autonoomiat ja tehisintelligentsi. Fenomeni kirjeldab autori hinnangul piisavalt hästi AI HLEG definitsioon [36, lk 44]:

> *„[AIS-id][1] on inimese projekteeritud tarkvaralised ja võimalik et ka riistvaralised süsteemid, millele antakse keerukas eesmärk ja mis toimivad füüsilises või digitaalses mõõtmes, tajudes oma keskkonda andmehõive abil, tõlgendades kogutud struktureeritud ja struktureerimata andmeid, tehes teadmuse põhjal järeldusi või töödeldes kõnealustest andmetest saadud teavet ja otsustades, millised on antud eesmärgi saavutamiseks parim(ad) toiming(ud). [AIS-id] võivad kasutada kas sümbolreegleid või õppida selgeks arvmudeli ning samuti saavad nad kohandada oma käitumist, analüüsides seda, kuidas nende varasemad toimingud mõjutavad keskkonda.“*

AIS koosneb kolmest põhikomponendist – (1) sisendite allikaks olevad sensorid, (2) AIS töötlemise süsteem ning (3) väljundit võimaldavad aktuaatorid. AIS-i sensoriteks ja aktuaatoriteks võivad olla nii tehnilised seadmed kui ka inimesed [14, lk 7]. Inimesed suhtlevad AIS-iga läbi kasutajaliidese [23, lk 15]. AIS võtmekomponendiks on töötlemise süsteem (nt masinõppealgoritm), mis algoritmi ja sensoritelt kogutavate sisendite abil annab väljundi aktuaatoritele (soovituste, ennustuste või otsustena), mis mõjutavad seeläbi keskkonda – STS-i või selle laiemaks keskkonnaks olevat ühiskonda [14, lk 6]. OECD on AIS-i kontseptuaalse skeemi esitanud järgmisena:

[1] AI HLEG kasutab mõisted „tehisintellekti süsteemid“. Käesoleva töö terminoloogilise ühtsuse huvides kasutab autor läbivalt IEEE eelistatud mõistet „AIS“.



Joonis 1. AIS-i kontseptuaalne skeem [14, lk 6].

AIS-ile autonoomiat ja tehisintelligentsi omadusi andvat AIS-i töötlemise süsteemi saab realiseerida erinevate vahendite abil. Käesoleval ajal on AIS-i töötlemise süsteemi realiseerimisel enimlevinud erinevate masinõppetehnoloogiate ja -algoritmide kasutamine [1, lk 43], [23, lk 9–10].

Masinõppe idee on pärit arvutiteadlaste unistusest luua kasutamise käigus iseenda töö tulemuslikkust parandav, nö iseõppiv arvutiprogramm [37, lk xv, 1]. Esialgsed püüdlused intelligentseid süsteeme luua olid üles ehitatud „if-then“ põhiste inimese poolt determineeritud teadmiste süsteemidel, mis üsna kiiresti ilmutasid oma ebaefektiivsust ning tõid kaasa entusiasmi raugemise 1970. aastatel kiiresti arenenud uurimisvaldkonnas [38, lk xi]. Masinõppe valdkonnale andis uue hingamise 1983. aastal ilmunud kogumik *Machine Learning: An Artificial Intelligence Approach* [39] ning 1997. aastal T. Mitchelli poolt avaldatud käsitlus *Machine Learning*, mis püüdsid vastata küsimusele „*kuidas luua arvutiprogramme, mis muutuvad automaatselt paremaks kogemuse omandamise kaudu*“ [37, lk xv]. Suurandmete levik on 2010. aastatel andnud masinõppe arengule uue hoo, kuna masinõpe on muutunud

huvipakkuvast uurimissuunast paratamatuks vajaduseks, et mõista ja hankida uusi teadmisi inimestele hoomamatust kogusest andmetest[1].

Masinõppe formaalne definitsioon pärineb Tom M. Mitchellilt [37, lk 2]:

> *„Arvutiprogrammi kohta saab öelda, et see õpib kogemusest E teatud liiki tegevuste T suhtes ja tulemuslikkusega P, kui programmi tulemuslikkus tegevuses T mõõdetuna P-s paraneb läbi kogemuse E.“*

Masinõppe uurimisvaldkonna eesmärgiks on luua programmid, mis püüavad ilma inimese sekkumiseta või võimalikult vähese inimese sekkumisega ise andmetest õppida ja seeläbi oma toimimist parandada selle asemel, et tugineda inimeste poolt eeldetermineeritud ja programmeeritud programmidele ja nendes sisalduvale teadmisele [40, lk 44], [41], [42, lk 3]. Masinõppel on erinevaid liike. Juhendatud õppe („*supervised learning“*) korral õpib algoritm valdkonna fenomene eristama inimese poolt eelsorteeritud andmete alusel. Juhendamata õppe („*unsupervised learning“*) korral leiab algoritm andmetest seoseid iseseisvalt. Stiimulõppe *(„reinforcement learning“*) korral õpivad algoritmid eelistama teatud tulemusi vastavalt eelnevalt määratud edukuskriteeriumitele või stiimulitele [4, lk 830–831], [43, lk 5], [44].

Kõigi nende lähenemiste puhul loovad masinõppevahendid ise algoritmi, s.o. reeglid, mida järgides langetab algoritm otsuseid uute sisendandmete alusel [42, lk 3]. Sõltuvalt kasutatud tehnoloogiast, võib arusaamine selliselt loodud algoritmi reeglitest ja nende muutumisest olla inimesele ülikeeruline või praktiliselt võimatu[2].

---

[1] 2009. aastal hindas IBM, et iga päev loodi maailmas 2,5 kvintiljonit baiti andmeid. Aastatel 2010-2016 loodi maailmas 90 % kõigist selleks ajaks inimkonna ajaloos loodud andmetest [56, lk 25]. IDC nimetab kogu inimkonna loodud, kogutud ja paljundatud andmehulka Globaalseks Andmesfääriks (*Global Datasphere*). IDC hindab, et Globaalne Andmesfäär kasvab 2018. aasta 33 zettabaidilt 2025. aastaks 2025 zettabaidini [148, lk 5].

[2] Inimesele mõistetavuse ühes otsas on otsustuspuu ja „*if-then*“ lausete põhised lahendused. Masinõppe tõeline võimekus ja täpsus avaldub aga inimesele mittearusaadavate sügavate närvivõrkude („*deep learning“*) toimimises. Moodsad masinõppe tehnoloogiad (otsustusmetsad, tugivektor-masinad, närvivõrgud, sügavad närvivõrgud ning nende meetodite kombinatsioonid[2]) on küll täpsemad, kuid nende otsustusprotsess ei ole inimesele jälgitav ega arusaadav või on seda väga piiratult [46]. Eriti keeruline on närvivõrkude ning sügavõppe otsustusprotsessi inimesele arusaadavus ja selgitamine [45, lk 6–7]. Seega on praeguse tehnoloogia arengu puhul suurema täpsuse saavutamiseks vajalik teha kompromisse otsustusprotsessi inimestele arusaadavuse arvelt.

Masinõppealgoritmi võime andmetest teadmisi omandada ning nende alusel oma toimimise loogikat muuta annabki algoritmile ja seeläbi AIS-ile teatava autonoomia. AIS-i tegevuse käigus toimuva otsuste vastuvõtmise protsessi muutumise tõttu ei ole AIS-i otsused süsteemi arendava inimese poolt eeldetermineeritud ega allu inimese vahetule kontrollile. Algoritmi otsuste vastuvõtmise protsessi keerukuse ja selle muutumise kiiruse tõttu ei ole see protsess alati inimese poolt mõistetav ega inimesele selgitatav [45, lk 6–7], [46].

AIS-i spetsiifiliste eetiliste ja sotsiaalsete riskide põhjused peituvadki just eelkirjeldatud autonoomias, selle raames toimuva algoritmi toimimise süsteemi muutumise inimesele arusaadavuse komplitseerituses ja sellest tulenevas süsteemi looja poolt seatud eesmärkide saavutamise ebakindluses [42, lk 3–4]. Eetiliste ja sotsiaalsete riskide tekitamise tõttu takistab sama autonoomia ka masinõppealgoritmi tehnilise, spetsifikatsioonile vastava toimimise eelkontrolli.

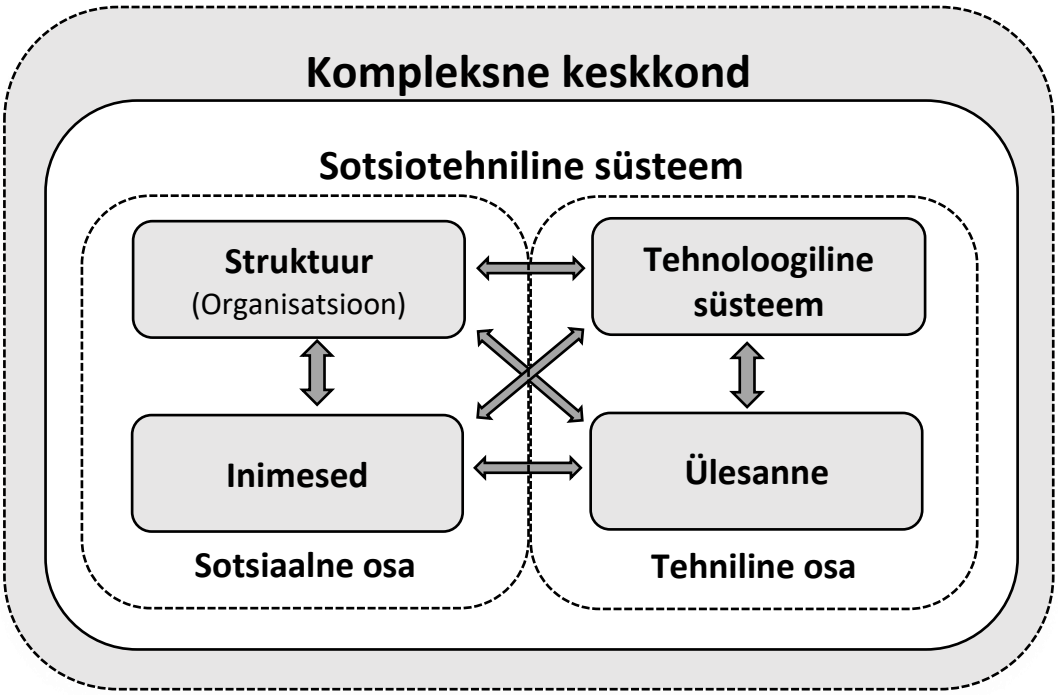
1.4 AIS osana sotsiotehnilisest süsteemist

Magistritöö esemeks on eetilised ja sotsiaalsed nõuded inimesi ja ühiskonda mõjutavale AIS-le. Selline AIS osaleb tehnilise komponendina sotsiotehnilises süsteemis (STS).

STS on selline süsteem, mis omab järgmisi põhitunnuseid:

- STS-il on üksteisest sõltuvad ja omavaheliste vastastikmõjudega komponendid;
- STS tegutseb, kohandub ja püüab oma eesmärke saavutada teatud väliskeskkonnas ja on vastastikuses mõjus selle väliskeskkonnaga [7, lk 5];
- STS-il on oma sisekeskkond, mis koosneb eraldiseisvatest, kuid üksteisest sõltuvatest tehnoloogilistest ja sotsiaalsetest alamsüsteemidest (inimesed, organisatsiooni struktuur) [6, lk 120];
- STS-i samu või samaväärseid eesmärke on võimalik saavutada mitmel erineval viisil, mida tuleb arvestada süsteemi kavandamisel;
- STS-i jõudlus sõltub tehniliste ja sotsiaalsete alamsüsteemide ühisest optimeerimisest [47, lk 201] [48], kusjuures ühe alamsüsteemi eelistamine teisele võib viia süsteemi jõudluse ja tulemuste halvenemisele [7, lk 5].

Seetõttu ei saa STS toimimist parandada vaid uute tehnoloogiate süsteemi lisamisega, vaid analüüsida ja optimeerida tuleb STS-i sotsiaalseid ja tehnilisi alamsüsteeme nende koosmõjus [49, lk 30] ja vastastikkuses mõjus STS-i ümbritseva keskkonnaga. Infotehnoloogilise komponendiga STS-i kontseptuaalne skeem on järgmine [50, lk 3]:



Joonis 2. Sotsiotehniline süsteem [50, lk 3].

Kui STS tehniline alamsüsteem hõlmab endas AIS komponenti, siis tuleb vastavale komponendile esitatavate nõuete analüüsimisel ning täitmise hindamisel arvestada ka STS-i sotsiaalse komponendi (inimesed, organisatsioon) ja ühiskonna kui keskkonna ootuste, võimaluste ja vajadustega [20, lk 5.4.1]. Seeläbi ulatuvad ühiskonna eetilised ja sotsiaalsed nõuded ka STS süsteemiks oleva AIS komponendini.

Sealjuures tuleb arvestada, et STS keskkond ning STS ise on pidevas muutumises. Seetõttu peab analüütik autori hinnangul suutma AIS kavandamisel ja käitamisel hinnata AIS komponendi eetiliselt ja sotsiaalselt hinnatavaid mõjusid STS-ile ning ühiskonnale koos nende mõjude muutumisega ajas. Vastav ootus kasvab koos konkreetse STS-i poolt inimestele ja ühiskonnale avalduva mõju kasvuga.

Peatüki kokkuvõttena järeldab autor, et tänase tehnika arengu juures tuleb AIS-idele esitatud eetilisi ja sotsiaalseid nõudeid valdavalt käsitleda kui nõudeid süsteemi arendajale või käitajale. AIS-ide mõningane autonoomsus ja intelligents raskendab küll

AIS-ide toimimisest arusaamist ja kontrolli AIS otsuste üle, kuid see ei ole piisav selleks, et vabastada AIS-i arendaja ja käitaja eetilisest ja sotsiaalsest hinnangutest ja vastutusest AIS-i poolt inimestele ja ühiskonnale põhjustatud mõjude eest.

Seetõttu peavad arendajad ja käitajad AIS-i mõistma süsteemi käitamise STS-ist ja keskkonnast tulenevaid eetilisi ja sotsiaalseid nõudeid vastavale süsteemile. Kuna erinevates keskkondades võivad olla erinevad eetilised ja sotsiaalseid nõudeid, siis tuleb neid nõudeid ja süsteemi nõuetelevastavust uuesti analüüsida süsteemi rakendamisel igas konkreetses keskkonnas. Eraldi väljakutseks on eetiliste ja sotsiaalsete nõuete ja AIS-i toimimise muutumine ajas, mistõttu tuleb ka pärast nõuetelevastava süsteemi rakendamist tagada nõuete püsimise ja nõuetelevastavuse kontroll.

2 AIS-iga seonduvad eetilised ja sotsiaalsed riskid

Käesolevas peatükis annab autor ülevaate AIS-iga seonduvate eetiliste ja sotsiaalsete probleemide ja riskide liikidest. Probleemide ja riskide mõistmine aitab analüütikutel valida sobivad meetodid või vahendid vastavate riskide haldamiseks AIS-i arendamise või käitamise sobivas etapis.

2.1 AIS-i eetilist ja sotsiaalset mõju omavate probleemide allikad

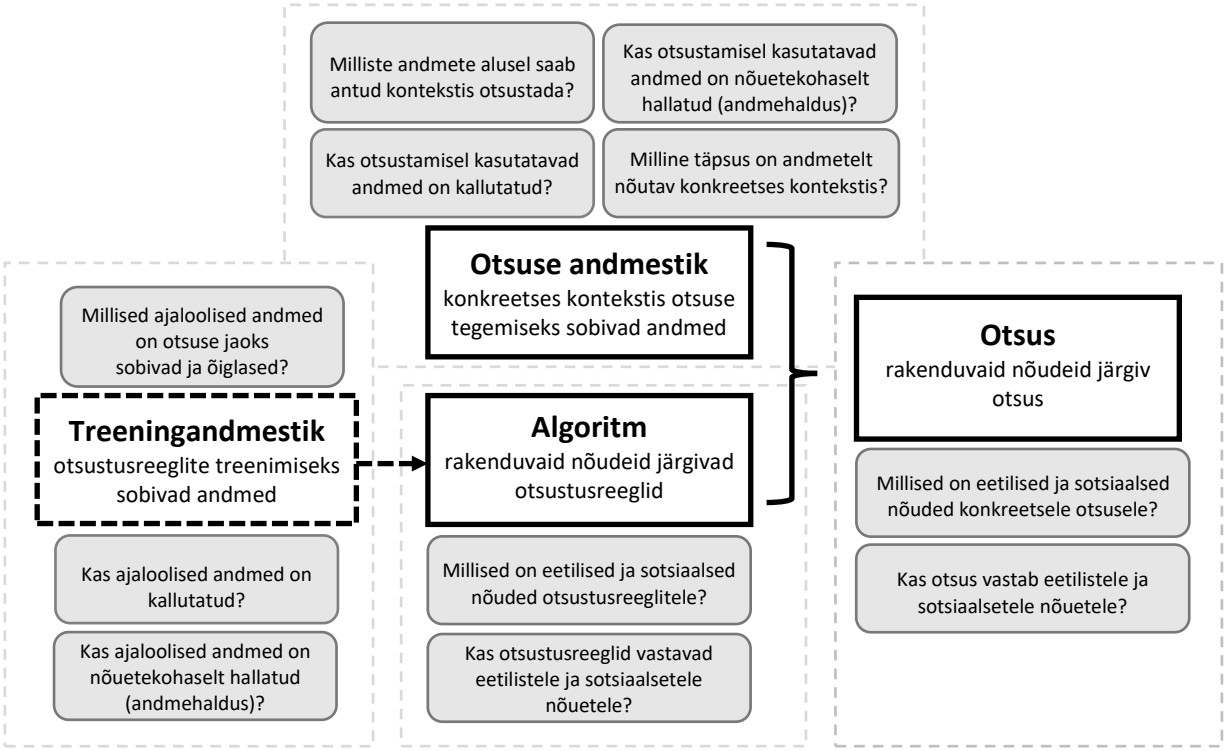
AIS rakendamisest oodatakse ühiskonnale suurt kasu läbi suutlikkuse analüüsida inimesele hoomamatuid andmehulkasid ja märgata neist inimesele märkamatuks jäävaid seoseid ja seaduspärasid järjest kiiremini ja madalama kuluga [51, lk 11], [52, lk 1252–1253]. Oodatakse ka, et formaalloogiliste reeglite, teadmiste ja andmete alusel töötavad AIS-id suudavad teha ekslikest ja kallutatud inimestest paremaid otsuseid.

Teadusuuringud tõepoolest kinnitavad, et AIS-id ja infosüsteemid eksivad otsuste ja soovituste tegemisel vähem kui inimestest eksperdid [51, lk 11]. Samas tuuakse kirjanduses välja mitmeid olukordi, kus AIS-i käitajad on põhjustanud tagajärgi, mis on ebaeetilised või sotsiaalselt taunitavad [53, lk 8–9, 106]. AIS-i võimsusest tulenevalt võivad ühe süsteemi ebaõiged otsused omada ülisuurt eetiliselt hinnatavat mõju ühiskonnale [54, lk 846]. Mitmed rahvusvahelised organisatsioonid, standardiasutused ning valdkonna praktikud on samuti möönnud AIS kaasnevate eetiliste ja sotsiaalsete riskide olemasolu ning nende haldamise vajadust (vt alapeatükk 3). Käesolevas punktis selgitab autor AIS-iga seonduvate eetiliste ja sotsiaalsete riskide allikaid ning nende avaldumise vorme.

Esimeseks AIS-i eetiliste ja sotsiaalsete riskide allikaks on põhjendamatu eeldus, et erinevalt inimese otsustest puudub selliste infosüsteemide otsustes automaatselt kallutatus[1] [45, lk 3–4]. Ehkki AIS-ide kasutamist õigustatakse sageli väitega, et sellised

[1] Infosüsteemi kallutatus („*bias*“) on süsteemi omadus diskrimineerida teatud isikuid või isikute gruppe võrreldes teistega ebaõiglaselt ja süstemaatiliselt. Ebaõiglase diskrimineerimisega on tegemist juhul, kui süsteem jätab isiku või isikute grupi ilma hüvest või soodsast otsusest või teeb isiku suhtes negatiivse mõjuga otsuse põhjusel, mis on ebamõistlik ja sobimatu. Süsteem on kallutatud, kui ebaõiglased otsused

süsteemid on objektiivsed ning vabad inimesele omastest eelarvamustest ja kallutatusest, näitavad arvukad uuringud siiski seda, et erinevatel põhjustel on eelarvamused ja kallutatus omane ka AIS-ile [55]–[58]. AIS-i poolt otsuste tegemiseks kasutava andmestiku, algoritmidele tugineva töötlemise süsteemi ja selle alusel langetatud otsusega kaasnevad mitmed eetilise ja sotsiaalse dimensiooniga väärtusotsused ja hinnangud. Nende tõttu avaldab isegi neutraalsena kavandatud AIS eetiliselt või sotsiaalselt hinnatavat mõju. AIS-i kavandamisel ja käitamisel tehtavate väärtusotsuste ja hinnangute illustreerimiseks sobib järgmine joonis [54, lk 842]:



Joonis 3. Eetilised ja sotsiaalsed valikud AIS-i arendamisel ja käitamisel [54, lk 842].

Infosüsteemide otsuste kavandatud kallutatuse kõrval on keerulisemaks probleemiks AIS-i arendaja või käitaja poolt kavandamata ja märkamata kallutatus, so infosüsteemi toimimine mittekavandatud viisil. Mittekavandatud viisil toimimise risk on AIS puhul kõrgendatud, sest inimvõimete piiratuse tõttu on inimesel keeruline aru saada valdavalt masina loodud algoritmide ülesehitusest ja tööst ning märgata ja vältida sellise algoritmi

on korduvad ja süsteemi toimimist iseloomustavad; ühekordne süsteemi viga ei ole käsitletav kallutatusena [2, lk 332–333].

kallutatust, sotsiaalselt kohatut otsustusprotsessi[1] või kasutatud (suur)andmete kallutatust või ebakvaliteetsust. AIS-i töötlemise süsteemi piiratud mõistmise ning süsteemi eetiliste mõjude hindamise väljakutsed muutuvad järjest suuremaks vastavalt sellele, kuidas kasvab AIS-ide keerukus ja mõju [59, lk 87]. Olukorras, kus AIS arendaja ja käitaja peaks olema teadlik AIS-iga kaasnevatest riskidest, aga jätab need riskid haldamata, on AIS poolt inimesele või ühiskonnale põhjustatud mõju eetiliselt ja sotsiaalselt hinnatav ja negatiivne hinnang on omistatav AIS-i arendajale või käitajale.

Teiseks AIS probleemide allikaks on masinõppe algoritmidest tulenev autonoomia ning selliste algoritmide otsustuste vastuvõtmise protsessi inimese poolt kontrollimise keerukus või võimatus. Masinõppealgoritmid on sedavõrd keerulise ülesehitusega, et inimesel puudub reaalne võimalus jõuda selgusele masinõppe algoritmi toimimise sammudes ja nende põhjendatuses [40, lk 57]. Kirjanduses nimetatakse masinõppe algoritme „mustadeks kastideks“, mis iseloomustab nende põhilist erinevust valdavalt inimese loodud algoritmidest: masinõppe puhul on inimesel võimalik tajuda masinõppealgoritmile antud sisendeid ja selle väljundeid, kuid masinõppealgoritmi enda valitud sisendid, neile inimese otsese abita valitud osatähtsused ning sisenditest väljunditeni jõudmiseks algoritmi enda moodustatud reeglid jäävad inimestele raskesti hoomatavateks või arusaamatuteks.

Kolmandaks loovad eetilisi ja sotsiaalseid probleeme AIS-ide ja masinõppe statistikal ja tõenäosusteoorial põhinevad töömeetodid. Selles kontekstis on põhiküsimuseks, millistel tingimustel on eetiline ja sotsiaalselt lubatav teha konkreetset inimest mõjutavaid otsuseid (a) laiema inimgrupi käitumist iseloomustavate või (b) inimese mineviku käitumist iseloomustavate andmete ja järelduste alusel. Inimväärikuse ja seda kaitsva eetika[2] seisukohast on masinõppe otsustega kaasnevad riskid sarnased statistilise otsuste alusel otsustamise riskidega [60, lk 1411] – mõlemad kätkevad ohtu, et isikut diskrimineeritakse teda mõjutavate otsuste tegemisel tema grupitunnuste või tema mineviku käitumise alusel.

---

[1] Sotsiaalselt kohatu on selline otsustusprotsess, mis võtab inimest või ühiskonda mõjutava otsuse tegemisel aluseks sotsiaalselt ebakohased andmed (nt tundlikud tunnused või käsitletava sotsiaalse fenomeniga mitteseonduvad andmed) või annab kasutatud andmetele sotsiaalselt ebakohase osatähtsuse.

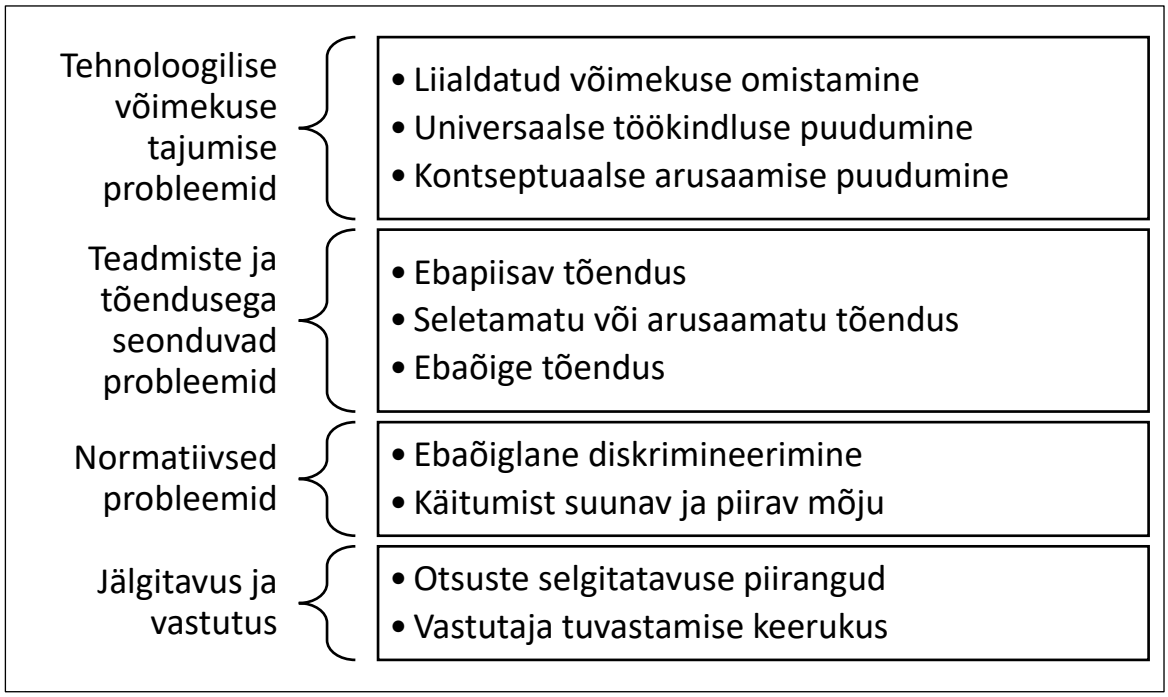
[2] Inimväärikust kaitsvast eetikast lähtuvad kõik käesoleva töö peatükis 3 käsitletavad AIS eetilised raamistikud.

Ehkki käesolevas töös kirjeldatud meetodid – eriti eetilise AIS-i elukaare haldus (vt alapeatükk 3.4) – võimaldavad hallata ka AIS-i teadliku kallutamise ja tehnoloogia kuritarvitamise riski, on magistritöö käsitluse keskmes just selliste riskide teadvustamine ja haldamine, mis põhjustavad süsteemi arendaja või käitaja poolt kavandamata eetiliselt ja sotsiaalselt hinnatavaid tagajärgi.

2.2 AIS-i eetilise ja sotsiaalse mõjuga probleemid

Kirjanduses on AIS-i ja valdavalt masinate loodud algoritmiga seonduvaid eetilisi ja sotsiaalseid probleeme välja toodud mitmeid. Probleemidel on erinevad allikad, olemus ning erinev ilmnemise koht AIS-i elukaarel. Analüütikud peavad probleemide liike, olemust ja põhjusi mõistma selleks, et hallata vastavaid riske selleks sobivate meetodite ja vahenditega AIS-i elukaare õiges etapis.

Ülevaatlikult on AIS-i valdavalt masinate loodud algoritmide, konkreetsemalt masinõppealgoritmide kasutamisega kaasnevad eetilised probleemid kaardistanud Mittelstadt jt. Selle ja teiste käsitluste alusel [40, lk 31–39], [42, lk 4] on autor koostanud järgmise ülevaate valdavalt masina loodud algoritmide kasutamisest tulenevatest eetilise ja sotsiaalse mõjuga tüüpprobleemidest:



Joonis 4. AIS-i eetilise ja sotsiaalse mõjuga tüüpprobleemid [40], [42].

Eelneval joonisel esitatud eetilist ja sotsiaalset mõju omavate tüüpprobleemide olemus ja allika on alljärgnevad:

1) **Liialdatud võimekuse omistamine** *(„fundamental overattribution error“).* Inimesed kipuvad AIS-e usaldama rohkem kui see on AIS-i tänaste võimekuste juures põhjendatud [40, lk 18–21]. Autori hinnangul on üheks liialdatud ootuse väljundiks ka uskumus, et AIS-ide otsused ja soovitused on „objektiivsed“ või „neutraalsed“ üksnes seetõttu, et tegemist on „andmepõhiste“ otsustega.

2) **Universaalse töökindluse puudumine** *(„lack of reliabiliy“).* AIS-id annavad üsna häid tulemusi teatud olukorras. Sama ja sarnase andmestiku alusel võib masinõppealgoritm toimida väga hästi. Samas võib algoritmi täpsus halveneda kohe, kui otsuseid tuleb teha erinevate sisendandmete alusel[1]. Valdavalt masina loodud algoritmi inimese poolt arusaadavuse ja kontrollitavuse piiratuse tõttu ei ole inimestel võimalik ka eelnevalt kontrollida, millistes olukordades algoritmi kasutav AIS töötab nõuetekohase tulemuslikkusega ja millal mitte [40, lk 57].

3) **Kontseptuaalse arusaamise puudumine**. Kuna masinõppealgoritmi jõuab sisendandmetest otsuseni olulisel osal statistilistel meetoditel, siis ei pruugi algoritmil andmete alusel tekkida kontseptuaalset teadmist ja arusaamist andmetega kirjeldatud fenomenist [40, lk 90]. See halvendab AIS-i suutlikkust toimida erinevates või piirolukordades. Samal ajal ei pruugi masinõppealgoritmidel tekkida ka kasutaja poolt soovitud lahenduse kontseptuaalse mõistmise võime. Seetõttu võib masinõppealgoritmi kasutav AIS

[1] Näiteks isesõitvate autode treenimisel ilmnevad raskused ja algoritmi täpsuse langus juba algoritmi kasutamisel erinevates linnades [149, lk 7]. Samuti on praktikas näiteid, kui isesõitev auto võib ilusa tavaolukorras liigelda ohutult ja nõuetekohaselt, kuid ei suuda tuvastada peatunud vilkuritega politseiautot takistusena ja põhjustab liiklusõnnetuse [150].

hakata lahendama püstitatust erinevat probleemi[1] või rakendada soovitust erinevat, kasutaja poolt mittesoovitud lahendust[2].

4) **Ebapiisav tõendus**. Ebapiisav tõendus on valdavalt masina loodud algoritmi poolt loodava teadmisega seonduv (epistemoloogiline) probleem, mis seisneb algoritmi ebaõigetes otsustes või soovitustes. Masinõppe algoritmid ja andmekaeve jõuab otsusteni läbi andmestikus tuvastatud seoste leidmise. Sellised otsused rajanevad paljuski üldistava statistika ja masinõppe tehnoloogiatel, mis annavad tõenäosuslike, kuid sageli ebakindlaid teadmisi [42, lk 4]. AIS-i otsuse ja soovituste tegemisel ei tuvastata tingimata kausaalsust või põhjuslikku seost andmetest leitud seoste ja algoritmi soovituste vahel. Ehkki andmetes avalduvate seoste alusel otsuste vastuvõtmine on sageli tulemuslik ka ilma põhjusliku seose tõendamiseta [61, lk 376], on korrelatsioonide pinnalt otsuste langetamine siiski ülimalt probleemne. Kui ühiskonnast kogutud üldise andmestiku alusel tuvastatud korrelatsiooni alusel tehakse STS-is otsuseid konkreetse inimese kohta, siis autori hinnangul ei erine sellise korrelatsiooni alusel üksikindiviidi kohta tehtud otsus oluliselt inimese poolt üldistusel rajaneva eelarvamuse alusel tehtud otsusest. Inimese poolt tehtavad otsused, mis rajanevad laiemas grupis esinevate üldistatud omaduste indiviidile ülekandmisel, on ilmselgel suutelised rikkuma otsuse põhjendamise kohustust ning diskrimineerimise keeldu. Samamoodi võivad korrelatsioonil rajanevad otsused olla üksikindiviidi suhtes põhjendamatud ja diskrimineerivad. Isegi kui AIS-i otsuseid ei rakendata indiviidide suhtes, vaid isikugruppide suhtes, siis selline otsus võib ikkagi olla ebaõiglane samasse gruppi langeva isiku suhtes, kelle muud omadused ei õigusta sarnast kohtlemist ülejäänud grupiga [42, lk 5].

---

[1] Näiteks luges melanoomi tuvastamiseks kasutatud pildituvastustarkvara ohtlikeks sünnimärgid, mille fotodel olid joonlauad [151]; pildituvastusalgoritm eristas fotosid huntidest ja *husky* tõugu koertest vastavalt sellele, kas fotol oleva looma taustal oli või polnud lund [110, lk 8].

[2] Näiteks õppis kiireima liikumisviisiga „mehhanismide“ arendamisele suunatud masinõppealgoritm „ehitama“ järjest kõrgemaid „konstruktsioone“ ning neid ümber lükkama kiire kukkumiskiirenduse saavutamiseks ja jalgpalli mängimise eesmärgil pallipuuteid saavutama treenitud algoritm hakkas palli löömise asemel puudete maksimeerimiseks palli vastas vibreerima. Kumbki eesmärkidest ei olnud algoritmi arendaja poolt soovitud [152].

5) **Seletamatu või arusaamatu tõendus**. Seletamatu või arusaamatu tõendus on algoritmi loodud teadmisega seonduv probleem, mis väljendub algoritmi otsuse otsustuste vastuvõtmise protsessi inimesele mittearusaadavuses ehk läbipaistmatuses. AIS-i otsuste läbipaistmatus on AIS eetiliste probleemidega seonduvas kirjanduses üheks enimkäsitletud probleemiks. Kui mistahes järelduse või otsuse tegemisel on aluseks informatsioon, siis on ühiskonna mõistlikuks ootuseks võimalus tutvuda informatsioonist järelduse või otsuseni viiva argumenteeritud põhjenduskäiguga. See on vajalik nii otsuse mõistmiseks kui ka otsuse kritiseerimiseks või vajadusel vaidlustamiseks. Kui see otsustustee ei ole iseenesest mõistetav, siis on seda probleemi võimalik ületada täiendava selgitamisega. Masinõppealgoritmide tööpõhimõtete tõttu ei ole otsuse põhjendamise ootus automaatselt täidetud ning selle täitmine on ka raskendatud [42, lk 4]. Esmalt raskendab põhjendamist otsuse tegemiseks kasutatava andmestiku kvantiteedi (algandmete töötlemine ja hindamine käib inimesele üle jõu eelkõige mahu tõttu) ja kvaliteedi (esinduslikkus, täielikkus, usaldusväärsus, kallutatus jne) probleemistik. Veelgi keerulisemaks muudab probleemi aga inimesele väljakutse tõlgendada ja interpreteerida seda, kuidas ja mil määral konkreetsed andmestiku hulka kuuluvad andmed mõjutavad algoritmi alusel tehtud otsust. Algoritmi otsuste põhjenduste mõistetavus ja läbipaistvus on vajalik selleks, et tagada algoritmide sisuline tulemuslikkus ja kindlus. Mittemõistevalt toimivate algoritmide puhul on piiratud ka nende nendega seonduvate riskide hindamine ja haldamine. Kirjanduse kohaselt on läbipaistvusel kaks komponenti: ligipääsetavus („*accessibility*“) ja arusaadavus („*comprehensibility*“). Ligipääsetavuse osas on algoritmidega seonduvateks probleemideks puudulik informatsioon algoritmide funktsioneerimise põhimõtete kohta – selle põhjuseks võib olla lihtsalt vajaduse mittenägemine või ka teised ristuvad huvid (ärisaladus, riigisaladus, vajadus vältida algoritmi petmist jne) [62, lk 3–5]. Arusaadavusega seonduvad probleemi on seotud masinõppe tuumolemusega – toimimise reeglid ei pruugi täies ulatuses olla inimese poolt loodud, otsust mõjutavate muutujate hulk ja nende osakaalude hulk võib olla inimesele hoomamatu ja õppimise käigus otsustusmehhanism muutub [62]. Sarnased probleemid võivad esineda ka inimese determineeritud reeglite järgi otsuseid langetavate algoritmidega, kuid seal on probleemi algeks inimesele iseenesest mõistevate reeglite rohkus – rohked reeglid ei ole inimesele

hoomatavad ja seetõttu arusaadavad, või on reeglid koostatud suurte meeskondade poolt pikema aja jooksul, mistõttu kaob terviklik ülevaade reeglitest ja seega nende mõjudest. Algoritmi mittemõistetavusest omakorda algavad mitmed probleemid – alustades töökindluse ja kvaliteedijuhtimise probleemidest kuni kasutajate ja ühiskonna usalduse puudumiseni AIS-i otsuste või soovitustes suhtes [42, lk 6–7]. Algoritmi ja selle otsuse selgitatavus tuleb tagada tasemel, et otsus on arusaadav isikule, keda otsus mõjutab [42, lk 14, 17].

6) **Ebaõige tõendus**. Ebaõige tõendus on algoritmi loodud teadmisega seonduv (epistemoloogiline) probleem, mis sageli avaldub algoritmi kallutatud otsustena [42, lk 7]. Valdavalt inimese loodud algoritmid ja nende otsused sisaldavad vältimatult nende loomises osalenud inimeste väärtusotsuseid ja väärtushinnanguid; samamoodi võivad kallutatuse põhjused tuleneda valdavalt masina loodud algoritmi õpetamisel kasutatud andmete puudustest või tehtud tehnilistest valikutest – nii näiteks põhinevad mitmed näotuvastustehnoloogiad kas kontrasti- või värvituvastusalgoritmidel, mille toimimise tulemuslikkus sõltub nahavärvist [63, lk 4974]. Sõltumata tehnilisest põhjusest võib nahavärvipõhine tulemuslikkuse erinevus viia diskrimineerimise ja algoritmi kasutaja vastutuseni [64], [65]. Kirjanduses on välja toodud järgmised infosüsteemi kallutatuse liigid [2, lk 333–336]:

Tabel 1. AIS-i kallutatuse liigid ja põhjused [2, lk 333–336].

	KALLUTATUSE LIIGID	**KALLUTUSE ALAMLIIGID**
1.	**Eeleksisteeriv kallutatus** tuleneb tehnoloogia rakendamisele eelnevalt selle rakendamise kontekstiks olevates sotsiaalsetes institutsioonides, praktikates ja hoiakutes peituvatest sotsiaalsetest väärtustest. Kui infosüsteem võtab üle süsteemi käitamise sotsiaalses kontekstis eeleksisteeriva kallutatuse, siis infosüsteem reeglina võimendab vastava kallutatuse mõju. Selline kallutatus võib süsteemi omaduseks saada nii isikute või organisatsioonide teadliku kui ka tahtmatu tegevuse tõttu; sageli ka vaatamata parimatele pingutustele kallutatust vältida	1.1. Kallutatuse põhjustajast sõltuv: 1.1.1. *Individuaalne kallutatus*, mille on põhjustanud süsteemi käitamist oluliselt mõjutanud aktor (isik või organisatsioon – nt süsteemi arhitekt või süsteemi tellija kavandab süsteemi, mis eelistab teatud rahvusest laenusaajaid) 1.1.2. *Ühiskondlik kallutatus*, mille on põhjustanud ühiskond laiemalt (vastava valdkonna praktikad, institutsioonid, või kultuur laiemalt, nt tarkvaratööstus arendab hariduslikku tarkvara, mis meeldib rohkem poistele kui tüdrukutele) 1.2. Kallutatuse tekkemehhanismi alusel: 1.2.1. *Andmete sotsiaalne kallutatus*, mille allikaks on sotsiaalselt mitteneutraalsed treeningandmed või otsuste sisendiks olevates andmetes esinev sotsiaalne kallutatus

	KALLUTATUSE LIIGID	KALLUTUSE ALAMLIIGID
		1.2.2. *Andmete algoritmiline kallutatus*. Masinõpealgoritmide õpetamiseks kasutatakse treeningandmestikke, mis peaksid olema loodud inimese poolt või kogutud algoritmist endast sõltumatult. Samas on praktikas ilmnenud, et suur osa kasutatavatest treeningandmestikest on loodud teiste masinõppealgoritmide poolt. Sellisel juhul on oht, et kui masinõppealgoritm on treeningandmete loomisel teinud vigu, siis kanduvad need järgmisesse algoritmi, mille tegevuse tulemusel loodavad andmed muutuvad omakorda sisendiks järgmise algoritmi õpetamisele ning seeläbi süveneb esialgne viga veelgi [40, lk 35]
2.	**Tehnoloogiline kallutatus** tõusetub läbi tehnoloogia rakendamise kas tehnoloogilistest piirangutest või valikutest. Eriti masinõppe puhul võivad sellise kallutatuse põhjuseks olla ka algoritmi treeningandmestiku puudused [42, lk 7]	2.1. *Arvutitööriistade põhjustatud kallutatus* tuleneb riist-, tarkvara ja välisseadmete piirangutest (nt kuna kasutaja eelistab otsingu esimesi tulemus, siis esimesele ekraanivaatele mittemahtuvad tulemused on koheselt halvemas positsioonis) 2.2. *Kontekstitundetu algoritmi kallutatus*, mis tuleneb algoritmi suutmatusest kohelda kõiki isikuid ja gruppe õiglaselt kõikides algoritmi kasutamise kontekstides (nt lennukitele õhkutõusu akna eraldamine lennufirmade nimede tähestikulises järjekorras) 2.3. *Juhuarvude genereerimise kallutatus*, mis tuleneb juhuarvude genereerimise vigadest olukordades, kus on oluline otsuste või soovituste juhuslik jaotamine (nt defitsiitsete ravimite eraldamine nimekirjas olevatele patsientidele mitte juhuslikult, vaid eelistades nimekirja lõpus või keskel olevateid patsiente) 2.4. *Inimesele omaste kontseptsioonide masinloetavuse puudustest tulenev kallutatus*, mis tuleneb infosüsteemi puudulikust suutlikkusest mõista ja adekvaatselt rakendada inimesele mõistetavaid kontseptsioone ja väärtushinnanguid (nt juriste abistavate ekspertsüsteemide liigne jäikus, mis ignoreerib inimese otsuse paindlikkust)
3.	**Süsteemi tekitatav kallutatus**, mis tekib alles süsteemi konkreetses kontekstis käitamisel. Sageli tekkib selline kallutatus süsteemi pikaajalisema kasutamise käigus süsteemi enda loodud sotsiaalsete või käitumuslike muutuste või arengute mõjul. Vastava kategooria mõjud ilmnevad ennekõike süsteemi ja kasutajate interaktsioonides, nende omapärades või nende puudumises	3.1. *Uue teadmise puudumine* on kallutatus, mille põhjustab infosüsteemi suutmatus arvesse võtta valdkonnas loodud uusi teadmisi (nt ekspertsüsteem, mis soovitab ravivõtteid loomise aegse parima teadmise kohaselt, kuid mis ei õpi operatiivselt uuest teadmisest) 3.2. *Süsteemi kasutajate ja süsteemi kokkusobimatus* põhjustab kallutatust, kui süsteemi kasutajate oskused või muud omadused erinevad oluliselt süsteemi loomisel eeldatud kasutajate oskustest või omadustest:

	KALLUTATUSE LIIGID	KALLUTUSE ALAMLIIGID
	(nt avaliku teabe edastamisel interneti võimalustele keskendumine on põhjustanud selle ligipääsuta inimeste madalama informeerituse).	3.2.1. Oskuste erinevus on probleemiks, kui süsteem on arendatud tegelikest kasutajatest erinevate oskustega kasutajate jaoks (nt kasutajad ei valda süsteemi kasutajaliidese keelt) 3.2.2. Väärtuste erinevus on probleemiks, kui süsteem on arendatud tegelikest kasutajatest erineva väärtus-süsteemiga kasutajate jaoks (nt individuaalsel võistlusel rajanev õppeprogramm ühiskonnas, mis individualismi asemel on kantud kollektiivsetes väärtustest) 3.2.3. Kasutajate kuritarvitused ja manipulatsioonid on probleemiks, kui kasutajad tahtlikult manipuleerivad (nt Microsoft Tay tekstirobotile rassistlike väljendite õpetamine [66]) või kuritarvitavad algoritmi

Eelnevast liigitusest tulenevalt tuleb infosüsteemi kallutatust mõista kui süsteemi poolt põhjustatud mõju. Vastavate mõjude vältimine nõuab infosüsteemi analüüsi faasis esmalt kallutatuse tuvastamise ja analüüsimise suutlikkust ning seejärel kallutatuse kõrvaldamiseks vajalike meetodite ja vahendite rakendamist.

7) **Grupitunnustest põhjustatud diskrimineerimine**. AIS teeb konkreetset indiviidi mõjutavaid otsuseid isikute gruppide kohta tuvastatud korrelatsioonide alusel. Isikute grupi omaduste ülekandmine isikule ja isiku enda omaduste ja privaatautonoomiaga mittearvestamine on autori hinnangul iseseisvaks eetiliste probleemide allikaks. Mõned autorid peavad sellist grupitunnuste alusel isiku kohta otsuste langetamist otsesõnu diskrimineerimiseks [[42, lk 9] kaudu [67]]. Diskrimineerimise-probleemistikuga seonduv kirjandus keskendub tundlike tunnuste alusel otsuste tegemise probleemistikule. Tundlikud tunnused on sellised isikut iseloomustavad andmed, mille alusel isikute eristamine ja grupeerimine on reeglina keelatud[1] (nt on reeglina keelatud palkamisel valikukriteeriumine lähtuda inimese rahvusest või soost) ja lubatud üksnes erandlikes oludes (nt on sugude eristamine põhjendatud günekoloogiliste või

[1] Näiteks on isikud Eestis seaduse alusel kaitstud diskrimineerimise eest rahvuse, rassi, nahavärvuse, usutunnistuse või veendumuste, vanuse, puude või seksuaalse sättumuse alusel [98]

androloogiliste haiguste diagnoosimisel). Kirjanduse kohaselt on küll võimalik AIS-is kasutatud andmetest elimineerida otseselt lubamatud või sensitiivsed tunnused (nt rahvus), kuid mitmekesistes suurandmetes võib sisalduda mitmeid andmeid, mis kaudselt viitavad ikkagi lubamatutele või sensitiivsetele tunnustele (nt viitab rahvusele selle rahvuse poolt domineeritud elanikkonnaga piirkonna postikood) ja seeläbi võivad kaudselt ikkagi viia diskrimineeriva mõjuga AIS otsuste või soovitusteni. Sellise kaudsete tunnuste alusel diskrimineerimise ja selle vältimise problemaatikaga tegeleb diskrimineerimis-teadliku andmekaeve uurimisvaldkond [9], [68]. Valdav osa kirjandusest peab grupitunnustest lähtuva AIS-i diskrimineerivaid mõjusid ebaeetiliseks, kuna see võib viia isiku suhtes isetäituvate ennustuseni, stigmatiseerida teatud ühiskonnagruppide liikmeid, õõnestada nende autonoomiat ja õigust eneseteostusele ning piirata aktiivset osalemist ühiskonnaelus [42, lk 9].

8) **Käitumist suunav ja piirav mõju**. AIS otsused ja soovitused võivad suunata ja piirata indiviidide ja ühiskonna käitumist ja vaba tahet. Näiteks võivad AIS-id inimeste käitumist suunata valikulise või teatud isikugruppidele erineva informatsiooni esitamisega (nt maksejõulisematele isikutele kuvatakse ainult kallimaid pakkumisi). Infosüsteemi poolt isiku varasemate valikute või grupitunnuste alusel personaliseeritud informatsiooni pakkumise korral on probleemiks õhkõrn piir inimese abistamise ja inimese kontrollimise vahel – paradoksaalsel kombel võib personaliseeritud informatsioon inimest nii aidata asjakohase informatsiooni leidmisel kui tema informeeritust ja käitumist piirata, peites tema eest teisi käitumisvõimalusi [69, lk 123] [70, lk 223]. Vaieldamatult on eetiliselt probleemne olukord, kus inimesele informatsiooni kuvamisel lähtutakse mitte inimese enda, vaid kolmanda isiku huvidest. Kuid inimväärikuse ja vaba tahte seisukohalt on probleemne ka olukord, kus info valikuline esitamine piirab vaba tahte kujunemist.

9) **Jälgitavus ja vastutus**. Ühiskonnas vajaliku usalduse eelduseks on võimalus mõista AIS-i otsuste põhjendusi ning vastutuse tagamine ebaõigete otsuste eest. Inimese poolt determineeritud reeglite alusel toimivate infosüsteemide puhul on võimalik süsteemi toimimist nii mõista kui ka leida negatiivsete mõjude eest vastutav isik – reeglina on selleks süsteemi arendaja või käitaja [42, lk 10–11]. AIS korral ei ole selline vastutusahel tingimata kohane – masinõppealgoritmi

võime õppimise käigus oma toimimise parameetreid muuta põhjustab olukorra, kus „*mitte kellelegi ei ole masina tegevuse üle piisavalt kontrolli, et vastutada masina tegevuse eest*“ [71, lk 181]. Erinevus süsteemi kavandaja kontrolli ja algoritmi tegeliku käitumise vahel tekitab vastutuse tühimiku, mis võib muuta vastutuse kandja leidmise keeruliseks. Autori hinnangul on seda tühimikku võimalik ületada normatiivsete ettekirjutustega, mis panevad vahetu vastutuse AIS-i negatiivsete mõjude eest AIS-i käitajale. Vastavat suunda pooldavad ka Euroopa Liit, IEEE kui ka OECD (vt alapeatükk 3.4). Oluline osa kirjandusest ongi pühendatud algoritmile eetilise ja moraalse vastutuse omistamise küsimusele [42, lk 10–12], kuid autori hinnangul on see diskussioon minetamas praktilisust, kuna vastutust reguleerivad normatiivaktid asetavad vastutuse AIS-i toimimise eest ühemõtteliselt selle käitajale. Küll aga on relevantne nõue, et algoritmid peavad olema nii selgitatavad („*explainable“*) kui ka inimesele arusaadavad („*understandable“*). Neid nõudmisi võib olla võimalik osaliselt täita ka otseselt, tagades indiviidile vahetu võimaluse saada selgitusi otsuste põhjenduste kohta; praktilisest vaatepunktist võivad efektiivsemad olla aga kaudsed meetodid, kus vastav eesmärk saavutatakse läbi vahendatud tegevuse, nt auditeerimise, sertifitseerimise või riikliku järelevalve [42, lk 13].

Üldistatult on AIS-iga kaasnevate eetilisi ja sotsiaalseid mõjusid omavate probleemide allikad järgmised:

1) AIS-i võimetele ja omadustele seotakse ootusi, mida tehnoloogia põhimõtteliselt ei ole suuteline täitma (töökindlus, kallutatuse puudumine ja sellest tulenev automaatne otsuste eetilisus);

2) AIS-il puudub arusaamine käitamise keskkonna eetilistest ja sotsiaalsetest nõuetest ning võime eristada nendele nõuetele vastavaid ja mittevastavaid otsuseid, mistõttu võib AIS teha ebaeetilisi ja sotsiaalselt lubamatuid otsuseid, kui süsteemi käitaja ei rakenda neid riske haldavaid meetmeid;

3) AIS-iga töödeldava informatsiooni hulga ning algoritmi keerukuse tõttu on inimese arusaamine algoritmi otsuste põhjendustest piiratud ning seetõttu on AIS-i inimese poolt kontrollimine ja auditeerimine[1] keeruline.

AIS-i ja valdavalt masina loodud algoritmi otsuste põhjenduste arusaadavuse ja kontrollitavuse piiratus inimeste jaoks tekitabki peamised probleemid, mille lahendamise võimalusi käesolev töö käsitleb. Need probleemid ja lahendused piirnevad ühelt poolt isikuvabaduste ja põhiõigustega, kuid teisalt on taandatavad väga praktilisele küsimusele (kas infosüsteemis saab kasutada komponenti, mille toimimise põhimõtted ja töökindlus on ebaselge).

[1] Masinõppe kontekstis on auditeeritavusele omistatud järgnev tähendus: *"Auditeeritavus tähendab, et tehisintellekti süsteemi algoritme, andmeid ja projekteerimisprotsesse saab hinnata. […] See ei tähenda tingimata, et tehisintellekti süsteemiga seotud ärimudeleid ja intellektuaalomandit käsitlev teave peab olema alati avalikult kättesaadav. Jälgitavus- ja logimismehhanismide tagamine alates tehisintellekti süsteemi projekteerimise varajasest etapist võib aidata tagada süsteemi auditeeritavust."*[36, lk 44]

3 Eetilised ja sotsiaalsed nõuded AIS-ile

3.1 Ülevaade erinevatest algatuste liikidest

Tehisintellekti ja masinõppe strateegilised algatused on peaaegu kõikidel riikidel, mis soovivad mängida olulist rolli globaalsel areenil[1]. Tehisintellekti eetiliste ja sotsiaalsete riskide haldamisega tegelevad ka valdkonna ettevõtjad [15], [72] ja teadlased [73].

Fookus AIS-i eetikaküsimustele tuleneb mõistmisest, et uued tehnoloogiad kätkevad endas uusi eetilise ja sotsiaalse mõõtmega riske, ja teisalt tõdemusest, et klassikaline inseneriteadus tegeleb nende riskidega ebapiisavalt. Inseneriteadused on alati lähtunud insenerieetikast, millega omakorda seonduvad sellised insenerlikud baasväärtused nagu ohutus, turvalisus, funktsionaalsus. Samas sotsiaalsema mõõtmega küsimusi nagu kallutatus, õiglus, sõltuvus ning kaudsed ühiskondlikud kahjud ei ole traditsiooniliselt inseneriteaduste uurimisesemesse loetud.

Tehnoloogia ühiskondlike mõjude suurenemise tõttu on ootused inseneridele muutnud – enam ei ole aktsepteeritav sotsiaalseid mõjusid eirava tehnoloogia kasutajateni viimine ning negatiivsete mõjude ühiskonna lahendada jätmine. Ehkki inseneridelt ei oodata filosoofide, psühholoogide ja sotsioloogide töö tegemist, oodatakse inseneridelt sotsiaalsete ja eetiliste riskide haldamist käsikäes teiste teadusvaldkondade esindajatega [74, lk 1]. Seetõttu palkavad tehnoloogiaettevõtted rohkem sotsiaalteadlasi, ülikoolid laiendavad oma õppekavasid, õpetamaks insenere lahendama ka eetilisi küsimusi ja standardiorganisatsioonid töötavad välja AIS-i eetikastandardeid [75, lk 110].

Alljärgnevalt annab autor ülevaate AIS-ile, selle arendajatele ja käitajatele esitatavaid eetilisi ja sotsiaalseid nõudeid käsitlevatest algatustest. Autor keskendub Euroopa Liidu, OECD, EN ja IEEE algatustele, kuna need mõjutavad Eestit enim.

---

[1] Valdkonnas on tihe konkurents Ameerika Ühendriikide ja Hiina Rahvavabariigi vahel. Mõlemad riigid on välja töötanud mastaapsed plaanid tehisintellekti ja masinõppe alaseks teadustööks ning tehnoloogia laiapõhjaliseks rakendamiseks. Mõlema riigi algatused paistavad silma utilitaarsusega, keskendudes tehnoloogia arendamise ja rakendamise küsimustele. Palju algatused on pärit ka Euroopast, kus oma algatused on EL-il, liikmesriikidel ja organisatsioonidel. Euroopa algatused paistavad silma inimkeskse lähenemise ja eetika küsimuste esiplaanile seadmisega. Selle hinnaks on aga praktilisuse ja rakendatavuse piiratus. Eetilise AIS-i algatuste osas on selgelt alaesindatud Aasia riigid [10, lk 5].

3.2 EL-i eetilise ja usaldusväärse AIS-i eetikasuunised

3.2.1 Eetikasuuniste alused ja mõju

EL ja liikmesriigid on AIS-i[1] küsimustega ühiselt tegelenud alates Eesti eesistumise ajal 2017. aasta septembris korraldatud digitaalvaldkonna tippkohtumisest [76, lk 1].

2018. aastal sõlmisid Euroopa Liidu 24 liikmesriiki[2] ja Norra ühise „*Koostöödeklaratsiooni tehisintellekti valdkonnas*“, mis rõhutab vajadust tagada AIS-i arendamise ja käitamise usaldatavus („*trust*“), läbipaistvus („*transparency*“) ning vastutus („*accountability*“) kooskõlas Euroopa Liidu põhiväärtuste ja tunnustatud põhiõigustega[3] [77] [78, lk 1–3].

2018. aastal moodustas EK kõrgetasemelise tehisintellekti ekspertrühma (AI HLEG), mis peaks 2020. aastal jõudma Euroopa AIS eetikasuuniste lõpliku versioonini [79, lk 7], [80, lk 17]. Eetikasuuniste eesmärgiks on inimkeskse ja usaldusväärse AIS-i arendamine: AIS ei ole eesmärk iseenesest, vaid tegemist on tööriistaga, mille ülim eesmärk on inimkonna heaolu tõstmine. Seetõttu tuleb tagada AIS-ide usaldusväärsus inimese ja ühiskonna vaatepunktist. Usaldusväärsust tagab nõue, et ühiskonna toimimise aluseks olevad väärtused tuleb täielikult ja läbivalt integreerida AIS-i ja selle arendamisse („*ethical by design*“) [81, lk 23]. Integreerimist vajavateks väärtusteks on EL üldised väärtused – inimväärikus, vabadus, demokraatia, võrdsus, õigusriik ja inimõiguste austamine, sh vähemuste osas [79, lk 1–2].

Eetikasuuniste kohaselt ei ole vajalik ega võimalik nõuda AIS-ilt arusaamist eetilistest ja sotsiaalsetest väärtustest ja nendest juhindumist. Vastavad nõuded rakenduvad AIS

---

[1] Euroopa Liit kasutab mõistet “tehisintellekt”. Autor kasutab selle asemel töös läbivalt IEEE mõistet “AIS”, mis on autori hinnangul tehnoloogilistelt täpsem ja antropomorfistlike konnotatsioonide puudumise tõttu sotsiaalselt neutraalsem mõiste.

[2] Deklaratsiooni allkirjastasid Austria, Belgia, Bulgaaria, Tšehhi Vabariik, Taani, Eesti, Soome, Prantsusmaa, Saksamaa, Ungari, Iirimaa, Itaalia, Läti, Leedu, Luksemburg, Malta, Madalmaad, Poola, Slovakkia, Sloveenia, Hispaania, Rootsi, Ühendkuningriik, Norra; hiljem liitusid deklaratsiooniga Rumeenia, Kreeka, Küpros ning Horvaatia [77].

[3] Teatises „*Tehisintellekt Euroopa huvides*“ on selgitatud, et Euroopa Liidu põhiväärtused ja tunnustatud põhiõigused on kajastatud Euroopa Liidu lepingu artiklis 2, mis on kõigi Euroopa Liidus elavate inimeste õiguste alus, ning Euroopa Liidu põhiõiguste hartas, mis koondab ühte teksti kõigi Euroopa Liidus elavate inimeste isiku-, kodaniku-, poliitilised, majandus- ja sotsiaalsed õigused. [80, lk 15]

arendajatele ja käitajatele, kes peavad tagama AIS-ide käitamise, tulemuste ja nende kontrollitavuse vastavalt EL-i üldistele väärtustele. Seega lähtub EL-i käsitlus käesoleva magistritöö peatükis 1.2 käsitletud eetiliste hinnangute ja nõuete omistamise reeglitest.

2018. aasta teatises „Tehisintellekt Euroopa huvides“ [82, Lõik 20] soovib EL pakkuda oma eetilist ja õiguslikku raamistikku – ohutus- ja tootjavastutuse, võrgu- ja infosüsteemide turbe ning isikuandmete kaitse standardeid – tehisintellekti tehnoloogiate reguleerimise globaalseks standardiks [80, lk 3, 15]. Arvestades EL-i suutlikkust mõjutada majandustegevust väljaspool oma piire [45, lk 2], võib EL-i algatusel olla arvestatav mõju ka väljaspool EL-i tegutsevatele ettevõtetele ja organisatsioonidele [80, lk 19–20], [83], [84].

AIS valdkonnaga seonduvate EL-i poliitikadokumentide fookuses on soov kaitsta isikute põhiõigusi ja vabadusi AIS-i eetiliste ja sotsiaalsete riskide eest [36, lk 14–18], [79], [85, Lõik 13–15; 22], [86], kusjuures selle nimel on EL valmis aktsepteerima uudse tehnoloogia aeglasemat arengut. Seetõttu on oodata, et EL hakkab kehtestama rangeid nõudeid AIS-ide arendajatele ja käitajatele.

3.2.2 Usaldusväärse AIS-i tunnused

AI HLEG hinnangul peab usaldusväärne AIS vastama kolmele tunnusele [36, lk 2]:

1) **Seaduslikkus**. AIS-i käitaja peab järgima kõiki kohaldamisele kuuluvaid õigusnorme. AIS ei toimi seadusetus maailmas, vaid õigusnormid on osa AIS-i käitamise keskkonnast. Juba täna kehtivad AIS-ide suhtes erinevaid regulatiivsed nõudeid – lisaks AIS-i erinõuetele ka üldised nõuded (nt ohutus). Õiguslik raamistik kehtestab nii positiivseid kui ka negatiivseid kohustusi: AIS-i käitlemisel ei ole asjakohased vaid nende suhtes otseselt rakenduvad käsud ja keelud, vaid arvestada tuleb ka teiste isikute õigustega, mida AIS ei tohi rikkuda [36, lk 7].

2) **Eetilisus**. AIS-i käitamine peab olema eetiline, st austama ja järgima eetikapõhimõtteid ja -väärtusi. Eetikasuuniste kohaselt ei piisa sotsiaalmajanduslikult optimaalse tulemuse saavutamiseks üksnes õigusaktide nõuete järgimisest – seadusandlus ei jõua alati tehnoloogia arenguga sammu pidada või ei sobi eetiliste probleemide lahendamiseks. Seetõttu tuleb AIS-ide

käitamisel, sh mõjude hindamisel arvestada laiemate eetiliste nõuetega [36, lk 8]. Juhised defineerivad järgmised neli ranget eetilist nõuet [36, lk 13], mida usaldusväärne AIS peab järgima:

a. inimeste sõltumatuse austamine – AIS-ist mõjutatud inimesel peab olema võimalik säilitada oma tegelik vaba tahe; AIS ei tohi inimest allutada, petta ega temaga manipuleerida, vaid peab olema suunatud inimese kognitiivse, sotsiaalse ja kultuurilise suutlikkuse võimendamisele; süsteemide tööprotsesside peavad olema allutatud inimese järelevalvele ja kontrollile [36, lk 13] ning olema kontrolli ja järelevalve tagamiseks inimesele arusaadavad;

b. kahju tegemisest hoidumise nõue – AIS-ide käitamine ei tohi inimest kahjustada, kahju põhjustada ega seda suurendada; kahjustamise keeld hõlmab nii kitsamalt vaimse ja füüsilise puutumatuse kaitset kui ka üldisemat inimväärikuse kaitset, kusjuures suuremat tähelepanu tuleks pöörata haavatavatele isikutele[1], keda tuleb kaasata neid mõjutavate AIS-ide arendamisse. Arvestada tuleb ka looduskeskkonna ja kõigi elusolenditega [36, lk 14];

c. õigluse põhimõte – AIS-ide, kasutuselevõtmine ja käitamine peab olema õiglane nii materiaalses kui ka menetluslikus mõttes; materiaalse õigluse all peavad eetikasuunised silmas kasude ja riskide võrdset ja õiglast jaotumist, kohustust vältida kallutatust, diskrimineerimist ning häbimärgistamist üksikisikute ja gruppide suhtes; õigluse menetluslik mõõde nõuab, et mõjutatud inimesel peab olema võimalik vaidlustada ja saada tõhusat õiguskaitset AIS-i ja selle käitaja otsuste eest, kusjuures selleks tuleb tagada otsuse eest vastutava isiku tuvastatavus ning

[1] Haavatavate isikute määratlus varieerub sõltuvalt valdkonnast ja organisatsioonist [153, lk 7], kuid lihtsustatult loetakse haavatavateks isikuteks neid isikuid, kelle vanus, vaimne või füüsiline seisund või puue raskendab, vähendab või piirab nende võimet sotsiaalsetes protsessides tõhusalt osaleda [154, lk 1]. Eetikasuunised toovad selles kontekstis eraldi välja "*lapsi, puuetega inimesi ja muid rühmi, kes on ajalooliselt olnud ebasoodsas olukorras või tõrjutuse ohus, ja/või olukordadele, mida iseloomustab võimu ja teabe asümmeetria, näiteks suhted tööandjate ja töötajate või ettevõtjate ja tarbijate vahel*" [36, lk 15].

suutlikkus AIS-i otsuste põhjendusi inimesele arusaadavalt selgitada [36, lk 14];

d. selgitatavuse põhimõte – AIS-i otsuste põhjenduste selgitatavus ja mõistetavus on AIS-i laiaulatuslikuks kasutamiseks vajaliku ühiskondliku usalduse eelduseks. Seetõttu peavad AIS-iga seonduvad protsessid olema läbipaistvad, informatsioon süsteemide suutlikkuse ja otstarbe kohta peab olema avalikult kättesaadav ning süsteemi otsused peavad olema selgitatavad neile, keda otsused mõjutavad. Kui AIS-i otsuste loogika ei ole inimesele mõistetav siis tuleb selgituste andmiseks kasutada muid meetmeid (nt auditeerimine, sertifitseerimine) [36, lk 14–15].

3) **Töökindlus**. AIS peab olema tehniliselt ja sotsiaalselt töökindel. Siin tähendab tehniline töökindlus süsteemi toimimist vastavalt spetsifikatsioonile ja sotsiaalne töökindlus AIS-i käitamise konteksti, kasutajate ja keskkonna arvestamist, sh kasutajate informeerimist konkreetse AIS-i võimekustest ja puudustest. AIS-i käitamisel ei piisa headest kavatustustest, vaid tuleb vältida ka tahtmatult tekkivat kahju indiviididele ja ühiskonnale [36, lk 2]. Selleks tuleb ette näha kaitsemeetmed AIS-i tahtmatute kahjulike mõjude tuvastamiseks ja nende mõjude vältimiseks [36, lk 8], [79, lk 3]. Suuniste kohaselt on see nõue ennekõike usaldusväärse AIS teiste tunnuste – seaduslikkuse ja eetilisuse – tagamise vahend.

Usaldusväärsuse mõiste ei puuduta üksnes AIS-i tehnilist töökindlust, vaid ennekõike kogu süsteemi käitamise elukaare usaldusväärsust – tagada tuleb kõigi süsteemi elukaares osalevate inimeste ja protsesside usaldusväärsus [36, lk 46]. AIS arendajal ja käitajal on vastutus ka AIS-i mõjude eest. Seetõttu peavad eetilistele ja sotsiaalsetele nõuetele vastama ka AIS-i otsused ja soovitused ning see vastavus peab olema kontrollitav nii otsustusprotsessi kui selle väljundite tasemel.

3.2.3 Põhinõuded usaldusväärse AIS-i raamistikule

Usaldusväärse AIS tunnused on praktiliseks rakendamiseks liiga üldised. Seetõttu esitab AI HLEG eetikasuunistes ka põhinõuded („*key requirements*“) usaldusväärse AIS-i raamistikule [36, lk 2].

AI HLEG kirjeldatud põhinõuded on ülevaatlikult esitatud järgneval joonisel, kusjuures heptagrammi tippudes asuvad põhinõuded on kõik on ühtemoodi olulised ning toetavad teineteist läbi kogu AIS-i elukaare [36, lk 17]:

Joonis 5. EL-i usaldusväärse AIS-i põhinõuded [36, lk 17].

Joonisel 5 esitatud usaldusväärse AIS-i põhinõuete täpsem sisu on avatud järgnevalt:

1) **Inimese toimevõime („*human agency*“) ja järelevalve**

AIS-i käitamine peab toetama inimesi paremini informeeritud otsuste tegemisel. AIS-i käitamine peab võimendama, mitte aga vähendama, piirama või eksitama inimese toimevõimet [79, lk 4]. Inimeste toimevõime tagamiseks peab süsteem juhinduma põhiõigustest ning kasutajatele tuleb anda teadmised ja võimalus süsteemi toimimispõhimõtete hindamiseks. Vastava nõude oluliseks väljenduseks on ka inimjärelevalve nõue, mis võib avalduda läbi inimosaluse, inimsekkumise või

inimjuhitavuse[1] [79, lk 4]. See nõue on juba realiseerunud õigusliku keeluna teha inimeste suhtes õiguslike tagajärgedega või muud olulist mõju avaldavaid otsuseid üksnes inimese osaluseta automatiseeritud otsustena [85, Lõik 22]. Mida raskem on inimesel AIS-i üle järelevalvet teostada, seda rangemad peavad olema nõuded süsteemi kasutamisele ja riskide haldamisele [36, lk 18].

2) **Tehniline töökindlus ja turvalisus**

Usaldusväärne AIS peab olema töökindel ja turvaline sellisel tasemel, mis väldib AIS-i vigu ning haldab adekvaatselt vigadest tulenevaid riske [79, lk 4] Sellisena on töökindluse nõue tihedalt seotud kahju tegemisest hoidumise nõudega. AIS peab toimima täpselt – seda nii puhtalt süsteemi võimena teha õigeid otsuseid kui ka hästi välja töötatud arendus-, hindamis- ja riskihaldusprotsessidena, mis aitavad vältida, leevendada ja parandada AIS-i vigade või kavandamata mõjudega kaasnevaid riske [36, lk 18–19]. Täpsuse tagamise minimaalseks meetmeks on tegelikkusele vastava teabe andmine AIS-i täpsuse kohta [79, lk 5]. Töökindlust iseloomustab AIS-i tulemuste korratavus – AIS peab toimima sarnaselt, nõuetekohaselt ja ennustatavalt erinevate sisenditega ja olukordades [36, lk 19]; korratavus on vajalik süsteemi toimimise kontrollimiseks. Töökindluse tagamiseks peavad AIS-id käsitlema küberturvalisust süsteemi disaini aluspõhimõttena („*security by design*“) [79, lk 5].

3) Privaatsus ja andmehaldus

AIS-i käitamine peab tagama nii privaatsuse kaitse kui ka andmekaitse kogu süsteemi elukaarel – seda alates süsteemi tegevuse aluseks olevate andmete kogumisest, andmesubjekti informeerimisest kuni vastava otsuse üle järelevalve teostamiseni. AIS-i privaatsuse ja andmehalduse tagamise nõue on tihedalt soetud kahju tegemisest hoidumise põhimõtte [36, lk 19] ning autori arvates ka inimeste sõltumatuse austamise põhimõttega. AI HLEG hinnangul on privaatsuse tagamine ühiskondliku usalduse

[1] Vastavate juhtimismehhanismide tunnused on järgmised: inimosaluse (*HITL – human-in-the-loop*) korral saab inimene süsteemi tegevusse sekkuda igas otsustustsüklis, mis ei ole sageli praktiline ega eesmärgipärane; inimsekkumise (*HOTL – human-on-the-loop*) korral saab inimene sekkuda süsteemi kavandamisse ja jälgida süsteemi toimimist, kuid inimene ei pruugi vahetult osaleda süsteemi toimimises; inimjuhitavuse (*HIC – human-in-command*) korral saab inimene juhtida AIS kogu tegevust, kaasaarvatud selle laiemat majanduslikku, ühiskondlikku, õiguslikku ja eetilist mõju, ning inimese võimalust otsustada kas, millal ja kuidas AIS-i konkreetses olukorras kasutada) [79, lk 4].

põhieeldus. Usalduse tagamiseks on vajalik tagada andmesubjektide täielik kontroll oma isikuandmete üle [79, lk 5]. Andmehalduse osas peab AIS käitaja tagama süsteemi poolt kasutatavate andmete kvaliteedi [36, lk 20] – autor nõustub, et ebakvaliteetsete ja puudulike andmete alusel ei ole võimalik langetada usaldusväärselt AIS-il nõutavate omadustega otsuseid. AI HLEG rõhutab eetikajuhendis, et suurandmetena kogutud andmed võivad olla kallutatud, puudulikud või ebapiisavalt hallatud (nt puudub andmete kogumise ja selle konteksti kohta ammendav ja kontrollitav dokumentatsioon). AIS-iga töödeldava andmestiku osas tuleb tagada asjakohased testimise, kontrollimise ja ligipääsuõiguste süsteemid ning nende dokumenteerimine auditeerimist võimaldaval tasemel [79, lk 5]. AI HLEG peab vajalikuks andmete probleemide lahendamist enne andmete alusel AIS-i õpetamist ja otsuste langetamist[1] [79, lk 5].

4) **Läbipaistvus**

Läbipaistvuse nõue toetab AIS-i selgitatavuse nõuet ning hõlmab AIS-i kasutamise tervikkonteksti – andmete, süsteemi, ärimudelite – läbipaistvust. Läbipaistvuse nurgakiviks on jälgitavus: AIS-i otsuste aluseks olevad protsessid (sh algoritmide otsustusloogika) ja andmestik tuleb dokumenteerida vastavalt parimatele võimalikele standarditele, et tagada kogu AIS-i elukaare selgitatavus ja auditeeritavus. Selgitamise nõudele allub nii AIS-i tehniline toimimine kui seonduvad otsuseid, mis on langetatud inimese poolt. Selgitused peavad olema õigeaegsed ning vastama konkreetse AIS-i poolt mõjutatud osapoolte tegelikule kvalifikatsioonile [36, lk 20] – erineva kogemuse ja arusaamisvõimega osapooltele tuleb anda erineva ulatusega selgitusi. Organisatsiooni (sh ärimudeli) läbipaistvuse tagamiseks on tuleb AIS-i otsusest mõjutatud isikule selgitada seda, mil määral mõjutab AIS isiku suhtes tehtavaid otsuseid [79, lk 5]. Vastava nõude täitmiseks on vajalik ka aus kommunikatsioon AIS-i tegeliku suutlikkuse, piirangute ning kasutuselevõtmise põhjuste osas – nt võib AIS-i

[1] Autori hinnangul võib absoluutne nõue lahendada andmete kallutatuse probleem enne AIS-i õpetamist olla liialt ambitsioonikas ning negatiivse kaasmõjuna kaasa tuua innovatsiooni ning AIS-i täpsuse vähenemise. Seetõttu võivad tõhusamad olla meetmed, mis võimaldavad AIS-i kallutatust märgata ja korrigeerida süsteemi hilisema käitamise käigus (vt alapeatükki 5.3).

juurutamise põhjuse varjamine või eksitava põhjuse esitamine[1] olla vastuolus selle põhinõudega. AI HLEG peab läbipaistvuse nõude lahutamatuks osaks ka inimese õigust olla teadlik, et nad suhtlevad AIS-iga („*tehisintellekti inimesena esitlemise keeld*“) ning inimeste õigust pöörduda AIS-i asemel inimese poole, kui see on vajalik põhiõiguste järgimiseks [36, lk 20–21].

5) **Mitmekesisus, mittediskrimineerimine ja õiglus**

AIS-ide käitamine peab tagama optimaalse lahenduse kogu ühiskonna jaoks. Seetõttu tuleb AIS-ide käitamisel arvestada ühiskonna kõigi sidusrühmadega, tagada neile võrdne ligipääs tehnoloogia arengule ning võrdne kohtlemine AIS-ide poolt. AIS-ide õigluse tagamiseks ja kallutatuse vältimiseks on vajalik kallutatuse põhjuste mõistmine ning nende vältimise meetmete rakendamine juba süsteemide kavandamise faasis. AIS-ide suur jõudlus võib andmete kallutatust ja puudusi võimendada ning viia tahtmatute ebaõiglaste tulemusteni. Õigluse põhimõte nõuab, et selliste probleemide vältimisega tegeletakse juba süsteemi kavandamise algfaasis. AIS-ist tulenevate eeliste õiglase jaotuse tagamiseks peavad süsteemid olema juurdepääsetavad kõigile ühiskonnagruppidele, sõltumata nende vanusest, soost, võimetest või muudest omadustest. AI HLEG rõhutab eraldi universaaldisaini põhimõtet puuetega inimestele ligipääsu tagamise kontekstis[2] [36, lk 21].

6) **Ühiskondlik ja keskkonnaalane heaolu**

AI HLAG leiab, et AIS-ide käitamisel tuleb sidusrühmadena käsitleda lisaks inimestele ka teisi aistmisvõimelisi olendeid ja laiemalt keskkonda [79, lk 6]. Seetõttu tuleb soodustada AIS-ide kestlikkust ning ökoloogilist vastutustundlikkust. Igavikulisemate huvide (nt järeltulevate põlvede huvide arvestamine) kõrval on AIS-i usaldusväärsuse hindamiseks vajalik ka hinnata süsteemi mõju ühiskonnale, selle institutsioonidele ja demokraatiale [36, lk 22]. Seega süsteemid, mille ärimudel on suunatud või mille

[1] Eksitava põhjuse esitamise näiteks on AIS-i juurutamisega kaasnev turunduslik teavitus, et süsteem võetakse kasutusele organisatsiooni otsuste kvaliteedi tõstmiseks, kuid tegelikult on süsteemi juurutamise põhjuseks soov koondada varasemalt äriprotsessis osalenud inimesed ning seeläbi kulusid vähendada.

[2] Autori hinnangul on AI HLEG universaaldisaini võimalustele osundanud piiratult, kuna läbiva kasutamise korral aitab universaaldisain kaasa ka inimeste toimevõime, tehnilise töökindluse ning privaatsuse tagamisele.

kõrvalmõjudeks on ühiskonna lõhestamine või ühiskonnas paratamatult valitsevate vastuolude võimendamine, ei ole autori hinnangul käsitletavad usaldusväärse AIS-ina ning nendest põhjustatud riskide haldamiseks on põhjendatud süsteemiväliste, sh regulatiivsete meetmete rakendamine.

7) **Vastutus**

Eetikasuunised nõuavad AIS-i käitaja vastutust („*accountability*“) süsteemi käitamise ja mõjude eest [79, lk 6]. Vastutuse võtmise nõue tähendab esmalt AIS-i auditeeritavuse tagamist süsteemi algoritmide, andmestiku ja protsesside osas. Auditeerimisnõude täitmiseks ei ole tingimata vajalik algoritmide, andmestiku ja protsesside avalik kättesaadavus; tehnoloogia usaldusväärsuse tagamiseks on võimalik tugineda sisemiste ja väliste audiitorite hinnangutele ning hindamisaruannetele. Auditeerimisnõuded peavad olema rangemad süsteemide osas, mis mõjutavad isikute põhiõigusi ja turvalisust; sellisel juhul tuleb tagada sõltumatu auditeerimise võimalus. Vastutuse võtmise nõudest tulenevalt peab AIS-i käitaja tagama süsteemi ebasoovitavate eetiliste ja sotsiaalsete mõjude minimeerimise ning nendest teavitamise. Selleks rakendatavad meetmed[1] peavad vastama AIS-i mõjule ja sellest tulenevale riskile [36, lk 22–23]. Rakendatud meetmete sobivuse eest vastutab konkreetse AIS-i käitaja, kusjuures meetmete valik ja asjakohasuse hindamine peab olema osa AIS-i elukaart hõlmavast riskihaldusest. Riskihalduse raames tuleb välja selgitada AIS-ist mõjutatud eetilised ja sotsiaalsed väärtused ja huvid, hinnata nendevahelisi konflikte ning rakendada asjakohased meetmeid [79, lk 6]. Lõpetuseks tuleb AIS-i tekitatud ebaõigluse korral tagada mõjutatud osapoolele asjakohane õiguste kaitse [36, lk 23].

3.2.4 Eetikasuuniste mõju ja kriitika[2]

Eetikasuunised määratlevad ühiselt, et EL-is käideldavate ja siin mõju omavate AIS-ide arendajad ja käitajad peavad järgima EL-i põhiväärtusest ning tunnustatud põhiõigustest

[1] AI HLEG toob meetmete näitena välja mõjuhinnanguid, nn punaste meeskondade kasutamise (*red teaming*) ja mitmesugused algoritmilised mõjuhinnanguid.

[2] Kuna EL-il on pädevus anda oma territooriumil kohustuslikke õigusakte ning kehtestada nendega siduvaid nõudeid, siis käsitleb autor EL eetikasuuniste mõju ja kriitikat teistest algatustest põhjalikumalt.

tulenevaid eetilisi ja sotsiaalseid nõudeid. Need nõuded on koondatud erinevatesse EL-i raamdokumentidesse [77], [78, lk 1–3], [79, lk 1–2].

Eetikasuuniste keskendub ennekõike eetikasuuniste olemusele. Kriitikud kahtlevad, kas sedavõrd oluliste küsimuste reguleerimiseks piisab mittesiduvatest soovitustest [87, lk 9]. Kriitika on osaliselt õigustatud, kuid eirab vajadust enne ulatuslike õigusmuudatuste rakendamist analüüsida usaldusväärse AIS-iga seonduvaid riske ning nende haldamise vahendeid. Teiseks ei ole eetikasuuniste nõuded üksnes soovitusliku iseloomuga – mitmete eetikasuunistes kajastatud nõuete täitmine on vähemalt osaliselt kohustuslik juba täna kehtiva seadusandluse alusel (nt ohutuse tagamise nõue, isikute kahjustamise keeld, privaatsuse kaitse). Täna kehtiv seadusandlus paneb mitmete AIS-i eetiliste ja sotsiaalsete riskide realiseerumise tagajärgede eest vastutuse AIS arendajatele ja käitajatele. Seonduva debati alusel on oodata, et EL-i paneb AIS-i arendajatele ja käitajatele selge kohustuse tagada kasutatava lahenduse nõuetekohane toimimine, õiglus, kallutatuse puudumine ning AIS-i otsuste põhjenduste inimesele arusaadavus.

Teiste kriitikute arvates on AI HLEG puudulikult käsitletud spetsiifilisi tehnoloogilisi riske ning nende haldamise vahendeid ja meetodeid [11, lk 4]. Autor nõustub, et spetsiifiliste meetodite ja vahendite käsitlemine lihtsustaks usaldusväärse AIS-i arendamist ja käitamist. Samas on autori hinnangul kiiresti arenevas ja olemuslikult riskihaldusega seonduvas valdkonnas otstarbekas esmalt sõnastada selged üldnõuded ning jätta nende täitmise vahendite valik valdkonna osapooltele – arendajatele, käitajatele, standardimisasutustele. Alles üldnõuete tasemel reguleerimise läbikukkumisel või üldtunnustatud tehnoloogiliste lahenduste väljakujunemisel on põhjendatud kohutuslike meetmete ja vahendite kehtestamine.

Autori hinnangul on AI HLEG liialt optimistlikult hinnanud AIS-i käitavate osapoolte riskihaldamise suutlikkust. Eetikasuunistes kajastatud riskihaldamise protsess eeldab küpsete võimekustega organisatsiooni, millel on vastavateks tugiprotsessideks piisavalt ressursse. Olukorras, kus suur osa AIS-iga seonduvast innovatsioonist toimub *startup* või väike- ja keskmise suurusega ettevõtetes[1], on vastavate juhiste adekvaatne

[1] Ettevõtted, mille töötajate arv on väiksem kui 250, mille käive on kuni 50 miljonit eurot ja bilansimaht kuni 43 miljonit eurot [155, lk 3]

rakendamine raskendatud nii oskuste ja ressursside piiratuse kui ka teiste prioriteetide tõttu. Ka ei ole AI HLEG piisavat tähelepanu pööranud võimalikule konfliktile kirjeldatud riskihaldamise protsessi ja agiilselt tarbijateni jõudvat innovatsiooni soodustavate äri- ja arendusmeetodite [88], [89] vahel – juhul kui AI HLEG soovitused ja nõuded ei ole praktilises elus järgitavad (ega kohustuslikud), siis võib tehnoloogia ja ärimeetodite areng eetikasuuniste soovitustest lihtsalt mööda minna.

Kokkuvõttes peab autor eetikasuuniseid oluliseks dokumendiks, mis selgitab tänaseid eetilisi ja sotsiaalseid nõudeid AIS-i käitamisele ning viitab tulevikus konkretiseeruvatele ning karmistuvatele nõuetele. Eetikasuunistes sõnastatud eesmärkide ja nõuete tõttu ei saa enam AIS-i kavandamisel ja käitamisel eirata nõudeid süsteemi toimimise eetilisusele.

3.3 OECD nõuded usaldusväärsele AIS-ile[1]

22. mail 2019. aastal kiitis OECD heaks AIS-iga seonduvad soovitused. OECD lähenemise keskmes on AIS-i inimkeskne ja usaldusväärne käitamine [90, lk 2].

OECD käsitleb usaldusväärsena ja eetilisena AIS-i, mille käitamisel on elukaare kõigis etappides täidetud järgmised eetilised ja sotsiaalsed nõuded [90, lk 5–7]:

[1] OECD kasutab mõistet "tehisintellekti süsteem" ("*Artificial Intelligence System"* või *"AI system"*). Autor on asendanud selle mõistega AIS magistritöö terminoloogilise ühtluse tagamiseks.

Tabel 2. OECD nõuded usaldusväärsele AIS-ile [90, lk 5–7].

KAASAV MAJANDUSKASV, SÄÄSTEV ARENG JA HEAOLU
• AIS-ide käitamise eesmärk peab olema püüdlus inimeste ja planeedi heaolu suunas ja seeläbi üleüldisesse heaolusse panustamine [90, Lõik 1.1]**.**
INIMKESKSED VÄÄRTUSED JA ÕIGLUS
• AIS elukaares aktiivselt osalevad isikud ja organisatsioonid (AIS aktorid) peavad austama seadust, inimõigusi ja demokraatlikke väärtusi läbi kogu süsteemi elukaare. • AIS-i aktorid peavad rakendama inimese otsustusvabadust („*human determination*“) kaitsvaid meetmeid ja turvavõtteid, mis sobivad konkreetse AIS-i käitamise konteksti ja vastavad tehnika tasemele („*state of the art*“) [90, Lõik 1.2].
LÄBIPAISTVUS JA SELGITATAVUS
• AIS-i aktorid peavad edendama AIS-ide üldist mõistmist. • AIS-i kohta antav informatsioon peab olema asjakohane, konteksti sobituv ning vastama tehnika tasemele. • AIS-i osapooled peavad olema informeeritud (kaasaarvatud töösuhtes) sellest, et nad suhtlevad tehisintellekti komponendiga. • Võimaldada tuleb AIS-i otsuste arusaadavus nendest mõjutatud osapoolte jaoks. • AIS-i otsustest mõjutatud osapooltel peab olema otsuste vaidlustamise võimalus, mille eelduseks on selgete ja arusaadavate põhjenduste kättesaadavus [90, Lõik 1.3].
VASTUPIDAVUS, TURVALISUS JA OHUTUS
• AIS-id peavad olema vastupidavad, turvalised ja ohutud kogu nende elukaare vältel, nii et normaalse kasutamise, eeldatava kasutamise või väärkasutuse või muude ebasoodsate tingimuste korral toimiksid need nõuetekohaselt ega põhjustaks põhjendamatut riski ohutusele ja turvalisusele. • AIS aktorid peavad tagama kõigi AIS-i elukaares kasutatud andmekogumite, protsesside ja tehtud otsuste jälgitavuse, et võimaldada AIS-i tulemuste ja nende põhjenduste analüüsi ja mõistmist. • Kõik AIS aktorid peavad AIS-i elukaare igas etapis rakendama süstemaatilist riskihaldust ka eetiliste ja sotsiaalsete nõuete täitmise osas [90, Lõik 1.4].
VASTUTUS
• AIS-i osapooled peavad vastutama AIS-i korrektse toimimise ja eelnevalt kirjeldatud nõuete järgimise eest. • Nõuete järgmise vahendid ja meetmed peavad kogu AIS-i elukaare vältel arvestama konkreetse süsteemi kasutamise konteksti ja vastama tehnika tasemele [90, Lõik 1.5].

Eelnevast nähtub, et OECD eetilised nõuded AIS-ile on sarnased EL-i nõuetele. Samuti on OECD poolt sõnastatud eetiliste ja sotsiaalsete nõuete täitmine AIS-i arendajate ja käitajate kohustus. Autori hinnangul on OECD käsitluse olulisemaks erinevuseks võrreldes teiste käsitlustega OECD lähenemise selgem elukaare-põhisus, mida on täpsemalt avatud käesoleva töö alapeatükis 4.2.

3.4 EN soovitused seoses AIS-i mõjudega

8. aprillil 2020. aastal võttis Euroopa Nõukogu (EN) vastu soovitused seoses AIS mõjudega inimõigustele [91]. EN seisukohad on olulised, kuna seovad AIS eetika küsimused otseselt inimõiguste ja põhivabaduste tagamise ja kaitsega[1].

EN nõuab, et AIS-ide käitamisel tuleb (i) kaitsta ja austada inimeste põhiõigusi ja -vabadusi, (ii) tagada kinnipidamine diskrimineerimise keelust, (iii) tagada AIS-ide töökindlus ja turvalisus, (iv) tagada läbipaistvus, neutraalsus ning vaimne terviklikkus ning (v) tagada AIS-ide toimimine kasutajate kontrolli all [92]. Vastavate nõuete täitmine ja põhiõiguste kaitse on osapoolte ühine kohustus, mis tuleb tagada läbi tasakaalustatud regulatsioonide, AIS kavandamise ning andmestiku analüüsi [92].

AIS arendajatelt ja käitajatelt nõuab EN inimeste põhiõiguste austamist [91, lk 10, 14]. Eraldi rõhutab EN inimeste privaatsust, andmete turvalisust ja kallutatuse vältimist võimaldava andmehalduse vajadust ning kohustust tagada otsustuste vastuvõtmise protsessi läbipaistvus, vastutus ning õiguste kaitse võimalus [91, lk 15–16].

3.5 IEEE nõuded eetikat arvestavale AIS-ile[2]

IEEE „Eetikat arvestava disaini“ („*Ethically Aligned Design*“) algatus tegeleb AIS-i seonduvate eetiliste ja sotsiaalsete mõjude mõistmise ja haldamisega [1, lk 10]. IEEE algatus koondab valdkonna asjatundjate parima teadmise, eesmärgiga formuleerida

[1] EN on inimõiguste ja põhiõiguste kaitsega tegelev organisatsiooni, mis on ellu kutsunud Euroopa inimõiguste ja põhivabaduste kaitse konventsiooni ja Euroopa Inimõiguste Kohtu (EIK). EN seisukohad asetavad AIS-ide eetilised mõjud konventsiooni konteksti ja viitavad võimalusele saada põhiõigusi rikkuvate AIS-i mõjude eest kaitset EIK-s.

[2] IEEE kasutab mõistet “autonoomsed ja intelligentsed süsteemid“, mida autor kasutab magistritöö terminoloogilise ühtluse tagamiseks läbivalt.

nende kaudu AIS-ide standardiseerimise aluseks olevad soovitusliku nõuded. Käsitluse haare on muljetavaldav – AIS-i tehnilistest nõuetest kuni nendega loodava heaolu („*well-being*“) mõõtmise metoodikate väljapakkumiseni [93]. Lähenemise aluseks on klassikaline eetika, kuid soovitused hõlmavad praktilisi juhiseid väärtuste AIS-idesse integreerimise ning eetikast juhinduva teadus- ja arendustegevuse osas [1, lk 8].

IEEE hinnangul ei ole AIS-ide käitamine üksnes tehniline tegevus, vaid selle käigus tuleb teha eetilisi valikuid ja otsuseid [94, lk 69]. IEEE „Eetikat arvestava disaini“ aluspõhimõtetest tuleneb üheselt, et eetilised ja asotsiaalsed kaalutlused on olulised ka inseneriteaduste seisukohalt. AIS-ide eetiline ja sotsiaalne mõõde ei ole üksnes sotsiaalteadlaste ja filosoofide mure, kuna sotsiaalteadustest on küll abi tehnoloogiaga seonduvate probleemide mõistmisel, kuid neist ei piisa selliste probleemide lahendamiseks [1, lk 197]. IEEE seisukohtadest tulenevalt on AIS-ide kavandamisel oluline silmas pidada süsteemide laiemat sotsiaalset mõju (nt inimõigustele ja sotsiaalsele heaolule) ning konkreetselt tagada kavandatava süsteemi toimimise arusaadavus, selgitatavus ning kontrollitavus kolmandate isikute poolt [1, lk 249].

IEEE nõustub, et eetilise AIS-i arendamine ja käitamine on võimalik üksnes siis, kui riskide haldamine hõlmab kogu konkreetset STS-i ja selle osaks oleva AIS-i elukaart [1, Lõik 62–63; 78–79; 155–156]. IEEE soovitab, et AIS-i riskide haldamine algaks elukaare võimalikult varajases faasis, kuna nii saab riske hallata läbivalt ja efektiivsemalt [1, lk 69]. IEEE soovitab süsteemi elukaare erinevate faaside kohta välja töötada vastava faasi riske haldava ja selgitavad poliitikad [1, Lõik 155; 199]. Seega tuleb ka AIS-i eetilisus, vastutus ja läbipaistvus tagada läbi kogu süsteemi elukaare.

AIS-i käitamisega seonduvate IEEE standardite eesmärgiks on AIS-ide eetiline arendamine ja käitamine, mis juhindub järgmistest nõuetest [1, lk 6–9; 22–32]:

Tabel 3. IEEE nõuded eetikat arvestavale AIS-ile [1, lk 6–9; 22–32].

INIMÕIGUSED
• AIS-id tuleb kavandada ja rakendada austama ja järgima inimõigusi. • AIS peab olema kontrollitavalt ohutu ja turvaline kogu elukaare lõikes. • Kui AIS tekitab kahju, siis peab olema võimalik tuvastada kahju põhjus. • Lähitulevikus ei tohiks AIS-idele omistada inimestele omaseid õigusi ja omadusi. • AIS-id tuleb hoida inimeste kontrolli alla. • Välja tuleb töötada eelnevaid eesmärke toetavad standardid ja õigusnormid.
HEAOLU
• AIS-ide kavandamisel ja kasutamisel tuleb prioriteediks seada kasutajate mõõdetav heaolu.
VASTUTUSTUNDLIKKUS JA VASTUTUS
• AIS-ide arendajad ja käitajad peavad olema vastutustundlikud ja neid peab saama vastutusele võtta. • AIS-ide arendajad ja käitajad peavad arvestama kultuurilise mitmekesisusega ja väärtuste varieeruvusega erinevates kogukondades. • Kehtestada tuleb selged keelud tegevuste ja otsuste osas, mida ei tohiks AIS-idele üle anda. • Kaaluda tuleks AIS-ide registri loomist, mis võimaldaks iga komponendi osas arendaja ja vastutava isiku tuvastamist.
LÄBIPAISTVUS
• AIS-ide toimimine peab olema hinnatavalt läbipaistev ja arusaadav. Selleks on väljatöötamisel IEEE standard P7001™. • AIS-i kasutatavad andmed ja algoritmid peavad olema järelevalvajatele kättesaadavad ja allutatud riskihaldusele ja rangele („*rigorous"*) testimisele. • AIS-ide käitamisel tuleb säilitada otsuste kontrollimist võimaldav auditijälg. • Avalikkusele peab olema teada, kes langetab AIS-iga seonduvaid eetilise mõõtmega otsuseid.
TEADLIKKUS VÄÄRKASUTUSEST
• AIS-ide väärkasutuse riski tuleb minimeerida nii teadlikkuse tõstmise, hariduse kui õiguslike regulatsioonide kehtestamise kaudu [1, lk 6–7]

Eelnevalt kirjeldatud nõuete järgimise protsessi kirjeldamiseks ning sertifitseerimise toetamiseks on IEEE alustanud 14-st AIS-i probleemistikuga tegelevast standardist koosneva IEEE P7000™ standardiseeria väljatöötamist [95]. IEEE tahab 2020. aastal jõuda maailmas kõige esimesena AIS toodete, teenuste ja süsteemide sertifitseerimiseni

kolmes valdkonnas: AIS-i läbipaistvus, AIS-i vastutustundlikkus ning AIS-i algoritmi kallutatus(e haldamine) [93, lk 4]. Väljatöötatavad standardid ja neil rajanev sertifitseerimisprotsess hakkavad oluliselt suunama eetilise AIS-i arengut ja panevad paika valdkonna parimad praktikad.

IEEE nõuded ei erine oluliselt eelnevalt käsitletud poliitiliste ja rahvusvaheliste organisatsioonide käsitlustes esitatud nõuetest. Seega on AIS-iga seonduvad poliitika algatused kooskõlas valdkonna praktikute vastavatele algatustega ja vastupidi. Autori hinnangul toetab poliitikute, ametnike ja inseneride sarnane lähenemine tõdemust, et AIS-ide eetiliste ja sotsiaalsete riskidega arvestamise vajadus on üldtunnustatud ning sellega tuleb arvestada AIS-ide arendamisel ja käitamisel.

3.6 Eetiline AIS

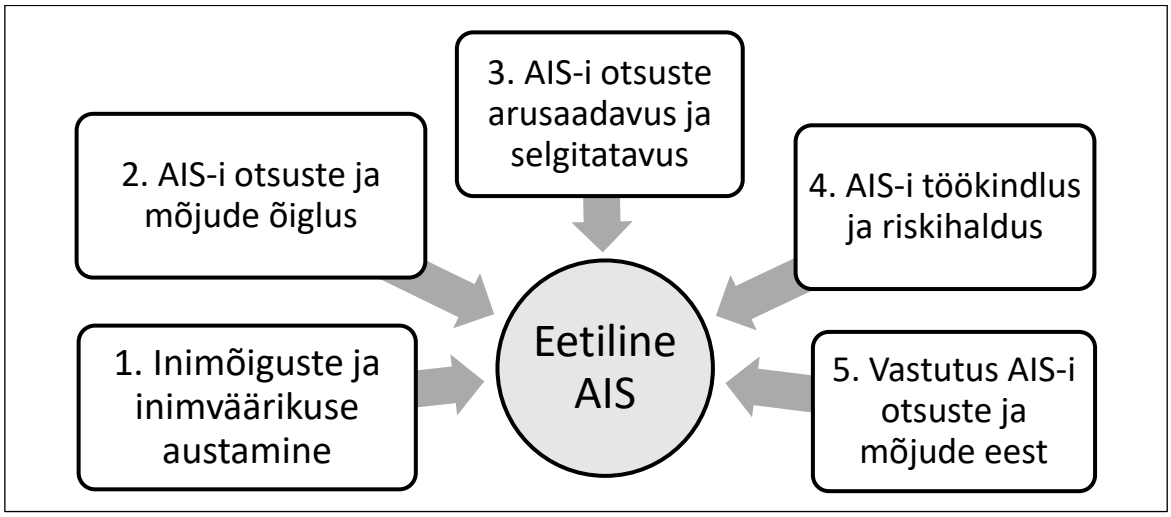
Käsitletud algatused ei omista ega nõua AIS-ilt iseseisvat võimet saada aru eetilistest nõuetest või olla eetiliste väärtuste kandjaks. Seega ei tähenda AIS-i eetilisuse nõue süsteemi võimet aru saada eetilistest nõuetest, langetada nende alusel väärtusotsuseid ning kanda vastutust ebaeetiliste otsuste eest.

Kõigi kirjeldatud algatuste järgi on AIS-i eetiliste ja sotsiaalsete nõuete täitmine süsteemi arendaja ja käitaja kohustus. Autori hinnangul asetavad käsitlused vastutuse eetiliste ja sotsiaalsete nõuete täitmise eest primaarselt AIS-i käitajale, kelle eesmärkide saavutamise raames AIS ühiskonda mõjutab. Sekundaarselt on vastavad kohustused asetatud AIS-i arendajale, kes peab nõudeid täitma selleks, et AIS-i käitajal oleks võimalik vastavaid nõudeid täita. Kõik käsitletud algatused võimaldavad nii AIS arendaja kui ka käitaja vastutust ka selliste AIS-i (ebasoovitavate) mõjude eest, mis ei ole arendaja või käitaja poolt põhjustatud teadlikult, vaid teadmatusest. Selline vastutuse jaotamine ei ole eetilise hinnangu omistamise seisukohalt ehk korrektne, kuid sotsiaalsest aspektist võib olla otstarbekas ja põhjendatud AIS-i arendamise ja käitamise negatiivsete mõjude eest vastutuse asendamine nendele osapooltele, kes saavad AIS-ide arendamisest ja käitamises vahetult enim eeliseid.

Käsitletud AIS eetika algatused panevad nõuete keskmesse inimese, tema vabadused ja õigused. Käsitlustes esitatud nõuded sisaldavad nii eetilisi kui ka sotsiaalseid nõudeid, neid alati korrektselt eristamata. Samas on sõnastatud eetilised ja sotsiaalsed nõuded

sarnased ja seetõttu võib autori hinnangul käsitleda neid nõudeid põhiosas universaalsetena. Erisused on ennekõike rõhuasetustes või inimõigustega piirnevates küsimustes[1]. Kõikide käsitluste kohaselt on aga lisaks eetilistele ja sotsiaalsetele põhinõuetele oluline ka nende kaitset tagav nõue, et süsteemi arendaja ja käitaja peavad tagama AIS-i toimimise vastavalt seatud eetilistele ja sotsiaalsetele nõuetele, haldama AIS-iga kaasnevaid riske ja kandma vastutust nende realiseerumise korral.

Eelnevate nõuete alusel käsitleb autor eetilist AIS-i kui sellist AIS-i, (1) mille käitaja on tuvastanud konkreetses käitamise kontekstis asjakohased eetilised ja sotsiaalsed nõuded, (2) määranud nende alusel nõuded süsteemi toimimisele ja (3) mis täidab neid nõuetekohaselt. Vastavad nõuded erinevad käsitluste lõikes [10, lk 7], kuid on teatav ühisosa, mille alusel esitab autor järgmise eetilise AIS-i põhinõuete skeemi:



Joonis 6. Eetilise AIS-i põhinõuded.

Autori hinnangul eristuvad eetilise AIS-i nõuete seast (3) arusaadavuse ja selgitatavuse ning (4) töökindluse ja riskihalduse nõude, mis on selgemalt insenertehnilised nõuded. AIS-i otsuste inimesele arusaadavuse ja selgitatavuse nõue on spetsiifiline nõue sellise süsteemi toimimisele. Selle täitmise lähtepunktiks on õigete tehniliste lahenduste valik. Samuti kuulub inseriteaduste valdkonda nõue tagada AIS-i toimimine vastavalt seatud nõuetele. See nõue leiab eetilise AIS-i valdkonnas läbivat rõhutamist, kuna osapooled

[1] Mõned raamistikud toovad sisse ka üldisemad nõuded nagu sotsiaalse heaolu kasvatamine, solidaarsus, keskkonnakaitse, säästev areng jms [10, lk 7], [72][72].

näevad eetiliste ja sotsiaalsete nõuete (piisava) määratlemise ja riskihalduse puudumises AIS-i valdkonda läbivat probleemi. Eetiline AIS ei tohi olla üksnes deklaratsioon, vaid seatud nõuete täitmist peavad toetama töökindluse ja riskihalduse meetmed. Teised nõuded kuuluvad suuremal määral eetika või ühiskonnateaduste valdkonda.

Eetiliste ja sotsiaalsete nõuete määratlemine ja täitmine on keeruline ka inimesele. Ka on eetilised väärtused ajas muutuvad ning kultuuriti erinevad. Asjakohased eetilised väärtused, neile omistatav osakaal ning võimalike konfliktide lahendus sõltub konkreetsest olukorrast. Seetõttu ei ole võimalik arendada ega käitada universaalselt eetilist AIS-i, vaid oluline on analüüsi faasi tuvastada konkreetse AIS-i käitamise kontekstis olulised eetilised ja sotsiaalsed nõuded ja tagada nende järgimine nõuetena süsteemi toimimisele [1, lk 33, 90, 166]. Tulemusliku analüüsi teostamiseks peab analüütikul olema arusaamine üldistest nõuetest eetilisele AIS-ile, analüüsi meetoditest ning suutlikkus tuvastada konkreetse süsteemi puhul üldnõudeid täiendavad või neist kõrvalekalduvad eetilised ja sotsiaalsed erinõuded.

Eetilised ja sotsiaalsed nõuded AIS-ile ei ole valdavas osas küll täna formuleeritud kohustuslike õiguslike või tehniliste nõuetena. Samas eksisteerib ka täna kohustus tagada toote ja teenuse ohutus [80, lk 3] [96, Lõik 5(1)], üleüldine teiste isikute kahjustamise keeld [97, Lõik 1043] ning diskrimineerimise keeld [98, Lõik 2]. Seega võib eetilisuse nõuetele mittevastav AIS, eriti tegeliku kahju ilmnemisel, tekitada AIS käitlevale organisatsioonile varalist (nt trahvid, kahjuhüvitised, süsteemi muutmise kohustus) kui ka mainekahju [99].

Autori hinnangul on märgata AIS-i arendamise ja käitamise eetiliste ja sotsiaalsete nõuete konkretiseerumist ja oodatavat karmistumist. Vastavate nõuete järgimine muutub ajas järjest olulisemaks nii tulevikus loodavate kui ka juba toimivate süsteemide jaoks[1].

[1] Kui arvestada keskmise IT-süsteemi kasulikuks elueaks[1] 3-10 aastat [156], siis võib eeldada, et paljude täna kasutavate ja eriti täna kavandatavate süsteemide eluea jooksul muutuvad AIS-i eetilisuse nõuded otseselt (läbi õigusaktide nõuete) või kaudselt (läbi standardite ja sertifitseerimise) kohustuslikuks. Seega tuleb nende nõuetega arvestada juba täna, kuid nende olulisus on kasvamas.

4 Eetilise AIS-i elukaare haldamise meetmed ja protsess

Käesolev peatükk annab ülevaate eelmises peatükis käsitletud AIS eetiliste ja sotsiaalsete nõuete täitmise haldamise meetmetest ning protsessidest. Autor pakub käesolevas peatükis välja ka eetilise AIS-i haldamise elukaare kirjelduse.

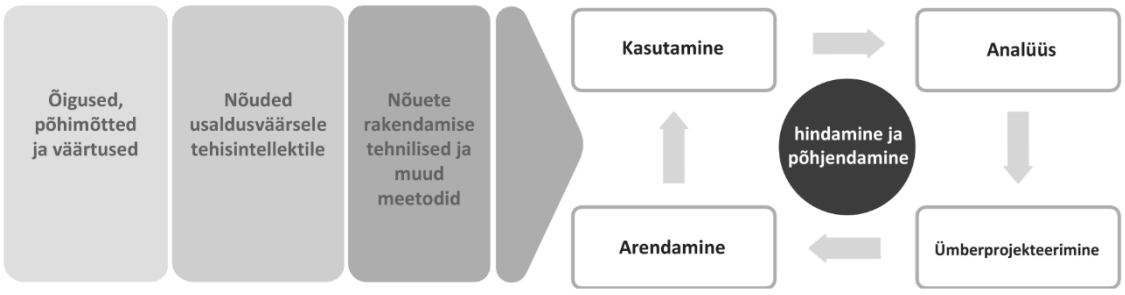
4.1 EL-i usaldusväärse AIS-i tagamise meetmed

4.1.1 Usaldusväärse AIS-i tagamise meetodid

AI HLEG hinnangul on usaldusväärse AIS-i põhinõuded universaalselt rakendatavad kõikidele AIS-idele [36, lk 17].

Samas tuleb usaldusväärse AIS-i põhinõuete täitmiseks hoolikalt analüüsida konkreetse süsteemi rakendamise konteksti ja vastavalt sellele välja töötada põhinõuete tasakaalustatud rakendamisel rajanev konkreetse AIS-i eripärasid arvestav riskide haldamise süsteem [79, lk 3]. Nt võib muusikasoovituse algoritmide puhul aktsepteerida oluliselt suuremaid riske kui meditsiinilisi diagnoose määravate algoritmide puhul. Seega lasub adekvaatsete ja proportsionaalsete meetmete valiku võimalus ja vastutus AIS-i arendajal ja käitajal.

AIS-i usaldusväärsuse tagamine ei ole AI HLEG hinnangul ühekordne tegevus, vaid kõigi põhinõuete järgimine peab hõlmama kogu konkreetse AIS-i elukaart. Tekkivatest kompleksetest vastandmõjudest ja vastuoludest arusaamine ning adekvaatsete kompromisside tegemine nõuab kogu AIS-i elukaare ja mõjude terviklikku analüüsi. AI HLEG kirjeldatud AIS-i elukaar, mille kõigis etappides tuleb tagada põhinõuete täitmine, on esitatud järgmisel joonisel [36, lk 23]:



Joonis 7. AI HLEG usaldusväärse AIS-i elukaar [36, lk 23].

Joonisel esitatud usaldusväärse AIS-i elukaare skeemi probleemiks on autori hinnangul eetilise ja sotsiaalse sisendi staatilisus. Skeem eeldab, et usaldusväärse AIS-i sisendiks olevad eetilised ja sotsiaalsed nõuded ei muutu pärast analüüsi ja rakendamist. See lähenemine eirab asjaolu, et eetilised ja sotsiaalsed nõuded on ajas muutuvad [1, lk 164–166], [22, lk 311]. Samuti on eetilised ja sotsiaalsed nõuded erinevates rakendamise kontekstides ja kultuuriruumides erinevad [1, lk 166; 193]. Seetõttu tuleb autori hinnangul AIS elutsüklis dünaamiliselt hinnata ka seda, kas süsteemile määratud eetilised ja sotsiaalsed nõuded on püsinud samana või tuleb vastavaid nõudeid ja nende täitmise meetmeid süsteemi käitamise käigus muuta või täiendada.

Usaldusväärse AIS-i põhinõuete täitmise tagamiseks ning AIS-i eetiliste ja sotsiaalsete riskide haldamiseks võib AI HLEG hinnangul kasutada nii tehnilisi kui ka muid meetmeid.

4.1.2 Usaldusväärse AIS-i tagamise tehnilised meetmed

Usaldusväärse AIS-i põhinõuete järgimise tehniliste meetmete näidiskataloog on AI HLEG käsitluses järgmine:

1) **Usaldusväärse AIS-i arhitektuurid**

AI HLEG ei paku usaldusväärse AIS-i arhitektuuri osas välja ei konkreetseid arhitektuure, spetsiifilisi nõudeid ega parimaid praktikaid. Piirdutud on üksnes tõdemusega, et AIS-i arhitektuuri aluseks võiksid olla valge nimekirja reeglid (käitumised või olekud, mida süsteem peab alati jälgima), musta nimekirja reeglid (käitumised või olekud, mida süsteem ei tohiks kunagi kasutada) ning nende kombinatsioonid. Eraldi on välja toodud õppimisvõimeliste süsteemide eripärad ning „taju-kavanda-tegutse“ („*sense-plan-act*“) tsükli olulisus nende käitamisel [36, lk 24]. Autori hinnangul puudub AI HLEG vastavas käsitluses igasugune tehniline mõõde. Süstemaatiline arusaamine selle nõude tehnilisest tähendusest puudub ka muus kirjanduses[1]. Ilmselt on arhitektuurinõuete meetodi üldine küpsusaste madal ka AI HLEG enda hinnangul. Autori ootaks tulevikus siiski AI HLEG-ilt põhjalikumat AIS-i

[1] Kirjanduses esitatud arhitektuuri käsitlused seonduvad ennekõike algoritmi selgitatavuse tagamisega (nt nõue esitada algoritmi otsuste põhjendused inimesele arusaadaval viisil naturaalkeeles) [100, lk 2].

arhitektuuriküsimuste käsitlust, sest üldtunnustatud arhitektuuripraktikate kujunemine toetaks sisuliselt usaldusväärse AIS arendamist.

2) **Lõime-eetika ja õigusriigi põhimõte**

Lõime-eetika ja õigusriigi meetod nõuab, et kooskõla eetika- ja õigusnormidega tuleb AIS-i sisse projekteerida. AIS-i käitajad peavad süsteemi kavandades hindama süsteemi eetilist ja sotsiaalset mõju ja tuvastama selle tegevusele kohalduvad õigusnormid, et ära hoida nende rikkumist ja minimeerida rikkumiste mõjusid [36, lk 24]. Autori hinnangul on ka selle meetodi puhul pigem tegemist deklaratsiooniga, äärmisel juhul mõne usaldusväärse AIS-i põhinõude täpsustusega. Olemuslikult on keeruline mõista, mida võiks AIS-i arhitektuuri juures tähendada eetika- ja õigusnormide sisseprojekteerimise nõue. Kindlasti ei tähenda vastav nõue, et AIS ise oleks suuteline aru saama vastavatest nõuetes. Pigem on tegemist nõudega, mis rakendub süsteemi elukaare kavandamisel ja elluviimisel osalevatele osapooltele, kellel lasub kohustus tagada AIS-i toimimise ja mõjude vastavus süsteemile kohalduvatele eetilistele ja sotsiaalsetele nõuetele.

3) **Arusaadavus ja selgitamise kohustus**

AIS-i usaldusväärsuse eelduseks on võimalus aru saada, miks süsteem käitus teatud viisil ja miks see jõudis konkreetse otsuseni. AIS-i käitaja peab suutma otsuste vastuvõtmise põhjustest ise aru saada ja suutma neid selgitada ka teistele süsteemist mõjutatud osapooltele. AIS-i otsustest mõjutatud isikutele tuleb anda teavet süsteemi suutlikkuse ja piirangute kohta. Eraldi on nõutav, et isikutele tuleb teada anda, et nad kasutavad AIS-i [36, lk 27]. AI HLEG ei esita selgituskohustuse osas konkreetseid juhiseid, kuidas seda nõuet täita, kuid osundab probleemi olulisusele ning täiendavate teadusuuringute vajadusele [36, lk 24–25]. Autori poolt teostatud kirjanduse analüüsi kohaselt on AIS-i otsuste selgitamise meetodite osas välja pakutud palju erinevaid vahendeid (vt alapeatükke 5.2 ja 5.3), kuid välja ei ole kujunenud üldiselt aktsepteeritavaid parimaid praktikaid või standardeid [100, lk 2]. Seda enam oleks autor soovinud, et AI HLEG esitanud kasvõi soovitusi metoodikate osas.

4) **Katsetamine ja valideerimine**

AI HLEG on seisukohal, et AIS-ide teatava autonoomia ja kontekstist sõltuva toimimise tõttu ei ole süsteemi toimimise ühekordne testimine või kontrollimine piisav. AI HLEG

osundab vajadusele testida ja kontrollida süsteemi ja selle kõigi komponentide – andmete, eeltreenitud mudelite, keskkonna – nõuetekohast toimimist nii AIS-i süsteemi arendamise, algoritmi treenimise, kasutuselevõtmise kui ka käitamise faasis, et veenduda süsteemi stabiilses ja töökindlas toimimises vastavalt seatud eetilistele ja sotsiaalsetele nõuetele [36, lk 25]. Autor oleks soovinud, et AI HLEG oleks seda järeldust põhjalikumalt selgitanud ning esitanud täpsema käsitluse selles osas, millises ulatuses on valdavalt inimese loodud algoritmide ja süsteemide testimise ja kontrollimise meetodid piisavad, kus ilmnevad selliste meetodite võimekuse piirid ning millises ulatuses on vaja välja töötada uusi meetodeid valdavalt masina loodud algoritmide testimise ja kontrollimise spetsiifiliste probleemide lahendamiseks.

5) **Teenuse kvaliteedi näitajad**

AIS-i loomise algfaasi iseloomustab sageli uudishimu, mängulisus ning sellest tulenevad eksperimentaalsed meetodid – kui eesmärgiks on avastada andmehulgast inimmõistusele märkamatuks jäänud seosed, siis ei ole vastava tegevuse alguses võimalik täpselt määratleda nõudeid otsitavale parameetrile ning täpsusele; võib juhtuda, et otsija küsimus jääb vastuseta ja vastuse leiab küsimus, mida eksperimendi alguses ei osatud sõnastada. Selles kontekstis on oluline AI HLEG seisukoht, et vaatamata algfaasi eksperimentaalsusele tuleb AIS-i juurutamisel siiski rangelt määratleda nõuded AIS-ile ja neid käitamisel järgida: seda nii AIS-i funktsionaalsete ja mittefunktsionaalsete nõuete kui ka süsteemi väljatöötamisel rakendatud kvaliteedikontrolli- ja riskihaldamise meetmete osas [36, lk 25]. Tarkvaranõuete defineerimise ja mõõtmise olulisuse rõhutamine ka valdavalt masina loodud algoritmide kontekstis on autori hinnangul oluline selleks, et vältida käitajale mittemõistetavate AIS-ide eksperimentaalset, spekulatiivset või teadmatusel põhinevat kasutamist inimeste põhiõigusi mõjutavate otsuste tegemisel.

AI HLEG küll rõhutab, et usaldusväärse AIS-i tagamise konkreetsete meetmete valik on konkreetse AIS-i arendaja ja käitaja ülesanne. Siiski on autori hinnangul kahetsusväärne, et AI HLEG on usaldusväärse AIS-i tagamise tehnilised meetmed markeerinud üksnes kontseptuaalsel tasemel ega ole meetmete täpsema sisu või valiku osas andnud sisulisi soovitusi. Vastavad soovitused abistaksid AIS-i arendajaid ja käitajaid vastavate nõuete täitmisel.

4.1.3 Usaldusväärse AIS-i tagamise muud meetmed

Usaldusväärse AIS-i põhinõuete järgimise muude meetmete näidiskataloog on AI HLEG käsitluses järgmine:

1) **Õigusnormid**

AI HLEG peab usaldusväärse AIS-i tagamisel oluliseks vastavust õigusaktide nõuetele, kuid jätab õiguslike nõuete konkretiseerimise teistele osapooltele [36, lk 25]. Autori hinnangul on siinkohal oluline rõhutada, et ehkki usaldusväärset AIS-i reguleerivaid õigusmuudatusi on oodata tulevikus [79, lk 5], [80, lk 15], siis juba täna reguleerivad AIS-ide arendamist ja käitamist erinevad õigusaktid valdkonnapõhistest nõuetest kuni üldiste vastutust reguleerivate normideni. Seega tuleb juba täna AIS-ide käitamisel arvesse võtta kehtivaid üldisi õigusnorme, mille ajalooline tehnoloogianeutraalsus tagab, et valdavat osa tõusetuvaid õiguslikke küsimusi ja riske on võimalik juba täna arvestatava täpsusega lahendada. Seega puudub praegu ülekaalukas vajadus nõrka tehisintellekti kasutavat AIS-i üldiselt reguleerivate õigusnormide järgi [81, lk 8].

2) **Käitumisjuhendid**

AI HLEG käsitleb usaldusväärse AIS-i tagamise meetmena käitumisjuhendeid, mis võivad olla formuleeritud nii AIS-i arendaja või käitaja enda sisekordadena (nt üldised eetikakoodeksid, spetsiaalsed usaldusväärse AIS-i tagamise juhendeid) kui ka valdkonnas üldiseks kasutamiseks väljatöötatud juhendite ja standarditena [36, lk 25–26]. AI HLEG ei osunda konkreetsetele juhenditele või standarditele, mida AIS-i arendajatel ja käitlejatel oleks lihtne rakendada. Autori hinnangul saab aga eetilise AIS-i nõuetele vastavuse esmaseks kontrolliks kasutada ka eestikasuuniste endi põhinõudeid ja kontrollküsimustikku (vt alapeatükk 5.2).

3) **Standardid**

AI HLEG tunnustab standardite suutlikkust suunata osapoolte poolt AIS-idele esitatud nõudeid ja mõjutada seonduvaid ostuotsuseid. Eetikasuunised osundavad ka asjakohastele ISO standarditele ning Elektri- ja Elektroonikainseneride Instituudi (IEEE) standardiseeriale P7000™ (vt alapeatükki 5.2). AI HLEG soovitab kaaluda usaldusväärse AIS-i märgise väljastamist usaldusväärseks loetud AIS-ile või selle käitajale [36, lk 26]. Autor soovinuks eetikajuhistes näha nii põhjalikumat ülevaadet

kogu maailma standardiseerimisalgatustest kui ka AI HLEG selgemat visiooni ootuste osas selliste standardite sisule ja lähenemisele. Näiteks on Hiina vastavates raamdokumentides esitatud märksa põhjalikum rahvusvaheliste standardimisalgatuste kaardistus [75, lk 110].

4) **Sertifitseerimine**

Usaldusväärse AIS-i sertifitseerimine on usaldusväärse AIS-i märgise formaalsem edasiarendus. Olukorras, kus AIS-i omadustest arusaamine ja nende hindamine laiema avalikkuse poolt on keeruline, saab usalduse ja läbipaistvuse loomiseks kasutada sertifitseerimisorganisatsioone, kes omavad AIS-i auditeerimise kompetentsi ja saavad laiemale üldsusele kinnitada konkreetse AIS-i ja sellega seonduvate protsesside nõuetelevastavust. Sertifitseerimine ei vabastaks AIS-ide arendajaid ja käitajaid vastutusest süsteemi nõuetelevastava toimimise eest [36, lk 26]. Autori hinnangul tuleks kehtestada ka selged nõuded sertifitseerimisorganisatsioonide vastutusele, et ennetada „Suurele Majandussurutisele“ [101] kaasaaidanud krediidireitinguagentuuride minetustega [102, lk 44] sarnaseid sertifitseerimisnõuete rikkumisi AIS-i sertifitseerimisel.

5) **Juhtimisraamistike rakendamine**

AI HLEG soovitab AIS-ide arendajatel ja käitajatel rakendada juhtimisraamistikke, mis tagaksid eetiliste ja sotsiaalsete nõuete täitmise AIS-i arendamisel ja käitamisel ning vastutuse nõuete rikkumise eest. Konkreetselt soovitab AI HLEG kaaluda AIS-iga seotud eetikaküsimuste eest vastutava isiku või eetikanõukogu määramist. Sellise isiku või nõukogu rolliks on eetiliste ja sotsiaalsete nõuete täitmise üle järelevalve teostamine kui ka nõustamine nõuete täitmisel [36, lk 26]. Autori hinnangul saab vastava isiku või nõukogu rolli määratlemisel eeskujuks võtta teistes valdkondades avaliku huvi tagamist toetavaid sarnaseid meetmeid (nt privaatsusnõuete järgimist toetav andmekaitseametnik [85, Lõik 35–36, 37–39, 83], eetikakomitee kliinilise ravimiuuringu läbiviimisel).

6) **Haridus ja teadlikkus**

AI HLEG peab usaldusväärse AIS-i tagamise oluliseks meetmeks sidusrühmade teadlikkuse tõstmist AIS-i eetiliste ja sotsiaalsete nõuete ja nende täitmise meetmete osas. AI HLEG peab siinkohal primaarseks AIS-i otsustest mõjutatud isikute ja

ühiskonna teadlikkuse tõstmist A [36, lk 26]. Autor on aga seisukohal, et teadlikkuse tõstmine peab olema ka vastassuunaline – professionaalsusest tulenevalt on just AIS-ide arendajate ja käitajate suhtes kõrgendatud ootused, et nad saavad aru nendest sotsiaalsetest protsessidest ja tegevustest, mida nende arendatavad ja käitatavad süsteemid mõjutavad. On vähetõenäoline, et ühiskonnas kujuneb üleüldine teadlikkus AIS-i riskidest ja nende haldamisest. Pigem on vastavate riskide ja nende konteksti mõistmine ja adekvaatne riskihaldus ikkagi AIS-i arendaja ja käitaja kohustus.

7) **Sidusrühmade osalus ja sotsiaaldialoog**

AI HLEG toob ühe usaldusväärse AIS-i loomise meetmena välja ka sidusrühmade osaluse ja sotsiaaldialoogi [36, lk 26–27]. Kaasamine ja kodanikuühiskonna osalus on ka oluline EL-i väärtus. Olemuslikult on tegemist meetmega, mille kaudu realiseerib *inimese toimevõime* põhinõue – inimese tahte arvestamine on võimalik üksnes sisukas dialoogis inimese või laiemat gruppi esindava organisatsiooniga. Teisalt on see meede autori hinnangul ennekõike poliitikakujundajate tööriistaks. Vastav meede võib aga praktilist väärtust pakkuda selliste AIS-ide arendajatele ja käitajatele, millel on sedavõrd ulatuslik ühiskondlik mõju või kliendibaas, et mõjutatud eetiliste ja sotsiaalsete nõuete mõistmiseks ei piisa enam tavapärastest kliendisuhtluse tööriistadest.

8) **Mitmekesisus ja kaasavad projektimeeskonnad**

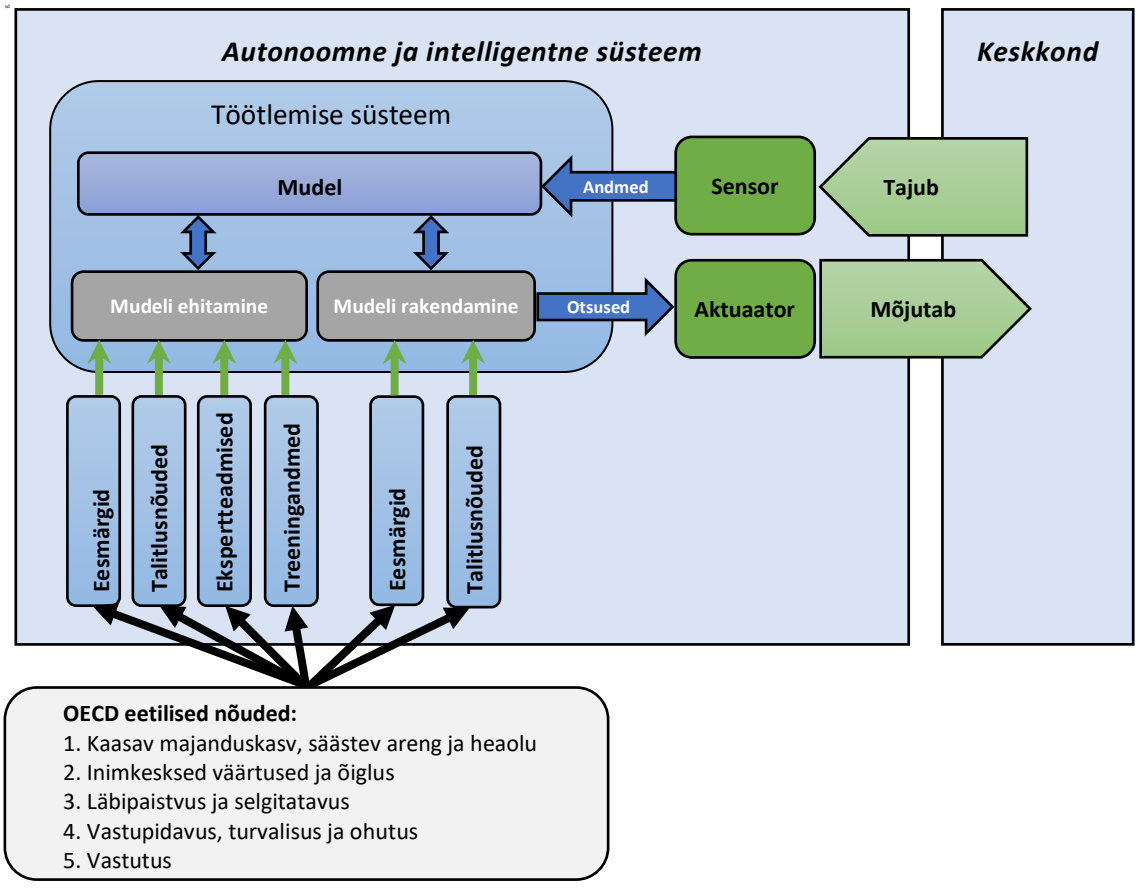
EL-i üheks põhiväärtuseks on ka mitmekesisuse austamine ning kaasamine. Seetõttu siis peab AI HLEG oluliseks, et AIS-ide arendamisse ja käitamisse on kaasatud süsteemist mõjutatud isikute ja kogu ühiskonna mitmekesisus [36, lk 27]. Mitmekesiste isikute osalemine tagab, et AIS-i arendamisel ja käitamisel võetakse arvesse erinevaid seisukohti, vajadusi ning ühiskondlikke piirjuhtumeid. Mitmekesisus ei ole üksnes sotsiaalne ja poliitiline deklaratsioon, vaid sisulise lähenemisega mitmekesisusele on tõepoolest võimalik vältida liiga kitsast lähenemisest põhjustatud eetilisi või sotsiaalseid probleeme [103], laiendada organisatsiooni kognitiivse taju diapasooni [104] ja tõsta organisatsioonide tulemuslikkust [105]; mitmekesisusega arvestamine on globaliseerumist arvestades isegi vajalik globaalselt edukate, erinevate kultuuriruumide ja -kontekstide nõudele vastavate AIS-ide arendamiseks ja käitamiseks.

Autori hinnangul on AI HLEG usaldusväärse AIS tagamise meetmete kirjeldus pigem mõtteraamistik, kui kõikehõlmav meetmete või tööriistade ülevaade. Samas on oluline

rõhuasetus, et usaldusväärse AIS-i tagamine ei sõltu vaid tehnilistest meetmetest, vaid olemuslikult tehnoloogia ja sotsiaalsete protsesside kokkupuutel ilmnevate eetiliste ja sotsiaalsete riskide haldamiseks saab kasutada nii tehnilisi meetmeid (nt arhitektuur, töökindlus), sotsiaalsed meetmeid (nt mitmekesisus, kaasavad projektimeeskonnad) kui ka meetmeid, mis sildavad lõhet tehnilise ja sotsiaalse domeeni vahel (nt haridus).

4.2 OECD eetilise AIS-i elukaare haldus

OECD seob AIS-i eetiliste ja sotsiaalsete nõuetega läbi süsteemi töötlemise süsteemi (nt masinõppealgoritmi) sisendite hoolika analüüsi ja mudeli poolt tehtavate otsuste nõuetelevastavuse hindamise. AIS-i arendamisel (sh mudeli ehitamisel) on sisendiks AIS-i eesmärgid, süsteemi talitlusnõuded, AIS-i käitamise valdkonna ekspertteadmised ning ajaloolised andmed. AIS-i otsuste hindamise aluseks on konkreetse süsteemi kavandamisel seatud eesmärgid ja süsteemi talitlusnõuded. Konkreetsete nõuete väljatöötamisel tuleb lähtuda OECD määratletud nõuetest (alapeatükk 3.3). OECD kirjeldab AIS-i ja eetiliste nõuete sidumist järgmiselt [14, lk 9]:

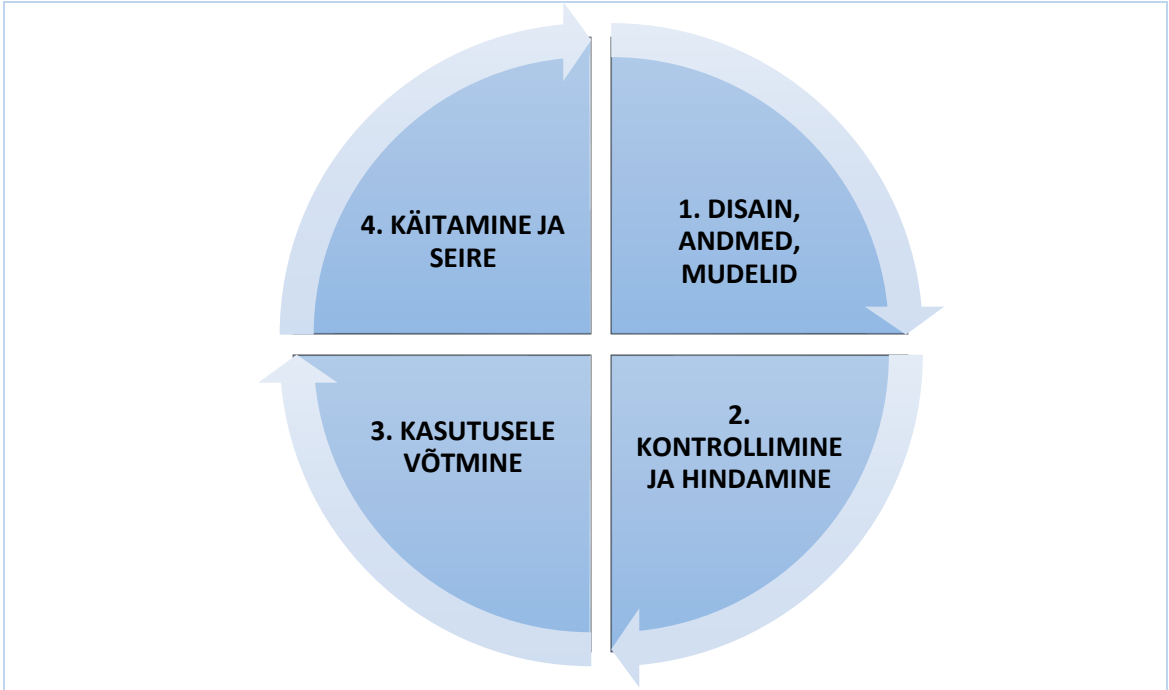


Joonis 8. OECD eetiliste nõuete sidumine AIS-iga [14, lk 9].

OECD lähtub teiste käsitletud lähenemistega võrreldes selgemalt AIS-i elukaare käsitlusest, kirjeldades AIS-i elukaare etappe, etappides ilmnevaid probleeme ja nende haldusmeetmeid [14, lk 2, 13]. OECD määratleb järgnevad AIS-i elukaare etapid:

> *„[AIS-i] elukaare etapid on: i) „disain, andmed, mudelid", mis on kontekstist sõltuv jada tegevusi, mis hõlmavad süsteemi analüüsi ja kavandamist, andmete kogumist ja töötlemist, samuti mudeli koostamist; ii) „kontrollimine ja hindamine"; iii) „kasutusele võtmine"; ja iv) „töö ja seire". Need etapid toimuvad sageli iteratiivselt ega pea olema järjestikused."* [14, lk 13], [90]

OECD käsitlusele vastav [14, lk 13] eetilise AIS-i elukaare kontseptsiooni on järgmine:



1. DISAIN, ANDMED, MUDELID
• AIS-i analüüs, kavandamine ja arhitektuuri • Andmete kogumine, analüüs (kallutatus jm), töötlemine ja dokumenteerimine · AIS-i mudeli ehitamine ja interpreteerimine
2. KONTROLLIMINE JA HINDAMINE
• Kasutusele võetavate mudelite kontrollimine ja täiendamine • Kasutusele võetavate mudelite testimine
3. KASUTUSELE VÕTMINE

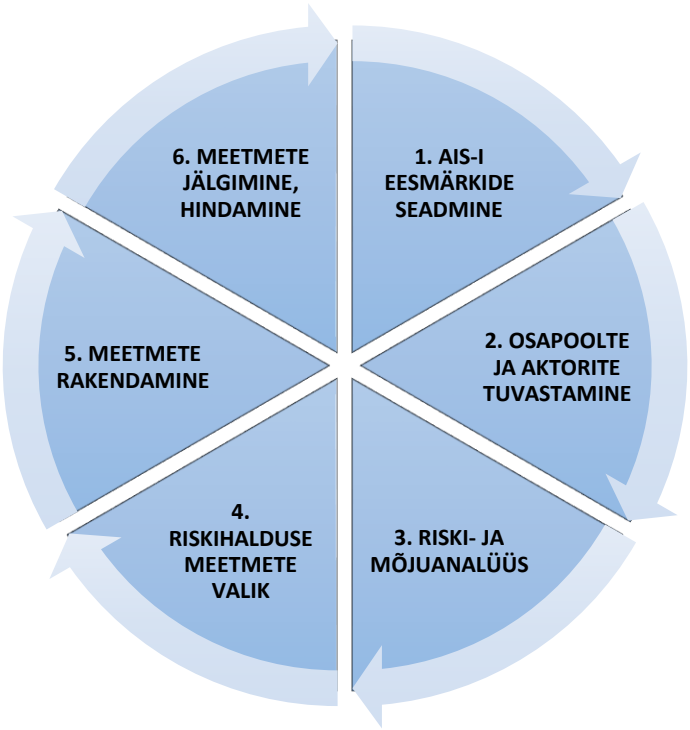
• AIS-i viimine kasutajateni • Varasemate süsteemidega tehnilise koostoime tagamine • Õigusaktide nõuetega vastavuse tagamine • Organisatsiooni muutuste juhtimine · Kasutuskogemuse jälgimine
4. KÄITAMINE JA SEIRE
• AIS-i käitamine ja kasutamine • Jooksev AIS-i kavandatud ja mittekavandatud mõjude (sh eetiliste) hindamine · AIS-i muutmine tegemine vastavalt mõjuanalüüsile

Joonis 9. OECD eetilise AIS-i elukaare kontseptsioon [14, lk 13].

Esitatud elukaare kontseptsioon on lihtsustav, kuna esitab ülevaatlikkuse huvides elukaare erinevad etapid iseseisvalt staatilistena ning teineteisele lineaarselt järgnevatena. Tegelikkuses võib iga kirjeldatud etappi vaadelda omakorda iteratiivse ja dünaamilise tsüklina. Samuti ei ole etappide mõjud lineaarsed, vaid hilisemad etapid võivad olla sisendiks varasema etapi iteratiivsesse tsüklisse (nt „kasutuselevõtmise“ faasis võivad ilmneda „disain, andmed ja mudelid“ faasis mittearvestatud mõjud, mis võib nõuda AIS-i korrigeerimist läbi varasemate etappide).

OECD kirjeldatud eetilise AIS-i nõuete järgimine peab arvestama konteksti, st nõuete järgimist ei saa hinnata universaalsest vaatepunktist, vaid hinnata tuleb konkreetse AIS toimimist konkreetses kontekstis. Lisaks tuleb arvestada, et igas AIS elukaare etapis avalduvad erinevad soovitud ja soovimatud mõjud.

Erinevate osapoolte huvide arvestamine ja mõjude hindamine läbi AIS-i elukaare on võimalik üksnes läbi kõikehõlmava riskihaldamise protsessi. OECD kirjelduse alusel on autor koostanud riskihaldamise tsükli kohta järgmise ülevaate [14, lk 14–17]:



1. AIS-I EESMÄRKIDE SEADMINE

- AIS-i eesmärkide ja mõjude seadmine, arvestades konteksti ja elukaare etappi.

2. OSAPOOLTE JA AKTORITE TUVASTAMINE

- Tuvastada igas elukaare etapis AIS-i poolt mõjutatud osapooled (vahetult mõjutatud isikud ja huvigrupid, kodanikuühiskond, järelevalveasutused jt).
- Määratleda AIS-i igas elukaare etapis osalevad AIS-i aktorid (andmeteadlased, analüütikud, süsteemi- ja tarkvarainsenerid, kasutajad jt).

3. RISKI- JA MÕJUANALÜÜS

- Süsteemi elukaare igas etapis osapooltele ning aktoritele avalduvate mõjude – positiivsete ja negatiivsete – ja riskide tuvastamine ja analüüs.
- Tuvastatud mõjude ja riskide hindamine eetilise AIS-i nõuete vastu.

4. RISKIHALDUSE MEETMETE VALIK

- Tuvastatud soovimatute mõjude ja riskide vältimise, vähendamise või ületamise meetmete valik ja väljatöötamine.
- Meetmed peavad arvestama ka mõjutatud osapoolte ja aktorite huvide, oskuste ja võimekustega.
- Meetmete valikul tuleb otsustada, milliseid meetmeid rakendada igas konkreetses AIS-i elukaare etapis.

5. MEETMETE RAKENDAMINE

- Väljatöötatud meetmete rakendamine – kes rakendab, millal rakendab, kuidas rakendab.

6. MEETMETE JÄLGIMINE, HINDAMINE

- Rakendatud meetmete mõjude jälgimine, hindamine ja tagasiside.

Joonis 10. OECD eetilise AIS-i riskihaldus [14, lk 14–17].

OECD käsitluse alusel on võimalik kirjeldada usaldusväärse AIS-i loomise ja käitamise elukaart. See algab konkreetse AIS-i käitamise aluseks olevate eetiliste ja sotsiaalsete eesmärkide ja osapoolte analüüsist, millele järgneb konkreetse AIS-i mõjudele vastava riskianalüüsi läbiviimine ning riskihaldamise meetmete kavandamine. Riskianalüüs ja rakendatavad meetmed peavad hõlmama kõiki AIS-i elukaare etappe.

4.3 EN inimõigusi arvestava AIS-i meetmed

EN lähenemine on teiste käsitletud lähenemisest rohkem suunatud poliitikate kujundamise suunamisele, eesmärkide ja põhinõuete sõnastamisele. Seetõttu on EN soovituste praktiline väärtus AIS-i arendajatele ja käitajatele piiratud.

Lisaks teiste raamistike juures käsitletud meetmetele on EN rõhutanud vajadust teostada AIS-i käitamise põhiõiguste mõjuanalüüse [91, Lõik 5.1, 5.2]. Nende läbiviimisel tuleks määratleda AIS-ist mõjutatud põhiõigused, hinnata vastavate mõjudega kaasnevaid riske põhiõigustele ning valida nende riskide haldamise meetmed.

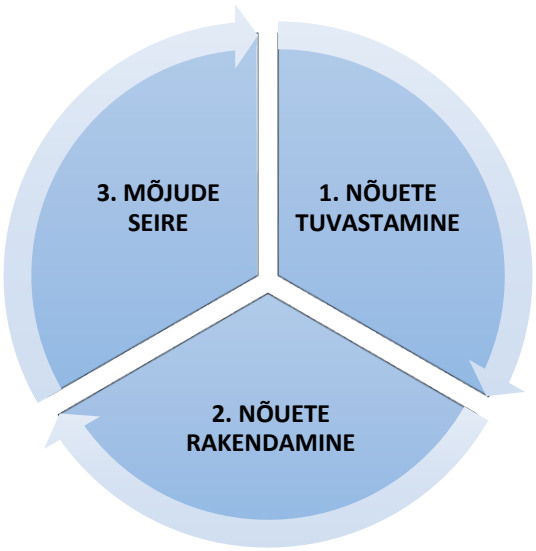
Põhiõiguste mõjuanalüüsid on vajalikud ennekõike kõrge riskiga AIS-ide käitamisel. EN peab kõrge riskiga AIS-ideks ennekõike avalike teenuste pakkumiseks kasutatavaid süsteeme [91, Lõik 3.6]. Seega on EN nõuded asjakohased ennekõike riikide ning avaliku sektori organisatsioonide jaoks.

4.4 IEEE eetilise AIS-i tagamise meetmed

4.4.1 Eetiliste ja sotsiaalsete nõuete AIS-i lõimimise tsükkel

IEEE suhtub kriitiliselt AIS-i ja selle eetika antropomorfsesse käsitlemisse. IEEE tõdeb, et masinatel ei ole (täna veel) suutlikkust aru saada eetilistest ja sotsiaalsetest nõuetes, vaid nad tegutsevad vastavalt sellele, kuidas nad on programmeeritud. Seega sõltub selliste nõuete täitmine AIS arendajate ja käitajate eetilisusest ja valikutest [1, lk 195].

AIS-i arendamisel soovitab IEEE juhinduda „eetika läbi disaini“ („*ethics by design*“) lähenemisest [1, lk 213]. IEEE poolt AIS-i arendajale ja käitajal antud soovituste alusel on autor loonud alljärgneva eetiliste ja sotsiaalsete nõuete AIS-i lõimimise sammudest [1, lk 33–54]:



1. NÕUETE TUVASTAMINE

- Tuvastada AIS-i käitamise kontekstis asjakohased eetilised ja sotsiaalsed nõuded.
- Tuvastada viisid, kuidas inimesed lahendavad asjakohaste nõuete vahelisi konflikte ja millised on ootused AIS-i poolt selliste konfliktide lahendamisele.
- Identifitseeritud eetilised ja sotsiaalsed nõuded ja konfliktide ületamise nõuded tuleb dokumenteerida.
- Eetiliste ja sotsiaalsete nõuete dünaamika tõttu tuleb võimaldada AIS-i eetiliste ja sotsiaalsete nõuete paindlik muutmine ning läbipaistvus muudatuste osas.

2. NÕUETE RAKENDAMINE

- Eetiliste ja sotsiaalsete nõuete AIS-is realiseerimise võimalusi on mitmeid (inimese determineeritud, masinõppepõhised, hübriid[1]).
- Nõuete täitmine peab toimima arusaadaval viisil ja süsteemi läbipaistvust toetades.
- Rakendada tuleb süsteemi nõuetele mittevastava toimimise riski haldavaid meetmeid.

3. MÕJUDE SEIRE

- Hinnata, kas rakendatud meetmed tagavad tuvastatud nõuete, sh osapoolte eetiliste ja sotsiaalsete ootuste täitmise.
- Hinnata, kas AIS-i mõjudes avaldub kallutatus, gruppide ebaõiglane kohtlemine või muu eetiliste ja sotsiaalsete nõuetega vastuolus olev mõju.
- Hinnata ja tagada, et AIS-i otsused on osapooltele arusaadavad, selgitatavad ja kontrollitavad (sh dokumenteeritus).

Joonis 11. IEEE eetiliste ja sotsiaalsete nõuete AIS-i lõimimise tsükkel [1, lk 33–54].

[1] Inimese determineeritud meetodi puhul määravad inimesed vastavad nõuded AIS-ile ettekirjutatud reeglitena; masinõppepõhise meetodi puhul õpetatakse AIS-ile eetiliselt ja sotsiaalselt lubatud ja mittelubatud käitumist andmete alusel; hübriidmeetod kombineerib mõlemat eelnevat meetodi [157]. Kõigil viidatud meetoditel on omad probleemid. Nt inimese determineeritud meetodit raskendab üldtunnustatud ja universaalselt sobiva eetilise teooria puudumine. Masinõppepõhise meetodi puhul võib AIS aga õppida eetilise käitumise asemel populaarset, enimlevinud käitumist [158, lk 3].

Kirjeldatud tsükli läbimine tagab, et AIS vastab käitlemise kõigis etappides käitlemise keskkonnas AIS-ile esitatavatele eetilistele ja sotsiaalsetele nõuetele. Eetilised ja sotsiaalsed nõuded AIS-i käitumisele võivad igas rakendusvaldkonnas, erinevates kogukondades ja eetikasüsteemides[1] olla erinevad ja ajas muutuvad [1, lk 164–166]. Seetõttu peab AIS-i väärtuste lõimimise tsükkel olema kontekstitundlik – süsteemi juurutamisel uues sotsiaalses kontekstis või sotsiaalse konteksti muutumisel tuleb tsüklit uuesti läbida; soovitatav on ka tsüklit jooksvalt käitada, et tagada süsteemi vastavus konteksti muutumisele või väärtuste muutumisele rakendusvaldkonnas või kogukonnas.

IEEE peab väärtuste AIS-i lõimimise oluliseks eelduseks organisatsiooni väärtusepõhist juhtimissüsteemi ja juhtimist, mis toetab eetiliste ja sotsiaalsete nõuete järgimist. Ainult sellises keskkonnas on võimalik luua väärtustega lõimitud AIS-e [1, lk 60–65]. Selliste juhtimispraktikate analüüs jääb käesoleva töö raamidest välja.

4.4.2 Eetilise AIS-i vastutustundlik arendamine ja käitamine

IEEE hinnangul on AIS-i vastutustundlikuks käitamiseks esmatähtis AIS-i ja selle komponentide seostatavus konkreetse arendajaga, arendajapoolsed adekvaatsed kasutusjuhised AIS-i käitajatele ning AIS-i nõuetekohase toimimise pidev tagamine.

Nende küsimuste lahendamiseks on IEEE välja pakkunud järgmised eetilise AIS-i arendamise ja käitamise nõuded ja meetmed [1, lk 155–156]:

Tabel 4. IEEE eetilise AIS-i arendamise ja käitamise nõuded [1, lk 155–156].

VASTUTUSE TAGAMINE
• AIS-id tuleb varustada tootja identifitseerimist võimaldava tähistusega, et alati oleks tagatud juriidilise vastutuse ahel. • Seadusandjad ja järelevalveasutused peavad looma efektiivsed menetlused, et vältida AIS-ide kuritarvitamist ja vastutuse vältimist AIS-e käitavate organisatsioonide poolt (sh kaaluda kohustuslikku kindlustust). • Kuna süsteemi projekteerija, tootja ja käitaja võivad olla erinevad isikud, siis tuleb luua konteksti arvestavad reeglid vastutuse jaotumiseks osapoolte vahel.

[1] IEEE toob oma käsitluses välja ennekõike Lääne eetika (Aristoteles, Kant), Idamaise eetika (Shinto, Konfutsius) ja Aafrika eetika (Ubuntu) [1, lk 2]. Lisaks tuleb autori hinnangul arvestada sellega, et ka nimetatud eetikasüsteemid ei ole homogeensed ning neis on olulisi erinevusi kultuuride, religioonide, ühiskonnagruppide, rahvuste, riikide jne lõikes.

KASUTUSJUHISTE DOKUMENTEERIMINE
• AI/S-ide arendajad ja käitajad peavad dokumenteerima ja tegema kättesaadavaks kirjalikud AIS-i kasutusjuhised (kuidas süsteem toimib, kuidas seda kasutada ja mõõta toimimise nõuetelevastavust, eeldused ohutuks kasutamiseks, sh koolitusvajadused). • Kasutusjuhised peavad olema piisavad, et AIS-i otsustest mõjutatud osapooltel ja kasutajatel oleks tegelikkusele vastav arusaamine AIS-i toimimisest ja suutlikkusest.
ALGORITMIDE JÄTKUV HOOLDUS
• AIS-i kahjulike mõjude ärahoidmise või kordumise vältimiseks peavad AIS-ide arendajad ja käitajad kaaluma algoritmide jätkuvat hooldust („*continued algorithm maintenance"*). • Algoritmide jätkuv hooldus peab olema üks AIS-ide käitamise olulisi lähtepunkte, mis tähendab süsteemi mõjude hoolikat jälgimist („*due diligence"*) ning ebasobivate mõjude jooksvat vältimist või kõrvaldamist. • Algoritmide jätkuvat hooldust peab teostama vastavate suutlikkustega isik või organisatsioon (vastava kohustuse edasidelegeerimine vastavate võimekusteta käitajale ei pruugi olla adekvaatne meede).
ÕIGUSRUUMI HARMONEERIMINE
• Vastutusega seonduv õiguslik regulatsioon tuleb rahvusvaheliselt harmoniseerida ja võimaldada piiriüleste õigusvaidluste pidamine. • Luua tuleb tootjavastutusega sarnana vastutusrežiimi, kus AIS-i defektide parandamine ei ole käsitletav vastutuse omaksvõtuna. • AIS-i kasutatavad andmed ja algoritmid peavad järelevalvajatele olema kättesaadavad, allutatud riskianalüüsile ja rangele („*rigorous"*) testimisele. • AIS-ide arendamisel ja käitamisel tuleb säilitada otsusete kontrollimist võimaldav auditijälg. • Avalikkusele peab olema teada, kes langetab süsteemiga seonduvaid eetilise ja sotsiaalse mõõtmega otsuseid.
TEADLIKKUS VÄÄRKASUTUSEST
• AIS-ide väärkasutuse riski tuleb minimeerida nii hariduse kui õiguslike regulatsioonide kehtestamise kaudu [1, lk 6–7]

Eelnevalt mainitud nõuetest ja meetmetest on autori hinnangul kõige olulisemad algoritmide jätkuva hoolduse, kasutusjuhendite kättesaadavaks tegemise ning AIS-i auditeerimise võimaldamise kohustused.

4.4.3 Eetilise AIS-i läbipaistvuse ja kontrollitavuse nõuded

IEEE hinnangul raskendab AIS-ide arendamise ja käitamise läbipaistmatus süsteemide eetilist käitamist ning selle järelevalvet. AIS-ide kontekstis on läbipaistmatuse allikaks IEEE hinnangul ennekõike algoritmide ja seonduvate otsustusprotseduuride puudulik dokumenteerimine, ebapiisav arusaamine, järelevalve ja sertifitseerimine [1, lk 68–70].

Läbipaistvusega seonduv eraldiseisev probleemistik kaasneb IEEE hinnangul selliste algoritmidega, mille funktsioneerimise põhimõtted ei ole süsteemi arendajale ja käitajale täielikult arusaadavad[1]. IEEE rõhutab, et tarkvarainsenerid peaksid selliseid läbipaistmatuid („*opaque*“) algoritme kasutama äärmise ettevaatlikkuse ning kõrgendatud nõuetega eetilisusele, kuna sellised komponendid põhjustavad tagajärgi, mida ei ole võimalik tavapäraste meetmetega täielikult inspekteerida, valideerida ja selgitada. Seetõttu kaasneb selliste süsteemidega kõrgendatud märkamatuks jäävate või ettenägematute vigade ja kahjulike mõjude risk. Vastutus selliste tagajärgede eest peab IEEE hinnangul jääma siiski AIS-ide arendajatele ja käitajatele [1, lk 70–71].

IEEE peab oluliseks kehtestada AIS-i läbipaistvusele selged nõuded, sh sätestada vastavad kohustused õigusaktides. IEEE käsitluses on nõuded AIS-i läbipaistvusele järgmised [1, lk 69–70; 158–160]:

Tabel 5. IEEE nõuded AIS-i läbipaistvusele [1, lk 69–70; 158–160].

SELGITUSKOHUSTUS
• AIS-ide kavandamisel tuleb püüelda süsteemi otsuste põhjuste selgitatavuse suunas. • STS-is peavad AIS-id proaktiivselt avaldama otsusest mõjutatud isikutele AIS-i otsust mõjutavad eeldused, riskid ja määramatused. • Selgituskohustus peab oleme kontekstitundlik – mida suurem on AIS-i abil langetatava otsuse mõju, seda olulisem on selgituskohustus; teatud riskide aktsepteerimine võib olla üldse lubamatu. • Vajalikud on standardid, nõuded ja meetodid, mis tagavad AIS-ide selgitatavuse (kohustuslikud logid, AIS-i järelevalvesüsteemid, läbipaistvuse tagamise meetodid). • Selgituskohustust aitab täita kogu AIS-i arendamise ja käitamise korrektne dokumenteerimine, AIS-ide ja komponentide eetikaauditid ja sertifitseerimine.
AVALIKUSTAMINE
• AIS-iga suhtlevad osapooled peavad teadma, et nad suhtlevad AIS-iga. • AIS-i käitajale ja otsusest mõjutatud isikule peab olema arusaadav, milliste andmete alusel AIS otsuseid langetab. • IEEE peab avalikustamise nõuet hetkel parimaks kättesaadavaks lahenduseks, kuna muud lahendused puuduvad või on osapooltele liiga keerulised mõista.
VASTUTUS
• AIS-i arendajad ja käitajad jäävad kandma süsteemi eetiliste ja sotsiaalsete nõuete mittetäitmise riski ja vastutust riskide realiseerumise mõjude eest.

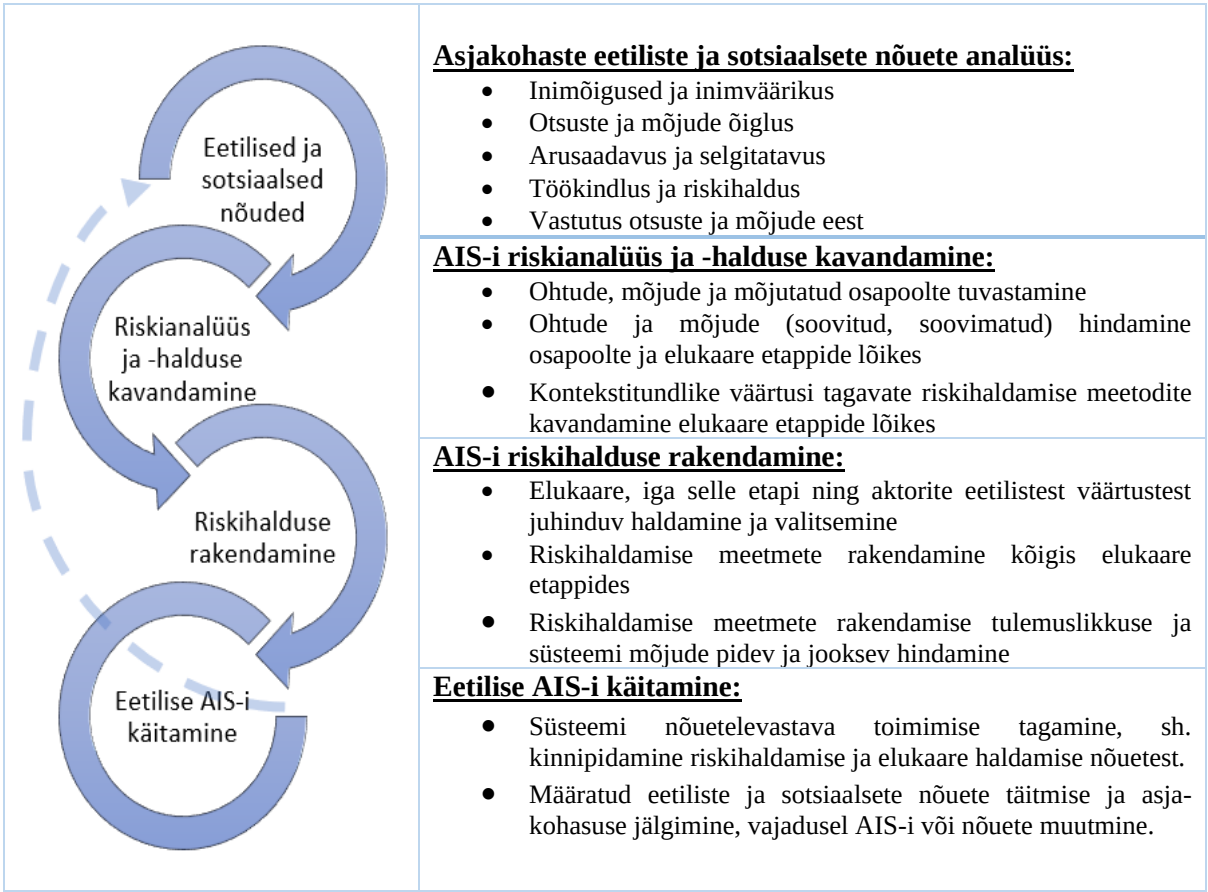
[1] Kirjanduses nimetatakse selliseid algoritme “musta kasti algoritmideks”.

Autori hinnangul on eelkäsitletud meetmetest eetilise AIS-i STS-ides käitamise seisukohast kõige olulisem kohustus tagada AIS-i otsustuste vastuvõtmise protsessi inimesele arusaadavus ja selgitatavus otsusest mõjutatud osapooltele ning vastava kohustuse täitmise meetmed (logid, dokumenteerimine jt).

Teiste käsitlustega võrreldes on IEEE eeliseks suurem konkreetsus nõuete ja nende tagamise vahendite osas. Kui teised käsitlused sõnastasid ennekõike üldisi eesmärke, siis IEEE keskendub konkreetsetele sammudele eetilise AIS tagamise teekonnal.

4.5 Eetilise AIS-i riskihalduse kontseptsioon

Eetilise AIS-i saavutamiseks tuleb AIS-i eetiliste ja sotsiaalsete põhinõudeid (vt Joonis 6, lk 56) ja konkreetsest käitamise kontekstist tulenevaid erinõudeid täita kõigis AIS-i elukaare etappides. Eetilise AIS-i elukaar ning eetiliste ja sotsiaalsete riskide haldamise süsteemi kontseptsioon on autori poolt esitatud alljärgnevalt:



Joonis 12. Eetilise AIS-i riskihalduse kontseptsioon.

Selliselt vaadates on AIS eetilisi ja sotsiaalseid nõudeid võimalik lihtsustada ühtedeks AIS-i kavandamise sisenditeks, kusjuures eetilised ja sotsiaalsed nõuded mõjutavad nii süsteemi funktsionaalseid kui ka mittefunktsionaalseid nõudeid. Esitatud eetilise AIS-i riskihalduse kontseptsiooni peamiseks eripäraks on eraldiseisev eetiliste ja sotsiaalsete nõuete analüüs, nende nõuete täitmist toetava riskihaldamise protsessi juurutamine ja selle järgimine läbi süsteemi elukaare. Vastavate sammude läbimine ei ole ühekordne tegevus. Eetilise AIS-i käitamisel tuleb regulaarselt kontrollida süsteemi arendamisel aluseks olnud eetiliste ja sotsiaalsete eesmärkide ja süsteemile seatud nõuete jätkuvat asjakohasust ning mõjude vastavust seatud või muutunud nõuetele.

Eetilise AIS-i elukaare haldamise eesmärgipäraseks toimimiseks on vajalik, et AIS-i arendamises, käitamises ja elukaare haldamises osalevad analüütikud ja muud spetsialistid omavad eetika valdkonna baasteadmisi ning oskavad neid rakendada konkreetse AIS-i kasutusvaldkonnas tekkivate eetiliste ja sotsiaalsete nõuete tuvastamiseks ning nende täitmise meetmete kavandamiseks ja rakendamiseks. Kirjeldatud AIS-i sotsiaalsete ja eetiliste riskide haldamise süsteemi võib rakendada iseseisvalt, kuid sageli võib olla otstarbekas selle lülitamine organisatsiooni või infosüsteemi üldisesse riskide haldamise protsessi.

5 Riskide haldamise metoodikad ja vahendid

Käesolevas peatükis süstematiseerib ja kirjeldab autor eetilise AIS-i elukaarel ilmnevate eetiliste ja sotsiaalsete riskide haldamise metoodikaid ja vahendeid.

5.1 Riskide haldamise metoodikate ja vahendite liigid

AIS eetiliste ja sotsiaalsete riskide haldamise metoodikaid ja vahendeid arendavad mitmed tehnoloogiavaldkonna osapooled. Oma algatusi omavad nii tehnoloogiaettevõtted (nt IBM: AI Fairness 360 [106], eetilise halduse arhitektuur [107], andmekaeve juhend CRISP-DM [108]; Google: Explainable AI lahendus [109]), rahvusvahelised organisatsioonid (nt OECD [14], EL [36], IEEE P7000™ standardiseeria), teadusringkonnad (algoritmi otsustusprotsessi selgitav LIME ja SP LIME [110], sisendite mõju kvantifitseerimise vahend QII [111], kallutatuse tuvastamise vahend FairML [56]) kui ka laiem tehnoloogiakogukond (ModelOps [112], MLOps [113][114]). Mitmed neist näidetest on sündinud teadlaste ja praktikute koostöös. See kinnitab, et AIS-i eetiliste ja sotsiaalsete riskide haldamine on oluline ja elujõuline uurimis- ning rakendusvaldkond, kus toimub nii aktiivne teoreetiline uurimine kui ka praktiliste meetodite ja vahendite arendamine.

AIS eetiliste ja sotsiaalsete riskide haldamise metoodikad ja vahendid jagunevad üldisteks metoodikateks ning tarkvaraliselt realiseeritud vahenditeks. Viimased omakorda jagunevad universaalseteks vahenditeks, mille eesmärk on samaaegselt hallata mitmeid AIS-iga elukaarel ilmevaid riske, ning spetsiifilisteks vahenditeks, mis pakuvad lahendust konkreetsetele probleemidele.

5.2 Riskide haldamise metoodikad

Eelneva liigituse alusel on autori hinnangul tähelepanuväärsemad järgmised AIS-i eetiliste ja sotsiaalsete riskide haldamise üldised metoodikad:

- *IEEE P7000™ standardiseeria*. Standardiseeria koosneb 14-st erinevast AIS seonduvast standardiseerimisalgatusest, millest 2020. aastal avaldatakse (i) AIS

eetiliste riskide haldamise (IEEE Standard P 7000™[1]), (ii) AIS läbipaistvuse (IEEE Standard P7001™[2]) ning (iii) algoritmi kallutatus(e haldamise) standardid (IEEE Standard P7003™[3]) [93, lk 4]. Lisaks on oodata mitmete eetilise AIS-iga seonduvate spetsiifiliste standardite valmimist (nt isikuandmete kaitse[4], laste ja õpilaste andmete töötlemine[5], kasutajate mõjutamine[6], näotuvastustehnoloogiate kasutamine[7]) [94, lk 70]. Standardiseeria eesmärgiks on luua nõuded eetilisele AIS-ile, mis toetaks vastavate nõuete täitmist ja võimaldaks sertifitseerimist ning auditeerimist [94, lk 72].

- *AI HLEG usaldusväärse AIS-i hindamise kontrollnimekiri*. AI HLEG on välja töötanud „*mitteammendava usaldusväärse tehisintellekti kontrollnimekirja […] usaldusväärse tehisintellekti kasutuslikule kujule viimiseks*“ [36, lk 27], mille lõplikku versiooni on oodata 2020. aastal [79, lk 7]. Kontrollnimekiri on universaalne enesehindamise raamistik, mida saab rakendada kõigis AIS-ide rakendusvaldkondades. Vajadusel korral saab raamistiku alusel luua täiendavaid, spetsiifilise valdkonna või konkreetse rakenduse vajadusi arvestavaid raamistike [36, lk 28]. Kontrollnimekirja alusel saavad AIS-i arendajad ja käitajad hinnata AIS-i ja seonduvate äriprotsesside vastavust AI HLEG-i eetikasuuniste nõuetele. Kontrollnimekiri annab hea ülevaate organisatsioonisisestest osapooltest (kaasaarvatud juhtimisraamistiku loomise eest vastutav nõukogu ja selle elluviimise eest vastutav juhatus), kes mõjutavad ja peavad olema kaasatud usaldusväärse AIS-i arendamise ja käitamisse. Kontrollnimekirja struktuur järgib magistritöö alapeatükis 3.2.3 kirjeldatud usaldusväärse AIS-i põhinõuete süsteemi. Iga põhinõude osas on AI HLEG esitanud enesehindamise küsimused.

---

[1] IEEE Standard P7000™ *„Model Process for Addressing Ethical Concerns During System Design“*

[2] IEEE Standard P7001™ *„Transparency of Autonomous Systems“*

[3] IEEE Standard P7003™ *„Algorithmic Bias Considerations“*

[4] IEEE Standard P7002™ *„Data Privacy Process“*

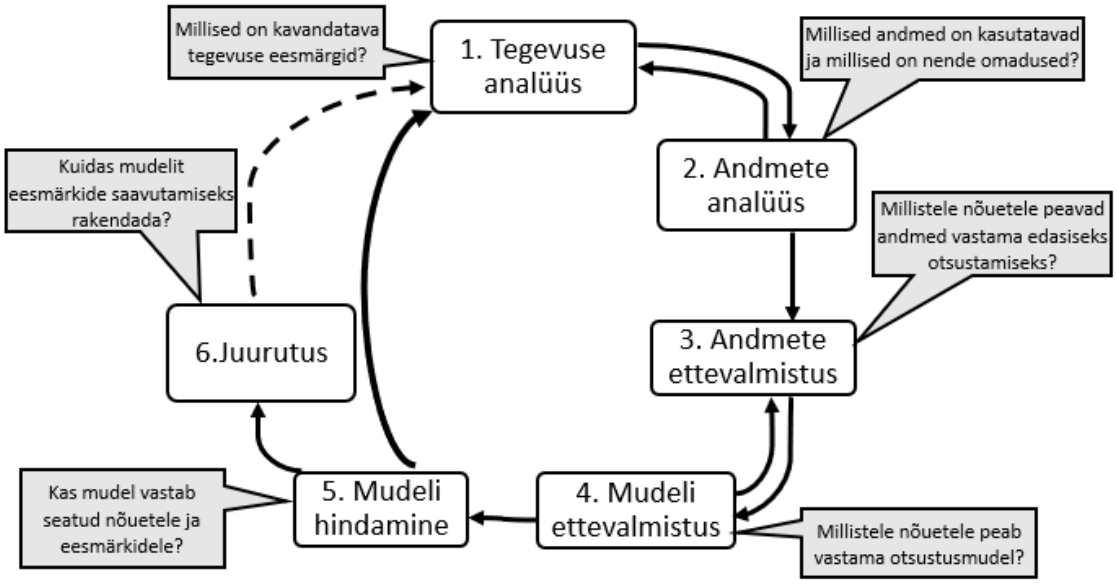
[5] IEEE Standard P7004™ *„Standard for Child and Student Data Governance“*

[6] IEEE Standard P7008™ *„Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems“*

[7] IEEE Standard P7013™ *„Inclusion and Application Standards for Automated Facial Analysis Technology“*

Küsimustik koosneb kokku 62 küsimusest koos alaküsimustega [36, lk 30–38]. Küsimustiku kaudu muutuvad eetikasuunistes sõnastatud abstraktsed põhinõuded kasutajatele praktiliselt mõistavateks. Ka saab kontrollnimekirja kasutada nii eetilise AIS-i arendamiseks ja käitamiseks vajalike organisatsiooni suutlikkuste hindamiseks kui ka nende arendamise planeerimiseks.

- *CRISP-DM (Cross-Industry Standard Process for Data Mining) andmekaeve metoodika ja protsessi mudel* [108], [115]. CRISP-DM on IBM-i loodud avatud standardil põhinev andmekaeve metoodika ja protsessi mudel, mis mõningate allikate kohaselt on andmekaeve valdkonnas enimrakendatav standardprotsess [116]. Ehkki 1995. aastal loodud CRISP-DM ei ole kavandatud AIS probleemistiku lahendamiseks, on kirjanduses metoodikat soovitatud nii AIS ja masinõppealgoritmide arendamise ja käitamise [117, lk 131] kui ka AIS-ide auditeerimise juhisena [118], [119]. AIS konteksti kohandatuna jagab CRISP-DM AIS-i elustsükli jätmistesse etappidesse: (1) tegevuse või ärivaldkonna analüüs, (2) andmete analüüs, (3) andmete ettevalmistus, (4) mudeli (algoritmi) ettevalmistus, (5) mudeli (algoritmi) hindamine ning (6) rakenduse juurutus. Ülevaate CRISP-DM protsessi mudelist ning AIS-i eetiliste ja sotsiaalsete riskide haldamise seisukohast asjakohastest kaalutlustest annab järgnev joonis:

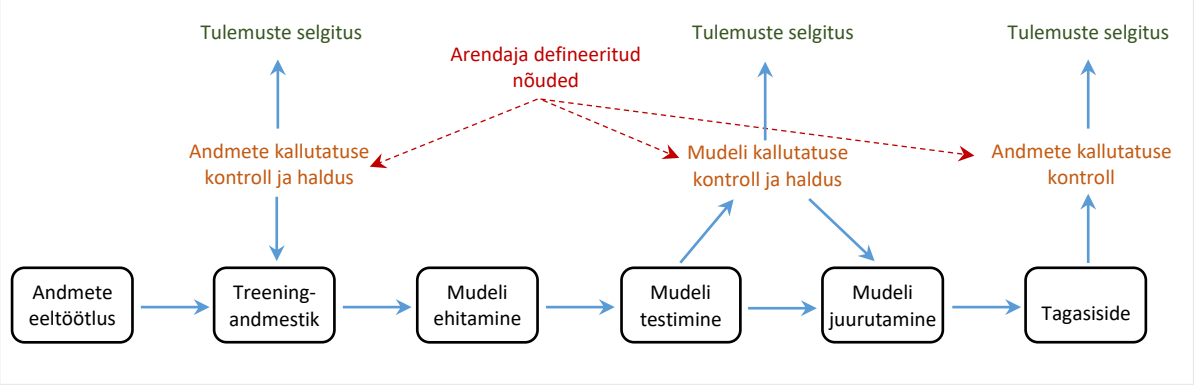


Joonis 13. CRISP-DM kontseptuaalne ülevaade [43, lk 9].

CRISP-DM protsessi läbimist tuleb jätkata ka pärast rakenduse juurutamist, et tagada juurutatud rakenduse vastavus tegevuse eesmärkidele.

Tarkvaraliselt realiseeritud AIS-i elukaare halduse vahendid ja nende aluseks olevad eetilise AIS tagamise meetodid on tänaseks toetatud kõigi suuremate masinõppelahenduste pakkujate (Google [120], IBM [121], Microsoft [122], Amazon [15]) teenustes. Spetsiifilisematest tarkvaraliselt realiseeritud vahenditest on autori hinnangul tähelapanuväärsemad:

- *AI Fairness 360 (AIF360).* AIF360 on IBM arendatav üheksast algoritmist koosnev metoodika ja rakendus, mis aitab tuvastada ja kõrvaldada AIS-is kasutatavate andmestike ja mudelite kallutatust ning parandada algoritmide selgitatavust [107], [123]. Vastava metoodika tarkvaralised realisatsioonid [124], [125] võimaldavad AIS-i eetiliste ja sotsiaalsete riskide haldamist tööstuslikuks kasutamiseks sobival tasemel. AIF360 hõlmab terviklikuks lahenduseks kallutatuse kõrvaldamise peamised meetodid – alates õiglasest andmete eeltöötlemisest („*fair pre-processing*“), õiglasest algoritmide treenimisest („*fair in-processing*“) kuni õiglasest tulemuste hindamiseni („*fair post-processing*“) [107]. Tervikliku lähenemise tõttu võimaldab see vahend AIS-i kallutatuse avastamist ja meetmete rakendamist sobivaimas AIS-i elukaare etapis.



Joonis 14. AIF360 kontseptuaalne ülevaade [123].

- *ModelOps.* ModelOps on AIS-i arendajate kogukonna ja IBM toetatav algatus, mille eesmärk on rakendada DevOps lähenemist eetiliste ja töökindlate AIS rakenduste arendamiseks ja käitamiseks. ModelOps eesmärk on saavutada AIS-ide arendamise automatiseeritus, nõuetelevastavus, usaldusväärsus, jälgitavus ja kvaliteedikontroll [112, lk 1]. ModelOps on spetsiifiliselt masinõppe

probleemide lahendamiseks arendatud võtete ja vahendite kogum. Seetõttu võimaldab ModelOps mitmete algoritmite samaaegse ja ülikiire treenimise ja juurutamisega seonduva töövoo ja probleemide haldamist – ModelOps autorid viitavad kasutajate poolt uute algoritmide käikulaskmisele keskmiselt 1,2-226 korda päevas [112, lk 3]. Lisaks mudelite ja andmestiku kiire muutumise riskidele võimaldab ModelOps toetada ka muude AIS-i riskide haldamist.

5.3 Riskide haldamise vahendid

Konkreetsete eetiliste ja sotsiaalsete riskide haldamise vahendid abistavad AIS-i arendajat ja käitajat peamiselt AIS-i otsuste kallutatuse kontrollimisel ja kõrvaldamisel ning algoritmi otsustuste vastuvõtmise protsessi selgitatavuse tagamisel.

AIS-i kallutatuse haldamise vahendid jagunevad kaheks suuremaks kategooriaks. Esimene kategooria vahendeid seab eesmärgiks kallutuseni viivate tundlikke tunnuste (nt sugu, rahvus) eemaldamise andmestikest enne nende kasutamist algoritmi treenimiseks või otsuste vastuvõtmiseks. Neid tööriistu eelistavate osapoolte hinnangul on ainult nii võimalik välistada tundlike tunnuste mõju algoritmi otsusele. Teised tööriistad eelistavad algoritmide treenimist ja kasutamist kallutatuse ja tundlike tunnuste osas eeltöötlemata andmetega ning kallutatuse kontrollimist ja eemaldamist algoritmi otsuste tasemel. Neid tööriistu eelistavate osapoolte hinnangul on tundlike tunnuste andmestikest kõrvaldamine ülimalt keeruline või lausa võimatu, seda eriti kaudsete tunnuste osas (nt postiindeks võib viidata vastavas piirkonnas elava isiku rahvusele)[126, lk 53]. Teiseks vähendab andmestike eelnev puhastamine tundlikest tunnustest selliste andmestike alusel treenitud algoritmide täpsust [107]. Autori hinnangul on oluline tagada mitte andmestike, vaid just AIS-i otsuste nõuetelevastavus. Seetõttu tuleb põhitähelepanu pöörata just nende kallutatuse hindamisele ja kõrvaldamisele. Kallutatuse tuvastamist toetavatest vahendites on enim käsitlemist leidnud FairML [118, lk 2]. FairML on auditeerimistööriist, mis mõõdab ja muudab inimesele arusaadavaks, kas ja mil määral mõjutavad andmestiku tundlikud tunnused konkreetse algoritmi tulemusi ([56, lk 93–94])[1]. Kirjanduses on osundatud ka

[1] Kallutatuse vältimist ja avastamist tutvustavates teadustöödes sisaldub sageli lisaks tehnilisele tutvustusele haarav arutelu algoritmi kallutatuse ja õigluse eetilise ja filosoofilise mõõtme üle. Teemast

järgmistele kallutatuse tuvastamise vahenditele Fairness Measures [127], FairTest [128], [129] ning Aequitas [130], [131]. Kallutatuse eemaldamist võimaldavad ka Themis-ML [132], [133] ning Fairness Comparison komplekt [134], [135].

Teine liik vahendeid toetab algoritmi otsuste selgitatavuse tagamist – need lahendused mõõdavad algoritmi sisendite muutuse mõju väljundile ja visualiseerivad inimesele arusaadavalt algoritmi poolt kasutavad sisendid ning sisendite mõju otsustele. Kirjanduses on käsitlemist leidnud LIME („*Local Interpretable Model-Agnostic Explanations*"), mis on mitmest algoritmist koosnev masinõppe klassifikaatorite selgitamise raamistik [110]. Muuhulgas võimaldab LIME superpiksli meetodil visualiseerida ka algoritmi poolt fotode klassifitseerimisel (pildituvastus) aluseks võetud pildi osasid [136, lk 5–9] ja korrigeerida vastavaid algoritmi vigu enne algoritmi kasutusse võtmist [110, lk 5, 9]. Masinõppe algoritmide selgitatavust käsitlev kirjandus näitab ilmekalt, kui põimunud on valdavalt masina loodud algoritmide otsuste inimesele arusaadavuse probleem algoritmi tehnilise töökindluse probleemiga. Masinõppealgoritmi otsuste põhjenduste mõistmise ja auditeerimise vahenditeks on ka Gestalt [137], Modeltracker [138], QII („*Quantitative Input Influence")* [111] ja TREPAN [139]. Seda tüüpi vahendite arendajad leiavad, et algoritmi otsuste selgitatavus on algoritmi otsuste usaldamise eelduseks [110, lk 10], [111, lk 1].

Autori hinnangul on AIS-i elukaare haldamise metoodikad varajases arengujärgus, madala küpsustasemega ning välja ei ole kujunenud üldtunnustatud lahendusi [112, lk 1]. Konkreetsete AIS-i eetilisusega seonduvate standarditeni jõudmisest on oodata 2020. aasta lõpus, mil IEEE avaldab P7000™ standardiseeria esimesed standardid. Kuid ka standardid esitavad üksnes nõuded ja jätavad kasutajatele võimaluse ja vastutuse valida konkreetsesse olukorda sobivad metoodikad ja vahendid.

Olemasolevate meetodite ja nende tarkvaraliste rakenduste abil on siiski võimalik oluliselt parandada käideldavate AIS-ide vastavust eetilise AIS-i nõuetele ning hallata selliste süsteemide käitamisega seonduvaid eetilisi ja sotsiaalseid riske.

---

sügavamalt huvitatutel soovitab autor tutvuda Adebayo [56] ja Barocase [51] käsitlustega ning Kamishima ettekannetega õiglus-teadlikkust andmekaevest [9]. Kamishima annab ülevaatliku käsitluse nii erinevatest relevantsetest õigluse vormidest (võrdsed võimalused vs võrdsed tulemused) kui ka nende tehnilistest väljendustest kitsamas masinõppe valdkonnas.

Kokkuvõte

Magistritöös uuris autor eetilisele AIS-ile esitatud eetilisi ja sotsiaalseid nõudeid, nende nõuete sisu ning nende täitmise viise.

Eetilisele AIS-ile esitatud nõudeid tuleb selliste süsteemide kavandamisel ja käitamisel arvesse võtta juba täna. Käesoleval hetkel ei ole veel kehtestatud spetsiifiliselt AIS-idele rakenduvaid kohustuslikke eetilisi ja sotsiaalseid nõudeid või standardeid, kuid need on peatselt tulemas. Ka on raamdokumentidest ning standardimisalgatustest näha ühiskonna ja valdkonna osapoolte ootus, et AIS-i arendajad ja käitajad peavad juba täna tagama eetilise AIS-i põhinõuete täitmise. Ka on mõned sellised põhinõuded juba sätestatud õigusaktides kas üldiste (nt üldine kahju tekitamise või diskrimineerimise keeld) või valdkonnapõhiste nõuetena (nt isikuandmete kaitse reeglid). Seega on eetilise AIS-i nõuded täna valdavalt soovituslikud, kuid osaliselt juba kohustuslikud.

Vaba tahte puudumise tõttu ei saa (täna veel) eetilise AIS-i nõuete täitmist nõuda AIS-ide endi poolt. AIS-idel on teatav autonoomia ja intelligents. Samas ei ole see piisav AIS-idele vaba tahte omistamiseks. Eetiliste ja sotsiaalsete hinnangute andmise eelduseks on aga just vaba tahte alusel toimuv põhjustamine. Siiski on AIS-ide autonoomia ja intelligentsus piisav selleks, et põhjustada mitmeid käesoleva magistritöö teises peatükis käsitletud spetsiifilisi probleeme ja riske. Autonoomsus annab AIS-idele võimekuse muuta süsteemi töö käigus ise oma toimimist. Teadmiste omandamisega (nt masinõppealgoritmi õpetamisega) kaasneb algoritmi otsuste vastuvõtmise protsessi muutus. Ka valivad AIS-id otsuse tegemiseks kasutatavad andmed ning neile omistatavad osakaalud autonoomselt või inimese piiratud osalusel. Sõltuvalt kasutatud tehnoloogiast, võib arusaamine algoritmi otsustamise aluseks olevatest reeglitest, andmetest ja nende muutumisest olla inimesele ülikeeruline või praktiliselt võimatu. See autonoomsus, intelligentsus ja inimesele arusaadavuse piiratus võib kaasa tuua AIS-i kõrvalekaldumise inimese poolt seatud eetilistest ja sotsiaalsetest nõuetest. Samad põhjused raskendavad ka AIS-ide töökindluse tagamist.

AIS-ile esitatud eetilised ja sotsiaalsed nõuded on sisuliselt nõuded AIS-i arendaja ja käitaja tegevusele. AIS-i arendaja ja käitaja peavad tagama, et AIS-i mõjud vastavad ühiskonnas ja konkreetses AIS-i käitamise kontekstis asjakohastele eetilistele ja sotsiaalsetele nõuetele. Eetilise AIS-i nõuete täitmine on ennekõike süsteemi käitaja

ülesanne, kuna AIS põhjustab inimestele või ühiskonnale vahetut mõju just käitaja tegevuse raames. Arendaja peab järgima eetilise AIS-i nõudeid, et anda käitaja kasutusse nõuetekohane süsteem.

Autori käsitluses on eetiline selline AIS, (1) mille käitaja on tuvastanud konkreetses käitamise kontekstis asjakohased eetilised ja sotsiaalsed nõuded, (2) määranud nende alusel nõuded süsteemi toimimisele ja (3) mis täidab neid nõuetekohaselt. Eetilise AIS-i käsitlused ei erista väga täpselt eetilisi ja sotsiaalseid nõudeid, vaid käsitlevad neid ühtse eetilise või usaldusväärse AIS-i nõuete paketina. Vastavad nõuded erinevad käsitluste lõikes, kuid on teatav ühisosa, mille alusel on autor välja toonud järgmised põhinõuded eetilisele AIS-ile: (1) inimõiguste ja inimväärikuse austamine, (2) AIS-i otsuste ja mõjude õiglus, (3) AIS-i otsuste arusaadavus ja selgitatavus, (4) AIS-i töökindlus ja riskihaldus ning (5) vastutus AIS-i otsuste ja mõjude eest (vt Joonis 6, lk 56). Eetilise AIS-i definitsiooni osaks on ka süsteemi töökindlus ja riskide haldamine, kuna nende ebapiisavus on AIS-ide mitmete eetiliste ja sotsiaalsete probleemide põhjuseks. Magistritöö kolmandas peatükis avatud eetilise AIS-i nõudeid tuleb konkretiseerida konkreetse AIS-i analüüsi ja kavandamise käigus ning valdavas osas on neid võimalik läbi tehniliste või muude meetmete praktiliselt rakendada.

AIS-iga seonduvatel eetilistel ja sotsiaalsetel riskidel ja mõjudel on väga erinevad põhjused. Need riskid ja mõjud võivad avalduda AIS-i elukaare erinevates etappides. Autori hinnangul leidis magistritöös kinnitamist hüpotees, et vastavaid riske on võimalik hallata üksnes konkreetse AIS-i kogu elukaart hõlmava riskihalduse abil. Autori kirjeldatud eetilise AIS-i riskihalduse kontseptsioon koosneb järgmisest etappidest: (1) asjakohaste eetiliste ja sotsiaalsete nõuete analüüs, (2) AIS-i riskianalüüs ja -halduse kavandamine, (3) AIS-i riskihalduse rakendamine ning (4) eetilise AIS-i käitamine (sh nõuete täitmise jälgimine). Sellist riskihaldust võib rakendada iseseisvalt, kuid sageli võib olla otstarbekas selle lülitamine organisatsiooni või infosüsteemi üldisesse riskide haldamise süsteemi. Täpsemalt on kirjanduses soovitatud riskihalduse meetmeid kirjeldatud magistritöö neljandas peatükis.

AIS-i elukaarel ilmnevate eetiliste ja sotsiaalsete riskide haldamise metoodikad ja vahendid on täna veel madala küpsustasemega [112, lk 1], mistõttu on eetilise AIS-i tagamine jätkuvalt keeruline väljakutse [74, lk 12]. Metoodikatest on hetkel vahetuma, ent ebapraktilise mõjuga EL-i eetikasuunised. Samas on 2020. aasta jooksul oodata

IEEE P7000™ standardiseeria eetilise AIS-iga seonduvaid standardeid, mis peaksid andma konkreetsemaid juhiseid. Eetilise AIS-i haldust toetavad vahendid on kättesaadavad nii integreerituna suuremate masinõppelahenduste pakkujate (Google, IBM, Microsoft, Amazon) teenustesse kui ka eraldiseisvalt (nt IBM AIF360). Spetsiifiliste riskide haldamise vahendid toetavad ennekõike AIS-i otsuste kallutatuse kõrvaldamist (FairML, Fairness Measures jt) või otsustusprotsessi selgitatavuse tagamist (LIME, Gestalt jt). Täpsem ülevaade on esitatud magistritöö viiendas peatükis.

Autor peab oluliseks eetiliste ja sotsiaalsete nõuete järgimist AIS-ide kavandamisel ja käitamisel. Eetilise AIS-i kavandamise lähtepunktiks on aga suutlikkus analüüsida AIS-iga kaasnevaid eetilisi ja sotsiaalseid mõjusid ja riske, mõista nende allikaid ja neid põhjustavaid ohte ning pakkuda välja mittesoovitud mõjude haldamiseks sobivad meetmeid. Seetõttu soovitab autor kaaluda infotehnoloogia õppekavade laiendamist, õpetamaks infosüsteemide arendajaid ja käitajaid märkama ja ennetama ka süsteemidega seonduvaid eetilisi ja sotsiaalseid ohte [73].

Inseneriteadused on alati lähtunud insenerieetikast, kuid sotsiaalsema mõõtmega eetikaküsimusi (kallutatus, õiglus, vabaduste piiramine) ei ole traditsiooniliselt loetud inseneriteaduste uurimisesemesse. Tehnoloogia ühiskondlike mõjude suurenemise tõttu peaksid aga muutuma ka ootused inseneridele – enam ei ole aktsepteeritav sotsiaalseid nõudeid eirava tehnoloogia kasutajateni viimine ning negatiivsete mõjude ühiskonna lahendada jätmine [106, lk 1].

Ka ei ole autori hinnangul põhjendatud jätta tehnoloogiliste riskidega seonduva diskussiooni juhtimine ilma tehnoloogia-alase ettevalmistuseta eetikutele, filosoofidele või ühiskonnateadlastele: reaalteaduste ja sellele tugineva tehnoloogia mittemõistmise tõttu võib diskussioon keskenduda ebapraktiliste (nt tehisintellektile õigussubjektsuse omistamine) probleemide käsitlemisele, samal ajal kui praktilist väärtust omavad, igapäevased probleemid (AIS otsuste selgitatavus) jäävad adekvaatselt käsitlemata.

Edasise uurimise jaoks peab autor huvitavaks AIS-i otsuste inimesele selgitatavuse ning otsuste kallutatuse kõrvaldamise praktiliste võtete käsitlemist. Nende probleemide lahendamine aitaks vältida suurt osa magistritöös käsitletud AIS-i eetilistest ja sotsiaalsetest probleemidest ja riskidest. Nende lahendamise käigus tõusetuvad koos tehniliste küsimusetega ka mitmed keerulised eetilised ja sotsiaalsed probleemid.

Kasutatud kirjandus

[1] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, „Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2“, 2017.

[2] B. Friedman ja H. Nissenbaum, „Bias in Computer Systems“, *ACM Transactions on Information Systems*, kd 14, nr 3, lk 330–347, 1996.

[3] European Commission: High-Level Expert Group, „High-Level Expert Group on Artificial Intelligence a Definition of Ai: Main Capabilities and Disciplines“, lk 7, 2019.

[4] S. J. Russell ja P. Norvig, *Artificial Intelligence A Modern Approach*, 3rd Ed. Upper Saddle River: Pearson Education, Inc., 2003.

[5] OECD, „Inter-governmental Organisations - OECD“, 2020. [Online]. Available at: https://www.oecd.org/gov/regulatory-policy/irc4.htm. [Vaadatud: 14-märts-2020].

[6] T. Herrmann, M. Hoffmann, G. Kunau, ja K. U. Loser, „A Modelling Method for the Development of Groupware Applications as Socio-technical Systems“, *Behaviour and Information Technology*, kd 23, nr 2, lk 119–135, 2004.

[7] G. Baxter ja I. Sommerville, „Socio-technical systems: From design methods to systems engineering“, *Interacting with Computers*, kd 23, nr 1, lk 4–17, jaan 2011.

[8] E. Tõugu, „On the Border Between Functional Programming and Program Synthesis“, *Proceedings of the Estonian Academy of Sciences*, nr 4, lk 119–129, 1998.

[9] T. Kamishima, „Fairness-Aware Data Mining“, 2019. [Online]. Available at: http://www.kamishima.net/. [Vaadatud: 16-veebr-2020].

[10] A. Jobin, M. Ienca, ja E. Vayena, „The Global Landscape of AI Ethics Guidelines“, *Nature Machine Intelligence*, kd 1, nr 9, lk 389–399, 2019.

[11] R. Csernatoni, „An Ambitious Agenda or Big Words ? Developing a European Approach to AI“, *EGMONT – Royal Institute for International Relations*, kd 117, 2019.

[12] The State Council of the People’s Republic of China, „Notice of the State Council Issuing the New Generation of Artificial Intelligence Development Plan“, 2017.

[13] Z. Sirui, „China Bets Big on AI: Summary of Central Government Policies - EqualOcean“, *EqualOcean*, 2019. [Online]. Available at: https://equalocean.com/ai/20190726-china-betting-big-on-ai-summary-of-central-government-policies. [Vaadatud: 02-märts-2020].

[14] OECD, „Scoping the OECD AI Principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD (AIGO)“, *OECD Digital Economy Papers*, kd 291, lk 28, 2019.

[15] G. Nott, „AWS is ethical about AI but ‘we just don’t talk about it’ say APAC execs | Computerworld“, *Computerworld*, 2019.

[16] International Organization for Standardization, „Artificial Intelligence (ISO/IEC JTC 1/SC 42)“, *ISO*, 2017. [Online]. Available at: https://www.iso.org/committee/6794475.html. [Vaadatud: 26-märts-2020].

[17] G. Tsafnat, P. Glasziou, M. K. Choong, A. Dunn, F. Galgani, ja E. Coiera, „Systematic Review Automation Technologies“, *Systematic Reviews*, kd 3, nr 74, 2014.

[18] M. Petticrew ja H. Roberts, *Systematic Reviews in the Social Sciences: A Practical Guide*. Oxford: Blackwell Publishing, 2006.

[19] V. Kalmus, A. Masso, ja M. Linno, „Kvalitatiivne sisuanalüüs | Sotsiaalse Analüüsi Meetodite ja Metodoloogia õpibaas“, 2015. [Online]. Available at: http://samm.ut.ee/kvalitatiivne-sisuanalyys. [Vaadatud: 28-märts-2020].

[20] International Organization for Standardization, „ISO 31000:2018(en), Risk management - Guidelines“, 2018. [Online]. Available at: https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-2:v1:en. [Vaadatud: 22-apr-2020].

[21] P. Lorents, *Süsteemide maailm*. Tartu: Tartu Ülikooli Kirjastus, 2006.

[22] I. Meos, „Filosoofia põhiprobleemid“, *Tallinn, Koolibri*, lk 360, 1998.

[23] M. Koit ja T. Roosmaa, *Tehisintellekt*. Tartu: Tartu Ülikooli Kirjastus, 2011.

[24] D. Rimkute, „The Impact and Regulatory Challenges of Artificial Intelligence“, *Citizen-Centred Digitalisation. Towards a Co-evolution of Technology and Society*, E. Liss, Toim Brussels: European Liberal Forum, 2018, lk 6–17.

[25] R. Palm ja R. Prank, *Sissejuhatus matemaatilisse loogikasse*. Tartu: Tartu Ülikooli Kirjastus, 2004.

[26] F. Kraemer, K. van Overveld, ja M. Peterson, „Is There an Ethics of Algorithms?“, *Ethics and Information Technology*, kd 13, nr 3, lk 251–260, sept 2011.

[27] P. Lorents, *Hulgad, valemid, algoritmid*. Tallinn: EBS Print, 2002.

[28] I. Kull ja M. Tombak, „Algoritmid ja lahenduvad hulgad ning nende rakendusi“, *Matemaatika ja kaasaeg. Abimaterjale matemaatika õpetajatele ja õppijatele*, kd XII, lk 44–63, 1967.

[29] E. Tõugu, *Arvutid, küberruum ja tehismõistus. Noppeid arvutite imepärasest eduloost*. Tallinn: Tallinna Ülikooli Kirjastus, 2018.

[30] J. G. Granström, *Treatise on Intuitionistic Type Theory*. Springer Science & Business Media, 2011.

[31] Y. Gurevich, „What is an Algorithm?“, *SOFSEM 2012: Theory and Practice of Computer Science.*, 2012, kd 7147 LNCS, lk 31–42.

[32] N. S. Yanofsky, „Towards a definition of an algorithm“, *Journal of Logic and Computation*, kd 21, nr 2, lk 253–286, veebr 2011.

[33] P. Lorents, „Konstruktsioonid ja algoritmid“, 2011. [Online]. Available at: https://lorents.ee/is_old/12 _KONSTRUKTSIOONID_JA_ALGORITMID.pdf. [Vaadatud: 20-märts-2020].

[34] D. E. Knuth, *The Art of Computer Programming. Volume 1. Fundamental Algorithms. Third Edition*. 1997.

[35] S. Jordan, R. Day, ja L. M. Ingram, „Glossary for Discussion of Ethics of Autonomous and Intelligent Systems , Version 1 A Glossary for Discussion of Ethics of Autonomous and Intelligent Systems , Version 1 Prepared for The IEEE Global Initiative for Ethically Aligned Design Glossary Committ“, 2017.

[36] Euroopa Komisjoni Sõltumatu Kõrgetasemeline Eksperdirühm, „Eetikasuunised usaldusväärse tehisintellekti loomiseks“, Brüssel, 2019.

[37] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[38] M. Kubat, *An Introduction to Machine Learning*. Springer International Publishing, 2017.

[39] John Robert Anderson, J. G. Carbonell, T. M. Mitchell, R. S. Michalski, ja S. Amare, *Machine Learning: An Artificial Intelligence Approach*. Palo Alto: Tioga Publishings Co., 1983.

[40] G. Marcus ja E. Davis, *Rebooting AI: Building artificial intelligence we can trust*. New York: Pantheon Books, 2019.

[41] „Machine Learning Definition | DeepAI“. [Online]. Available at: https://deepai.org/machine-learning-glossary-and-terms/machine-learning. [Vaadatud: 24-veebr-2020].

[42] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, ja L. Floridi, „The Ethics of Algorithms: Mapping the Debate“, *Big Data & Society*, kd 3, nr 2, dets 2016.

[43] L. Ilany, „How to build and deploy machine learning projects“, *Haifa Verification Conference*, 2017.

[44] T. Pungas, „Masinõpe: mittetehniline ülevaade“, 2017. [Online]. Available at:

https://pungas.ee/masinope-mittetehniline-ulevaade/. [Vaadatud: 15-märts-2020].

[45] B. Goodman ja S. Flaxman, „European Union Regulations on Algorithmic Decision Making and a ‘Right to Explanation’“, *AI Magazine*, kd 38, nr 3, lk 50–57, 2017.

[46] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, ja N. Elhadad, „Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission“, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, lk 1721–1730.

[47] M. Grijalvo, „Implementation of Quality Systems in Sociotechnical Systems“, *Quality Management and Practices*, InTech, 2012.

[48] E. Trist, *The Evolution of Socio-technical Systems: a Conceptual Framework and an Action Research Program*. Toronto: Ontario Quality of Working Life Centre, 1981.

[49] R. P. Bostrom ja J. S. Heinen, „MIS Problems and Failures: A Socio-Technical Perspective. Part I: The Causes“, *MIS Quarterly*, kd 1, nr 3, lk 17, 1977.

[50] R. Oosthuizen ja L. Pretorius, „Assessing the Impact of New Technology on Complex Sociotechnical Systems“, *South African Journal of Industrial Engineering*, kd 27, nr 2, lk 15–29, 2016.

[51] S. Barocas, M. Hardt, ja A. Narayanan, „Fairness in Machine Learning: Limitations and Opportunities“, 2019. [Online]. Available at: https://fairmlbook.org/pdf/fairmlbook.pdf. [Vaadatud: 09-veebr-2020].

[52] D. K. Citron, „Technological Due Process“, *Washington University Law Review*, kd 85, lk 1249, 2008.

[53] C. O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, 2016.

[54] K. Martin, „Ethical Implications and Accountability of Algorithms“, *Journal of Business Ethics*, kd 160, nr 4, lk 835–850, 2019.

[55] D. Krishna, „Managing Risks Raised by Algorithmic Applications“, *The Wall Street Journal*, 2018.

[56] J. A. Adebayo, „FairML: ToolBox for Diagnosing Bias in Predictive Modeling“, Massachusetts Institute of Technology, 2016.

[57] A. W. Flores, K. Bechtel, ja C. T. Lowenkamp, „False positives, false negatives, and false analyses: A rejoinder to ‘machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks’“, *Federal Probation*, kd 80, nr 2, lk 38–46, mai 2016.

[58] J. E. Johndrow ja K. Lum, „An Algorithm for Removing Sensitive Information: Application to Race-independent Recidivism Prediction“, *The Annals of Applied Statistics*, kd 13, nr 1, lk 189–220, märts 2019.

[59] A. Tutt, „An FDA for Algorithms“, *SSRN Electronic Journal*, kd 69, nr 1, lk 83–123, 2016.

[60] B. D. Underwood, „Law and the Crystal Ball: Predicting Behavior with Statistical Inference and Individualized Judgment“, *The Yale Law Journal*, kd 88, nr 7, lk 1408, juuni 1979.

[61] M. Hildebrandt, „Who Needs Stories If You Can Get the Data? ISPs in the Era of Big Number Crunching“, *Philosophy and Technology*, kd 24, nr 4, lk 371–390, 2011.

[62] J. Burrell, „How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms“, *Big Data & Society*, kd 3, nr 1, lk 12, jaan 2016.

[63] C. Sandvig, K. Hamilton, K. Karahalios, ja C. Langbort, „When the Algorithm Itself is a Racist: Diagnosing Ethical Harm in the Basic Components of Software“, *International Journal of Communication*, kd 10, lk 4972–4990, 2016.

[64] T. Simonite, „The Best Algorithms Struggle to Recognize Black Face“, *Wired*, 2019. [Online]. Available at: https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/. [Vaadatud: 18-märts-2020].

[65] S. Lohr, „Facial Recognition Is Accurate, if You’re a White Guy“, *The New York*

Times, 2018. [Online]. Available at: https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html. [Vaadatud: 18-märts-2020].

[66] O. Schwartz, „In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation - IEEE Spectrum", *IEEE Spectrum*, 2019. [Online]. Available at: https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation. [Vaadatud: 24-veebr-2020].

[67] K. de Vries, „Identity, Profiling Algorithms and a World of Ambient Intelligence", *Ethics and Information Technology*, kd 12, nr 1, lk 71–85, 2010.

[68] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, ja R. Zemel, „Fairness Through Awareness", *ITCS 2012 - Innovations in Theoretical Computer Science Conference*, 2012, lk 214–226.

[69] T. Zarsky, „The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making", *Science Technology and Human Values*, kd 41, nr 1, lk 118–132, 2016.

[70] E. Bozdag, „Bias in Algorithmic Filtering and Personalization", *Ethics and Information Technology*, kd 15, nr 3, lk 209–227, sept 2013.

[71] A. Matthias, „The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata", *Ethics and Information Technology*, kd 6, nr 3, lk 175–183, 2004.

[72] Algorithm Watch, „AI Ethics Guidelines Global Inventory – AlgorithmWatch", *Algorithm Watch*, 2019. [Online]. Available at: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/. [Vaadatud: 09-märts-2020].

[73] University of Cambridge, „The Centre for the Study of Existential Risk", 2019. [Online]. Available at: https://www.cser.ac.uk/. [Vaadatud: 16-märts-2020].

[74] D. Peters, K. Vold, D. Robinson, R. A. Calvo, ja S. Member, „Responsible AI-

Two Frameworks for Ethical Design Practice“, *IEEE Transactions on Technology and Society,* kd 1, nr 1, 2020.

[75] P. Yunhe *et al.*, „China AI Development Report 2018“, Peking, 2018.

[76] Euroopa Komisjon, „Tehisintellekti käsitlev kooskõlastatud kava COM(2018)795/F1 - ET (annex)“, Brüssel, 2018.

[77] Euroopa Komisjon, „EU Member States sign up to cooperate on Artificial Intelligence | Shaping Europe’s digital future“, 2018. [Online]. Available at: https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence. [Vaadatud: 26-veebr-2020].

[78] Euroopa Liit ja Norra Kuningriik, *Koostöödeklaratsioon tehisintellekti valdkonnas*. Brüssel, 2018.

[79] European Commisson, „Communication: Building Trust in Human Centric Artificial Intelligence“, 2019.

[80] Euroopa Komisjon, *Tehisintellekt Euroopa huvides. Komisjoni teatis Euroopa Parlamendile, Euroopa Ülemkogule, Nõukogule, Euroopa Majandus- ja Sotsiaalkomiteele ning Regioonide Komiteele*. Brüssel: Euroopa Komisjon, 2018.

[81] Majandus- ja Kommunikatsiooniministeerium, „Eesti tehisintellekti kasutuselevõtu eksperdirühma aruanne“, 2019.

[82] Euroopa Ülemkogu, „Euroopa Ülemkogu kohtumine (28. juuni 2018) - Järeldused“, Brüssel, 2018.

[83] Privacy International, „Why and how GDPR applies to people globally“, *privacyinternational.com,* 2018. [Online]. Available at: https://privacyinternational.org/long-read/2207/why-and-how-gdpr-applies-companies-globally. [Vaadatud: 29-veebr-2020].

[84] A. Bradford, „When It Comes to Setting Standards for the World, Europe Is No Fading Power“, *Foreign Affairs,* 2020. [Online]. Available at: https://www.foreignaffairs.com/articles/europe/2020-02-03/when-it-comes-markets-europe-no-fading-power. [Vaadatud: 26-veebr-2020].

[85] Euroopa Parlament, *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)*. OJ L 119, 4.5.2016, 2016, lk 1–88.

[86] European Commission, „Algorithmic Awareness-Building | Shaping Europe's digital future". [Online]. Available at: https://ec.europa.eu/digital-single-market/en/algorithmic-awareness-building. [Vaadatud: 29-veebr-2020].

[87] N. A. Smuha, „The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence", *Computer Law Review International*, kd 20, nr 4, lk 97–106, 2019.

[88] E. Ries, *Startup Way. How Modern Companies Use Entrepreneurial Management to Transform Culture & Drive Long-Term Growth*. New York: Currency, 2017.

[89] K. Korjus, T. Pungas, R. Pärnpuu, ja A. Heinla, „The Data is the Specification: A Manifesto for Iteratively Solving Complex Problems", 2019. [Online]. Available at: https://dataisspec.github.io/. [Vaadatud: 02-märts-2020].

[90] OECD, „Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449", 2019.

[91] Council of Europe, *Recommendation of the Committee of Ministers to Member States on the Human Rights Impacts of Algorithmic Systems*. 2020.

[92] P. Penninckx, „Human Rights & Artificial Intelligence", *Conference Octopus 2019: Cooperation against Cybercrime, Strasbourg, 22 November 2019*, 2019.

[93] IEEE Standards Association, „The IEEE Ethics Certification for Autonomous and Intelligent Systems (ECAIS) Industry Connections Activity Initiation Document (ICAID)", 2018.

[94] J. C. Havens ja A. Hessami, „From Principles and Standards to Certification", *Computer*, kd 52, nr 4, lk 69–72, 2019.

[95] Open Community of Ethics in Autonomous and Intelligent Systems, „IEEE P7000™ Projects - OCEANIS", 2020. [Online]. Available at:

https://ethicsstandards.org/p7000/. [Vaadatud: 07-märts-2020].

[96] Riigikogu, *Toote nõuetele vastavuse seadus*. Tallinn: Riigi Teataja, 2010.

[97] Riigikogu, *Võlaõigusseadus*. Tallinn: Riigi Teataja, 2001.

[98] Riigikogu, *Võrdse kohtlemise seadus*. Tallinn: Riigi Teataja, 2008.

[99] N. Vigdor, „Apple Card Investigated After Gender Discrimination Complaints - The New York Times“, *The New York Times*, 2019. [Online]. Available at: https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html. [Vaadatud: 16-veebr-2020].

[100] S. Narang, C. Raffel, K. Lee, A. Roberts, N. Fiedel, ja K. Malkan, „WT5?! Training Text-to-Text Models to Explain their Predictions“, *arXiv.org*, 29-apr-2020. [Online]. Available at: http://arxiv.org/abs/2004.14546. [Vaadatud: 03-mai-2020].

[101] B. Davis, „What’s a Global Recession? - Real Time Economics“, *The Wall Street Journal*, 2009. [Online]. Available at: https://blogs.wsj.com/economics/2009/04/22/whats-a-global-recession/. [Vaadatud: 16-märts-2020].

[102] National Commission on the Causes of the Financial and Economic Crisis in the United States, „Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States“, 2011.

[103] Bärí A. Williams, „The Answer to Tech’s Ethical Problems is Greater Diversity“, *FastCompany*, 2018. [Online]. Available at: https://www.fastcompany.com/90259821/the-answer-to-techs-ethical-problems-is-greater-diversity. [Vaadatud: 01-märts-2020].

[104] S. M. Johnson Vickberg ja K. Christfort, „Pioneers, Drivers, Integrators, and Guardians“, *Harvard Business Review*, 2017. [Online]. Available at: https://hbr.org/2017/03/the-new-science-of-team-chemistry#how-work-styles-inform-leadership. [Vaadatud: 01-märts-2020].

[105] S. Jang, „The Most Creative Teams Have a Specific Type of Cultural Diversity.“,

Harvard Business Review Digital Articles, 2018. [Online]. Available at: https://hbr.org/2018/07/the-most-creative-teams-have-a-specific-type-of-cultural-diversity. [Vaadatud: 01-märts-2020].

[106] S. Shaikh, H. Vishwakarma, S. Mehta, K. R. Varshney, K. N. Ramamurthy, ja D. Wei, „An End-To-End Machine Learning Pipeline That Ensures Fairness Policies“, *Bloomberg Data for Good Exchange Conference*, 2017.

[107] R. K. E. Bellamy *et al.*, „AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias“, *IBM Journal of Research and Development*, kd 63, nr 4–5, 2019.

[108] CRISP, „CRISP-DM Help Overview“, *Spss*, 1995. [Online]. Available at: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm. [Vaadatud: 06-veebr-2020].

[109] Google, „Explainable AI | Google Cloud“. [Online]. Available at: https://cloud.google.com/explainable-ai. [Vaadatud: 29-märts-2020].

[110] M. T. Ribeiro, S. Singh, ja C. Guestrin, „‘Why Should I Trust You?’ Explaining the Predictions of Any Classifier“, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, kd 13-17-Augu, lk 1135–1144, aug 2016.

[111] A. Datta, S. Sen, ja Y. Zick, „Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems“, *Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016*, 2016, lk 598–617.

[112] W. Hummer *et al.*, „ModelOps: Cloud-based lifecycle management for reliable and trusted AI“, *Proceedings - 2019 IEEE International Conference on Cloud Engineering, IC2E 2019*, 2019, lk 113–120.

[113] C. Breuel, „MLOps: Machine Learning Engineering | Towards Data Science“, *Towards Data Science*, 2020. [Online]. Available at: https://towardsdatascience.com/ml-ops-machine-learning-as-an-engineering-discipline-b86ca4874a3f. [Vaadatud: 29-märts-2020].

[114] K. Sato, „What is ML Ops? Best Practices for DevOps for ML (Cloud Next ’18) - YouTube“, 2018.

[115] C. Pete *et al.*, „CRISP-DM 1.0: Step-by-step data mining guide“, 2000.

[116] M. S. Brown, „What IT Needs To Know About The Data Mining Process“, *Forbes*, 2015. [Online]. Available at: https://www.forbes.com/sites/metabrown/2015/07/29/what-it-needs-to-know-about-the-data-mining-process/#7e6a5c6b515f. [Vaadatud: 29-märts-2020].

[117] B. D’Alessandro, C. O’Neil, ja T. LaGatta, „Conscientious Classification: A Data Scientist’s Guide to Discrimination-Aware Classification“, *Big Data*, kd 5, nr 2, lk 120–134, 2017.

[118] A. Clark, „The Machine Learning Audit: CRISP-DM Framework“, *ISACA Journal*, kd 1, lk 42–47, 2018.

[119] Smart Vision Europe, „What is the CRISP-DM methodology?“, 2018. [Online]. Available at: https://www.sv-europe.com/crisp-dm-methodology/. [Vaadatud: 07-märts-2020].

[120] S. Pichai, „AI at Google: Our Principles“, *2018-07-07*, 2018. [Online]. Available at: https://blog.google/technology/ai/ai-principles/. [Vaadatud: 29-märts-2020].

[121] A. Cutler, M. Pribić, ja L. Humphrey, „Everyday Ethics for Artificial Intelligence“, IBM Design Program Office, 2019.

[122] Microsoft, „Responsible AI principles from Microsoft“, 2020. [Online]. Available at: https://www.microsoft.com/en-us/ai/responsible-ai. [Vaadatud: 29-märts-2020].

[123] K. R. Varshney, „Introducing AI Fairness 360, A Step Towards Trusted AI“, *IBM Research Blog*, 2018. [Online]. Available at: https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/. [Vaadatud: 29-märts-2020].

[124] R. K. E. Bellamy, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, ja D. Saha, „GitHub - IBM/AIF360: A Comprehensive Set of Fairness Metrics for

Datasets and Machine Learning Models, Explanations for These Metrics, and Algorithms to Mitigate Bias in Datasets and Models“. [Online]. Available at: https://github.com/IBM/AIF360. [Vaadatud: 12-apr-2020].

[125] R. K. E. Bellamy, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, ja D. Saha, „AI Fairness 360 (Web Application)“, 2019. [Online]. Available at: https://aif360.mybluemix.net/. [Vaadatud: 12-apr-2020].

[126] R. T. Dive ja A. P. Khedkar, „An Approach for Discrimination Prevention in Data Mining“, *International Journal of Research in Engineering and Technology*, kd 03, nr 03, lk 412–414, 2014.

[127] M. Zehlike, C. Castillo, F. Bonchi, R. Baeza-Yates, S. Hajian, ja M. Megahed, „Fairness Measures: Datasets and Software for Detecting Algorithmic Discrimination“, 2017. [Online]. Available at: https://www.fairness-measures.org/. [Vaadatud: 29-märts-2020].

[128] F. Tramèr *et al.*, „FairTest: Discovering Unwarranted Associations in Data-Driven Applications“, *Proceedings - 2nd IEEE European Symposium on Security and Privacy, EuroS and P 2017*, 2017, lk 401–416.

[129] F. Tramer, „GitHub - Columbia/FairTest“, *GitHub*, 2020. [Online]. Available at: https://github.com/columbia/fairtest. [Vaadatud: 29-märts-2020].

[130] P. Saleiro *et al.*, „Aequitas: A Bias and Fairness Audit Toolkit“, 2018. [Online]. Available at: http://arxiv.org/abs/1811.05577. [Vaadatud: 29-märts-2020].

[131] P. Saleiro *et al.*, „Aequitas: Bias and Fairness Audit Toolkit“, 2020. [Online]. Available at: https://github.com/dssg/aequitas. [Vaadatud: 29-märts-2020].

[132] N. Bantilan, „Themis-ML: A Fairness-Aware Machine Learning Interface for End-To-End Discrimination Discovery and Mitigation“, *Journal of Technology in Human Services*, kd 36, nr 1, lk 15–30, 2018.

[133] N. Bantilan, „A Fairness-aware Machine Learning Library — Themis-ML 0.0.2 documentation“, 2017. [Online]. Available at: https://themis-ml.readthedocs.io/en/latest/. [Vaadatud: 29-märts-2020].

[134] Sorelle A. *et al.*, „Fairness Comparison: Comparing Fairness-aware Machine Learning Techniques“, *GitHub*, 2019. [Online]. Available at: https://github.com/algofairness/fairness-comparison. [Vaadatud: 29-märts-2020].

[135] S. A. Friedler, S. Choudhary, C. Scheidegger, E. P. Hamilton, S. Venkatasubramanian, ja D. Roth, „A Comparative Study of Fairness-enhancing Interventions in Machine Learning“, *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 2019, lk 329–338.

[136] L. Schallner, J. Rabold, O. Scholz, ja U. Schmid, „Effect of Superpixel Aggregation on Explanations in LIME -- A Case Study with Biological Data“, lk 1–12, okt 2019.

[137] K. Patel, N. Bancroft, S. M. Drucker, J. Fogarty, A. J. Ko, ja J. A. Landay, „Gestalt: Integrated support for implementation and analysis in machine learning“, *UIST 2010 - 23rd ACM Symposium on User Interface Software and Technology*, nr May 2014, lk 37–46, 2010.

[138] S. Amershi, M. Chickering, S. M. Drucker, B. Lee, P. Simard, ja J. Suh, „Modeltracker: Redesigning performance analysis tools for machine learning“, *Conference on Human Factors in Computing Systems - Proceedings*, 2015, kd 2015-April, lk 337–346.

[139] M. W. Craven ja J. W. Shavlik, „Extracting Thee-Structured Representations of Thained Networks“, *Advances in Neural Information Processing Systems*, lk 24–30, 1996.

[140] S. Jordan, R. Day, ja L. M. Ingram, „A Glossary for Discussion of Ethics of Autonomous and Intelligent Systems, Version 1“, 2017.

[141] M. Nõulik, „Turvateadlik tarkvaravalik haridusasutuses“, Tallinna Ülikool, 2016.

[142] D. P. Watson ja D. H. Scheidt, „Autonomous Systems“, *Johns Hopkins Applied Physics Laboratory Technical Digest*, kd 26, nr 4, lk 368–376, 2005.

[143] R. Arkin *et al.*, „Autonomous Weapon Systems: A Roadmapping Exercise“, 2019.

[144] C. Sandvig, K. Hamilton, K. Karahalios, ja C. Langbort, „Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms“, *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, 2014, kd 93, nr 1, lk 181–185.

[145] T. Petzinger, *Hard Landing*. New York: Random House, 1996.

[146] The Associated Press, „Track Officials Called Caster Semenya ‘Biologically Male,’ Newly Released Documents Show“, *The New York Times*, 2019. [Online]. Available at: https://www.nytimes.com/2019/06/18/sports/track-officials-called-caster-semenya-biologically-male-newly-released-documents-show.html?auth=login-google. [Vaadatud: 22-veebr-2020].

[147] K. Lutt, „Inglise-eesti väärtusterminite sõnastik. Magistriprojekt“, Tartu, 2008.

[148] D. Reinsel, J. Gantz, ja J. Rydning, „Data Age 2025: The Digitization of the World From Edge to Core“, *International Data Corporation*, nr november, lk 28, 2018.

[149] J. Hoffman, D. Wang, F. Yu, ja T. Darrell, „FCNs in the Wild: Pixel-level Adversarial and Constraint-based Adaptation“, 2016. [Online]. Available at: http://arxiv.org/abs/1612.02649. [Vaadatud: 24-veebr-2020].

[150] BBC, „Tesla Model 3 on Autopilot crashes into police car“, *BBC News Technology*, 2019. [Online]. Available at: https://www.bbc.com/news/technology-50713716. [Vaadatud: 24-veebr-2020].

[151] N. Wheeler, „Is the Media’s Reluctance to Admit AI’s Weaknesses Putting us at Risk?“, 2019. [Online]. Available at: https://towardsdatascience.com/is-the-medias-reluctance-to-admit-ai-s-weaknesses-putting-us-at-risk-c355728e9028. [Vaadatud: 24-veebr-2020].

[152] G. Branwen, „The Neural Net Tank Urban Legend“, *Gwern.net*, 2011. [Online]. Available at: https://www.gwern.net/Tanks. [Vaadatud: 24-veebr-2020].

[153] Rahvusvahelise Migratsiooniorganisatsiooni (IOM) Eesti esindus, „Eesti tagasipöördumissüsteemi kaardistamine“, 2011. [Online]. Available at:

http://iom.ee/avaldatud_materjalid/Tagasipoordumissysteemi kaardistamine.pdf. [Vaadatud: 29-veebr-2020].

[154] Euroopa Komisjon, *Haavatavate isikute menetluslikud tagatised kriminaalmenetluses. Soovitus.* OJ C 378 of 24.12.2013, 2013.

[155] Enterprise Estonia (EAS), „Väike- ja keskmise suurusega ettevõtja (VKE) definitsiooni selgitus vastavalt Euroopa Komisjoni määruse 800 / 2008 / EÜ lisa 1-le“, 2008.

[156] EY, „Worldwide Capital and Fixed Assets Guide 2018“, 2018. .

[157] W. Wallach ja C. Allen, *Moral Machines*. Oxford University Press, 2009.

[158] P.-H. Wong ja J. Simon, „Thinking About ‘Ethics’ in the Ethics of AI“, 2020. [Online]. Available at: https://revistaidees.cat/en/thinking-about-ethics-in-the-ethics-of-ai/?pdf=9399. [Vaadatud: 24-veebr-2020].