

Algoritmilise kallutatuse riskihalduse metoodika

Metoodika

Version 1.1

01.10.2025

D-16-572

Avalik

©2025 Justiits- ja Digiministeerium

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

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

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

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

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

Kuupäev	Versioon	Kirjeldus
13.06.2025	0.5	Esimene kallutatuse hindamise protsessi kattev mustand
01.07.2025	0.6	Testimise järel lisatud ohtude kandidaatide leht. Abimaterjalid toodud lähemale sinna, kus neid vaja on. Vormistuse ja sõnastuste parandused.
09.07.2025	1.0	Ohu kirjeldusse lisatud õigusnormi rikkumine.
01.10.2025	1.1	Mitmed täiendused ja parandused rakendajate tagasiside põhjal. Töölehele on lisatud korduma kippuvad küsimused.

Sisukord

1 Sissejuhatus.....	5
1.1 Eesmärk	5
1.2 Käsitlusala.....	5
1.3 Terminid ja lühendid	5
2 Metoodika kasutamisest.....	7
3 Tehisintellektisüsteemi passi koostamine	9
3.1 Kes, millal ja milleks?	9
3.2 Ettevalmistus	9
3.3 Töölehe täitmine	9
3.4 Järeltegevused	10
4 Tehisintellektiga seotud kasutuslugude kirjeldamine.....	12
4.1 Kes, millal ja milleks?	12
4.2 Ettevalmistus	12
4.3 Töölehe täitmine	13
4.4 Järeltegevused	15
5 Kallutatuse ohtude hindamine.....	16
5.1 Kes, millal ja milleks?	16
5.2 Ettevalmistus	16
5.3 Töölehe täitmine	17
5.4 Järeltegevused	18
6 Kallutatuse riskide käsitlus	19
6.1 Kes, millal ja milleks?	19
6.2 Ettevalmistus	19
6.3 Töölehe täitmine	19
6.4 Järeltegevused	20

1 Sissejuhatus

1.1 Eesmärk

Kui tehisintellekti ja algoritmiliste süsteemide kallutatuse hindamise juhend kirjeldab algoritmilise kallutatuse mõistete ruumi, kontseptsioone ja nende seoseid, siis metoodika on juhendi eluviimise abivahend.

Metoodikat järgides saab isik teha konkreetseid tegevusi ja samme, mis aitavad analüüsida algoritmilist kallutatust ning hallata riske AI-süsteemides ja algoritmilist otsustamist kasutavates süsteemides.

Metoodika on paindlik ning seda saab kasutada koos organisatsioonis rakendatud riskihalduse standarditega. Metoodika kasutamist toetab täidetav tööleht tabelitöötamise rakenduses.

1.2 Käsitlusala

Algoritmilise kallutatuse riskihalduse töövahend keskendub algoritmilise ja AI-süsteemide kallutatuse riskide haldamisele. Töövahend koosneb kolmest osast: juhendist, mis kirjeldab tehisintellektisüsteemide ja nende kallutatuse olemust ja tausta ning vaatleb kallutatuse tuvastamise ja leevendamise võimalusi; metoodikast, mis annab detailsed juhised riskihaldusprotsessi ülesseadmiseks ja läbiviimiseks; ning töölehest, mis lihtsustab riskihalduseks vajaliku teabe dokumenteerimist. Töövahendi käsitlusalas on eelkõige süsteemid, mida kasutatakse organisatsioonides. Erasisikute omaks tarbeks kasutatavatele algoritmilistele ja AI-süsteemidele kehtivad väiksemad nõuded. Vaatleme riskihaldusprotsessi üldisena, sest süsteemid võivad olla väga erinevad ja kallutatuse allikas võib olla erinev. Olenevalt süsteemi juurutusest, võib organisatsioonil puududa võimalus riske tehnoloogiliste vahenditega hinnata, seega peab neid käsitlema üldisemalt.

1.3 Terminid ja lühendid

AI-süsteem

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

Algoritm

astmeline eeskiri mingi tegevuse sooritamiseks

kaitsepiirded (*guardrails*)

tehnoloogia või raamistik, mis toimib turvakontrollpunktina, hinnates kasutaja sisendit ja mudeli genereeritud vastuseid lähtudes eelnevalt määratud ohutusreeglitest. Kaitsepiirete eesmärk on reaajas ennetada kahjuliku, sobimatu või teemavälise sisu genereerimist, kui mudeli ohutushäällestusest selleks ei piisa.

kallutatatus (*bias*)

teatud objektide, isikute või rühmade käsitlemise süstemaatiline erinevus teistega võrreldes; käsitlemine on ükskõik milline tegevus, sealhulgas tajumine, vaatlus, esitus, prognoosimine või otsustus

masinõppemudel

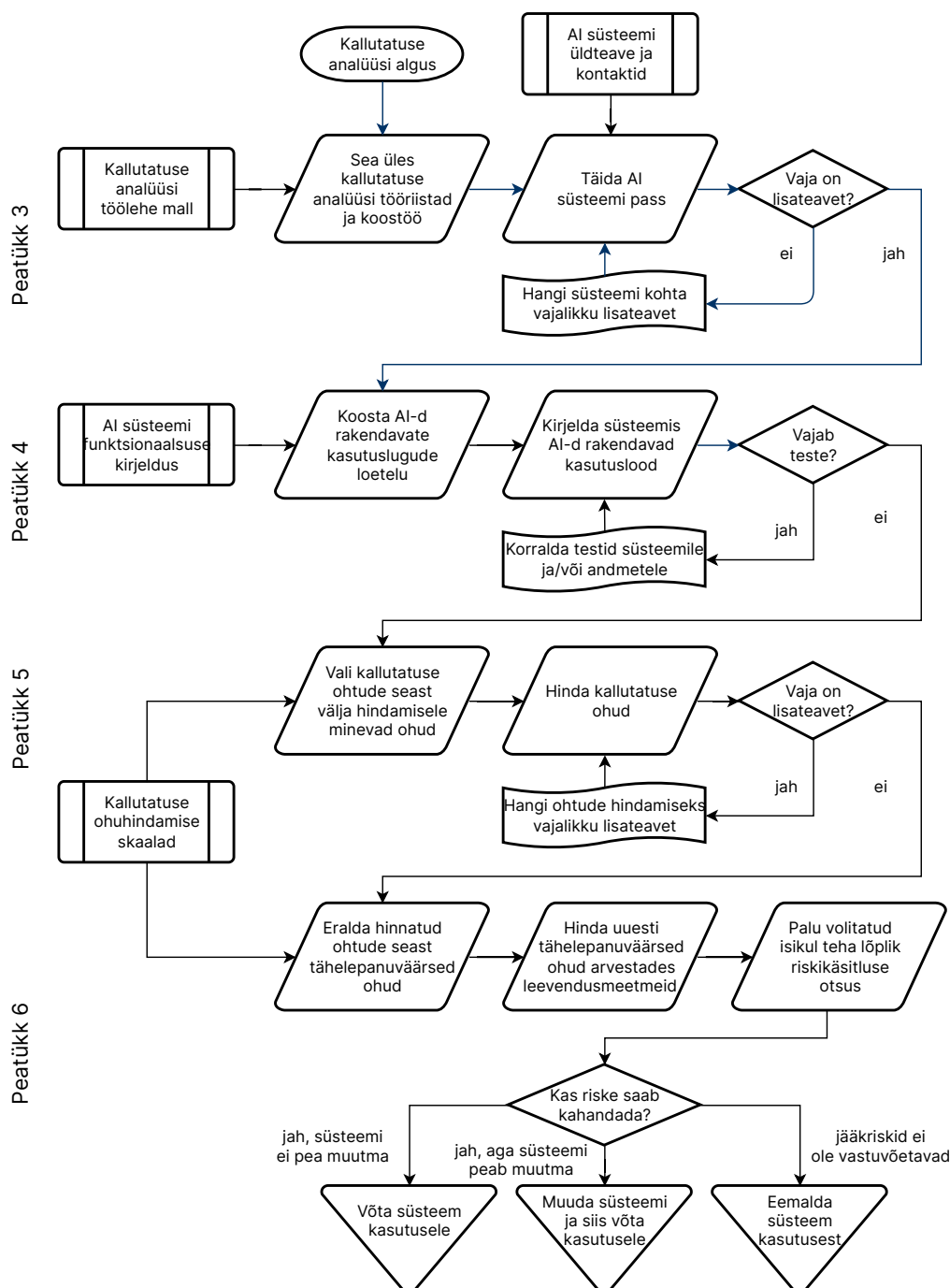
spetsiifiline algoritm või algoritmide kogum, mille abil saab teatud sisendväärtuste põhjal ennustada väljundeid

seletavus (*explainability*)

AI-süsteemi omadus väljendada oma tulemusi mõjutavaid olulisi tegureid inimestele arusaadaval viisil; on mõeldud vastama küsimusele "Miks?" püüdmata väita, et valitud tegevusviis on tingimata optimaalne

2 Metoodika kasutamisest

Tehisintellekti ja algoritmilise kallutatuse riskianalüüsi üldine protsess on joonisel 1. Toome ära ka metoodika peatükid, mis kirjeldavad, kuidas protsessi etappe läbida.



Joonis 1. Kallutatuse analüüsi töövoog

Metoodika katsetamise käigus küsisid kasutajad mitmeid küsimusi, mida usutavasti küsivad ka teised tulevased kasutajad. Anname siin neile vastused ja selgitused.

Mida teha, kui ma alles arendan süsteemi, mille paljud detailid on veel teadmata?

Alusta varakult, sest arendusfaasis süsteemi muutmine on odavam. Jäta töölehe lahtreid julgelt tühjaks, kasuta kavandites ja plaanides olevat infot ja muuda neid vastavalt riskianalüüsile!

Kuidas töölehte kasutada, kui kaalutakse olemasolevale süsteemile algoritmi või AI lisamist?

Juhul kui varasemaid otsuseid pole vaja ümber teha, käsitle vaid muudetavaid või uusi funktsioone, nii on töö jõukohasem.

Kui palju kasutuslugusid on mõistlik kirja panna?

1–4 võiks üldjuhul ära katta peamised AI funktsioonid. Samas proovi teha iga olulise AI-mudeli või algoritmi kohta üks kasutuslugu.

Kas ohtudest tuleb kindlasti mõelda kvantitatiivselt (protsentides või ohustatud inimestes) või võib hinnata ka kvalitatiivselt?

Võib hinnata kvalitatiivselt, aga siis on soovitatav leida rohkem hindajaid ning otsustada, kuidas ohtude tõsidust omavahel võrrelda.

Mida teha, kui minu organisatsioonis jääkriske aktsepteerida ei soovita?

Kõik inimeste poolt ettevõetav on riskidele avatud, riskideta elu on üldiselt võimatu või piiravalt kallis. Soovitame jagada teistega kogemusi praktiliste riskide hindamise teemal.

3 Tehisintellektisüsteemi passi koostamine

3.1 Kes, millal ja milleks?

Teostaja. Tehisintellektisüsteemi passi andmete täitmise eest vastutab tehisintellektisüsteemi omanik. Töö efektiivsuse jaoks võib ta selle delegeerida projektijuhile, tehnilisele eksperdile, turbe- või õigusmeeskonnale enda organisatsiooni sees või lepingulise arenduspartneri juures.

Aeg. Tehisintellektisüsteemi passi täitmisega võib alustada nii süsteemi lähteülesande loomisel, kavandamisel, arenduse käigus, riskianalüüsi alguses kui ka auditeerimisel. Mida varem alustatakse, seda vähem tehnilist teavet tehisintellektisüsteemist on saadaval, kuid samas on ka rohkem võimalusi süsteemi arengu suunamiseks, nõuete ja leevendusmeetmete kehtestamiseks.

Eesmärk. Tehisintellektisüsteemi passis on kirjas, millist süsteemi hinnatakse ja analüüsitakse. Süsteemist võib olla erinevaid versioone ja eri ajal tehtud hindamisi. Passi pannakse kirja nii süsteem kui ka hindamise ajahetkel sellega seotud isikud.

3.2 Ettevalmistus

Samm 1: valmista ette tööleht. Tee järgmised tegevused.

1. Leia uusim versioon tehisintellektisüsteemi kallutatuse analüüsi töölehe failist.
2. Loo töölehest uus koopia ja anna failile analüüsitava süsteemiga seostuv nimi.
3. Talleta loodud fail süsteemi dokumentatsiooni hoidlas, millele koostööpartnerid juurde pääsevad.

Samm 2: otsusta töölehe täitmise viis ja too kokku tööühm. Tee järgmised tegevused.

1. Selgita välja, kes võivad teada rohkem hinnatava tehisintellektisüsteemi tehnilisi detaile, õiguslikku raamistikku või riske (nt andmeteadlased, andmekaitsekespert, infoturbejuht, kvaliteedijuht, riskijuht, teised vastutavad juhid).
2. Selgita välja, kes peavad olema kallutatuse analüüsi kaasatud. Selleks uuri organisatsiooni kvaliteedisüsteemi, infoturvapoliitikat, andmekaitsepoliitikat, ühiskondlike mõjude poliitikat, riskihalduse protseduure ning üldiseid vastutusahelaid.
3. Otsusta, kas täidate töölehte üheskoos samaaegselt koosoleku vormis või igaüks omaette endale sobival ajal. Sea vastavalt sellele üles koostöövahendid töölehe täitmiseks.
4. Määra koostöövormile vastavalt istungite ja panustamise ajakava (istungite ajad, isiklike panuste tähtajad). Jaga tööülesanded kirja teel või tööde halduse süsteemi kaudu.

3.3 Töölehe täitmine

Samm 1: kirjelda AI-süsteemi, et hiljem oleks teada, mida täpselt hindasite. Täida alamlehel AI-SÜSTEEMI PASS (1-5) kollased lahtrid 1.1–1.X vastavalt töölehe juhiste, arvestades järgmiseid soovitusi.

1. Versiooni määramisel võid kasutada konkreetset versiooninumbrit või -nime, aastaarvu, koodihoidla kehtestuse koodi vms teavet, mille abil saab hiljem vältida segadusi teemal, millist varianti süsteemist hinnati.

2. Rakendava asutuse määramisel mõtle, kes on teenuse omanik, hankija või väljatöötaja – kelle käes on vastutus.
3. Lahtrisse 1.x sobivad viited. Veebiteenuse korral AI-süsteemi aadress (URI/URL) või seda tutvustav veebileht, äpi puhul viide rakenduse poele, vikileht, dokumendihoidla, lähtekoodi hoidla. Vajaduse korral lisa ridu.

Samm 2: täpsusta AI-süsteemiga seotud pooli, et teada oleks vastutajad ning mõjutatud isikud. Täida alamlehel AI-SÜSTEEMI PASS (1-5) kollased lahtrid 2.1–2.X vastavalt töölehe juhiste, arvestades järgmiseid soovitusi.

1. Euroopa Liidu tehisintellektimäärus toob välja rollid: teenustaja, tootja, juurutaja, importija ja turustaja. Vaata ka juhendi alajaotist 2.2.
2. Arenduspartnereid võib AI-süsteemil olla mitu. Lisa nimi, vastutusala ning aastate vahemik, millal partner aktiivne oli.
3. Peamiste mõjutatud isikutena loetle süsteemi peamine kasutajarühm (nt mõne asutuse taotluste menetlejad, otsuste tegijad), otsustest mõjutatud sihtrühm (nt haldusmenetluses osalev kodanik). Kui võimalik, lisa laiem ühiskonnarühm, keda süsteem puudutab (nt tervishoiuteenuse osutajad, lapsevanemad, alaealised, avalikus ruumis või hoonetes valvekaamerate vaatlusalas liikujad, sidesüsteemide kasutajad).

Samm 3: sõnasta AI-süsteemi eesmärgid, et oleks teada, miks süsteemi vaja on. Täida alamlehel AI-SÜSTEEMI PASS (1-5) kollased lahtrid 3.1–3.X vastavalt töölehe juhiste, arvestades järgmiseid soovitusi.

1. Otseste eesmärgidena pane kirja oodatavad funktsioonid. Mida peab süsteem tehisintellekti või algoritmi toel ära tegema? Milliseid otsuseid toetama? Millist inimlikuks peetavat ülesannet täitma? Vaata ka juhendi alajaotist 4.2.
2. Kaudsete eesmärgidena pane kirja, milleks neid funktsioone oli vaja teostada. Millist muudatust soovitakse ühiskonnas saavutada? Millist protsessi efektiivsemaks muuta?
3. Mõõdikuid ei pea ise välja mõtlema, need võivad tulla süsteemi omanikult.

Samm 4: pane kirja kontaktid, et oleks võimalik teha järelepärimisi. Täida alamlehel AI-SÜSTEEMI PASS (1-5) kollased lahtrid 4.1–4.7 vastavalt töölehe juhiste, arvestades järgmiseid soovitusi.

1. Kõiki kontakte ei pea passi märkima. Eesmärk on lihtsustada edasist tööd, nii et kontaktid oleks lihtsasti leitavad.

Samm 5: lisa viited tehnilistele dokumentidele, et hinnangute andmisel teaks kõik, kuidas süsteem on ette nähtud töötama. Täida alamlehel AI-SÜSTEEMI PASS (1-5) kollased lahtrid 5.1–5.9 vastavalt töölehe juhiste, arvestades järgmiseid soovitusi.

1. Dokumentide olemasolu sõltub sellest, millises elutsükli etapis tehisintellektisüsteem on. Kui süsteem on alles visioonifaasis, võib olemas olla visioonidokument. Kui süsteem on tootnuküps, on olemas rohkem dokumente.
2. Dokument võib olla konkreetse failitüübi, vormingu või standardiga. Peamine, et olemas on lahtri juhendis kirjeldatud sisu.

3.4 Järeltegevused

Samm 1: kontrolli kogutud teabe kvaliteeti koos töörühma ja partneritega. Tee järgmised tegevused.

1. Juhul kui töörühma kõik liikmed ei saanud töös osaleda, on pärast töölehe täitmist soovitatav teha lühikese tähtajaga kooskõlastusring. Eesmärk on materjalis olevate lünkade täitmine ning kõigile järgmiseks tegevuseks (süsteemis tehisintellekti kasutava funktsionaalsuse tuvastamiseks) ühtse inforuumi loomine.

4 Tehisintellektiga seotud kasutuslugude kirjeldamine

4.1 Kes, millal ja milleks?

Teostaja. Tehisintellektiga seotud kasutuslugudest peab teadlik olema tehisintellektisüsteemi omanik. Süsteemi funktsioonide ja tehniliste detailide väljaselgitamiseks on mõistlik kaasata teenuse omanik, süsteemianalüütik, arhitekt ja andmeteadlased. Edasise töö lihtsustamiseks soovitame kas tegevuse käigus või järeltegevusena tulemite kontrollimisel kaasata ka õiguseksperdi ja/või riskijuhi.

Aeg. Tehisintellekti rakendavad kasutuslood on selle metoodika peamine uurimisobjekt, sest kasutuslugudes ja nendega seotud talitlusmudelites on kõige paremini näha andmevood ja nendega seotud isikud. Süsteemi visiooni ja eelanalüüsi faasis võib detaile täpse analüüsi jaoks napida – sellisel juhul soovitame täita kasutusloo info nii hästi kui võimalik, ilma tehisintellektikomponendi tehniliste detailiteta, mis selguvad hiljem. Juhul kui kallutatust uuritakse juba teostatud süsteemis, on ka detaile rohkem saadaval ning ohtude hindamine selle võrra lihtsam.

Eesmärk. Tehisintellekti komponente kasutavate peamiste kasutuslugude kirjapanek võimaldab selgelt tuvastada isikud, keda kallutatus AI-süsteemis võib mõjutada. Lisatud tehnilised detailid aitavad hinnata kallutatuse ulatust.

4.2 Ettevalmistus

Samm 1: vajaduse korral tutvu tehisintellekti ja andmetega seotud kontseptsioonidega. Tee järgmised tegevused.

1. Korda üle, mis on tehisintellekt, masinõppemudelid ja otsusetoe algoritmid selleks, et neid süsteemis ära tunda. Seda aitavad teha juhendi alajaotised 2.1 ja 2.3.
2. Korda üle andmetega seotud määratlused (isikustatud andmed, isikustatavad andmed, eriliigilised isikuandmed ja ärisaladusega seotud andmed). Seda aitab teha juhendi alajaotis 3.2.
3. Võimaluse korral jaga enda leide ja sobivaid abimaterjale töörühmaga, et ka nemad oleks tööks ette valmistatud.

Samm 2: tutvu tehisintellektisüsteemi funktsiooni kirjeldavate allikatega. Tee järgmised tegevused.

1. Tutvu kallutatuse analüüsi töölehe jaotises 5 (tehnilised detailid) viidatud materjaliga.
2. Leia materjalidest funktsiooni kirjeldused (kasutusloomallid, talitluse mudelid ja skeemid, stsenaariumid ja muu materjal). Loe need mõttega läbi ja hinda, kas suudad neis tuvastada seotud pooled ja andmeelemendid. Tee endale tehisintellektiga seotud kasutuslugude mustand. Grupeeri sarnaseid seotud pooli hõlmavad või tehisintellektikomponenti/algoritmi sarnaselt kasutavad kasutuslood üheks kokku, et kallutatuse analüüsi töömahtu vähendada. Kasutuslugude tuvastamist toetavaid näiteid leiad juhendi alajaotisest 4.2.2.
3. Leia materjalidest teave selle kohta, milliste funktsioonide täitmisel kasutab analüüsiv süsteem tehisintellekti või otsusetoe algoritme. Leia viidatud dokumentatsioonis neid kirjeldav materjal (kui see on olemas). Tehisintellektikomponentide äratundmist toetavad juhendi ala-

jaotised 2.1 ja 2.3.

4. Vajaduse korral jaga oma leide ja sobivaid abimaterjale töörühmaga, et ka nemad oleks tööks ette valmistatud.

4.3 Töölehe täitmine

Samm 1: koosta tehisintellekti või otsusetoe algoritmi kasutavate kasutusjuhtumite loetelu. Täida alamlehel AI KASUTUSLOOD (6) kollased lahtrid 6.1–6.9 (lisa või eemalda neid vajaduse järgi) vastavalt töölehe juhiste, arvestades järgmiseid soovitusi.

1. Juhul kui teed tööd töörühmaga, esitle ettevalmistuse käigus tuvastatud kasutuslugusid, mille käigus kasutatakse süsteemi tehisintellektikomponenti või otsusetoe algoritmi. Kontrolli, kas teised süsteemi tundvad isikud nõustuvad tehisintellekti kasutavate kasutuslugude loeteluga! Arutlege, kas on vaja midagi lisada, eemaldada või milliseid kasutuslugusid saaks aja kokkuhoiu huvides vaadata ühtsena (samad kasutajad, mõjutatud isikud ning tehisintellektikomponent või algoritm).
2. Ideaalis oleks süsteemil 1–7 peamist tehisintellektiga seotud kasutuslugu, aga midagi pole katki ka siis, kui neid on rohkem. Plaani sellisel juhul lihtsalt töö jaoks rohkem aega.
3. Kui töölehe alamlehel AI KASUTUSLOOD (6) on jaotis 6 täidetud, siis kopeeri alamlehte AI KASUTUSLUGU A (KOPEERI MIND) nii mitu korda, et iga kasutusloo jaoks oleks üks alamleht ning iga alamlehe pealkiri oleks AI KASUTUSLUGU ?, kus ? on järjestikuline tähestiku suurtäht (nagu jaotises 6). Pane iga loodud alamlehe lahtrisse A5 kasutuslugu sümboliseeriv suur tähestikutäht. Võiks tekkida alamlehed AI KASUTUSLUGU A, AI KASUTUSLUGU B ja nõnda edasi.

Samm 2: kirjelda tehisintellekti või otsusetoe algoritme kasutavaid kasutuslugusid üldisel tasemel. Täida alamlehel AI KASUTUSLUGU A (ja ülejäänud alamlehtedel) kollased lahtrid A.1–A.X (seejärel järgmisel alamlehel B.1–B.X jne) vastavalt töölehe juhiste, arvestades järgmiseid soovitusi.

1. Kasutusloo algataja (A.2) võib olla süsteemi kasutaja (isik) või ka süsteemi automaatne komponent, kes või mis reageerib sisendandmetele (kellaajale, saabunud teatele, saabunud kaamerapildile). Mõlemal juhul on oluline kirja panna, mis on kasutusloo ja tehisintellekti käivitamist tingiv algsündmus (A.3).
2. Minimaalselt võiks tehisintellekti või otsusetoe algoritmi kohta teada, mis on selle sisendandmete (A.4), töötlemise eesmärgi (A.5) ja väljundandmete (A.6) üldine iseloom.
3. Kallutatuse analüüsiks on erakordselt oluline kirja panna (A.7), mismoodi tehisintellekti või algoritmi väljundit kasutatakse. Kas mõni süsteemi kasutaja tugineb sellele otsuse tegemisel? Kui see on nii, siis millises ulatuses: kas lihtsalt masina soovitusena või toimub kasutaja peas masinalt saadud andmete põhjal aktiivne järeletöötlus otsuseni jõudmiseks? Või teeb masin otsuse automaatselt ning viib selle (näiteks teiste süsteemidega koos) ellu?
4. Kui siin punktis kasutatakse täiendavaid allikaid lisaks varem nimetatule, pane need kirja lahtrisse (A.X).

Samm 3: täpsusta sisendandmete iseloom (nii palju kui võimalik). Täida alamlehel AI KASUTUSLUGU A (ja ülejäänud alamlehtedel) kollased lahtrid AA.1–AA.X (seejärel järgmisel alamlehel BB.1–BB.X jne) vastavalt töölehe juhiste, arvestades järgmiseid soovitusi.

1. Andmete koosseisu ja liikide (lahtrid AA.1–AA.4) tuvastamine peab ennekoike aitama mõista, millised andmed võivad olla seotud otseselt (isikustatud) või kaudselt (isikustatavad) konk-

reetsete füüsiliste või juriidiliste isikutega ja põhjustada nende diskrimineerimist automaatses või masina toetatud otsustusprotsessis.

2. Andmete kogumise viis (lahter AA.5) peab aitama selgitada nende kvaliteeti ja täpsust. On märkimisväärne vahe, kas sisendandmed tulevad mõnest riigi andmekogust digitaalse identiteediga tuvastatud isiku kohta või on need näiteks tuvastatud mürarikkamast kanalist (tekstisõnumitest, aga ka heli, pildi või video analüüsist). Kui andmed pärinevad mitmest andmekogust, loetle need, et hiljem saaks uurida, kui kvaliteetsed on sealsed andmed ning kas need sobivad andmete analüüsiks.
3. Tervikluse ja konfidentsiaalsuse kaitse meetmed (lahter AA.5) peavad toetama tehisintellekti või algoritmi sisendandmete korrektsust ja ka seda, kui kaitstud on andmed protsessi käigus. Õiguslikel või tervemõistuslikel põhjustel konfidentsiaalsusnõuetega andmeid tuleks töödelda täiendavate lisakaitsemeetmetega, et vältida nende lekkimist menetluse käigus ja võimalikku planeerimatut diskrimineerimist väljaspool ametlikke kasutuslugusid.

Samm 4: täpsusta tehisintellekti või algoritmiga töötlemise tehnilise detailid (nii palju kui võimalik). Täida alamlehel AI KASUTUSLUGU A (ja seejärel samamoodi B, C jne) kollased lahtrid AAA.1–AAA.X (seejärel järgmisel alamlehel BBB.1–BBB.X jne) vastavalt töölehe juhiste, arvestades järgmiseid soovitusi.

1. Esimesed neli lahtrit (AAA.1–AAA.4) aitavad tuvastada tehisintellektikomponendi tehnilist teostust. Sõltuvalt süsteemist võib siin olla toote nimi, masinõppemudeli nimi, algoritmi nimi raamatust või teadusartiklist. Versiooni (number, nimi, aasta – kui on asjakohane) teadmine aitab leida teavet konkreetse komponendi omaduste kohta. Siia võid panna ka mitu mudelit või nime, eriti kui need töötavad koos.
2. Mudeli või algoritmi liidestus süsteemiga (AAA.5) peab aitama mõista, kuidas sisendandmed mudelile või algoritmile ette antakse. Lisaks tuleb kirja panna, kelle kontrolli all algoritm või mudel töötab (süsteemi käitaja kontrolli all tema enda andmekeskuses kuni siis rakendusliidesena teenusena pilvelt teises äärmuses). Selle teadmine on oluline, sest aitab hinnata riski, et mudel või algoritm muutub ilma rakendaja teadmata ning süsteemi käitumine seetõttu muutub.
3. Kui on piisavalt teada mudeli või algoritmi loomisel kasutatud treening- ja testandmete kohta, siis selle info võib kirja panna lahtrisse (AAA.6). NB! See informatsioon ei ole kohustuslik, sest ei tarvitse olla kättesaadav. Kui sellele teabele lisada juurde teadaolevad kvaliteedimõdikud (AAA.7), saame selle järgi hiljem hinnata, milliste andmete kvaliteet või vead võivad väljundit mõjutada.
4. Ideaalsel juhul on tehisintellektikomponendi looja juba dokumentatsiooni osana selgitanud, milliseid kallutatuse vähendamise või kontrollimise meetmeid (lahter AAA.8) ta on rakendanud ning kas neid on kuidagi testitud (AAA.9). Võib ka juhtuda, et dokumentatsiooni meetmete kohta pole, aga testitulemused on ikka. NB! Oluline on mõista, et mõlema lahtri mõtestamiseks saab ka kallutatuse riskianalüüsi korraldaja tellida mudeli või algoritmi uuringu või testimise.

Samm 5: täpsusta tulemi elluviimist (nii palju kui võimalik). Täida alamlehel AI KASUTUSLUGU A (ja seejärel samamoodi B, C jne) kollased lahtrid AAAA.1–AAAA.X (seejärel järgmisel alamlehel BBBB.1–BBBB.X jne) vastavalt töölehe juhiste, arvestades järgmiseid soovitusi.

1. Väljundandmete koosseisu lahtrit (AAAA.1) lugedes peab olema aru saada, millised on masinõppemudeli või algoritmi väljastatud andmed, mille pealt inimene või masin otsust tegema hakkab. Lahtrites AAAA.2–AAAA.4 peab olema kirjas, millised neist andmetest aitavad tuvastada isikut, tema eriliigilisi andmeid või siis organisatsiooni ärisaladust või salastatud teavet.

2. Lahtris AAAA.5 soovitame lisada detaile võrreldes lahtriga A.7 ja keskenduda just seletavusele ja läbipaistvusele. Selle lahtri lugeja peab mõistma, kui palju saab mõjutatud isik teavet enda kohta tehtud otsuse protsessi kohta, sh kasutatud mudelite või algoritmide ja nende väljundi kohta.
3. Lahtris AAAA.6 tuleb kirjeldada kasutatavaid tulemi korrektsuse tagamise meetmeid. Kui kasutusel on hallutsineerimisvõimeline masinõppesüsteem, peab sellele lahtrile topeltsuurt tähelepanu pöörama. Lisaks tuleb kirja panna, mismoodi on kaitstud masinõppemudeli või algoritmi konfidentsiaalsusnõuetega andmed (olgu need seotud isikustatavate andmete, ärisaladuse või salastatud teabega).

4.4 Järeltegevused

Samm 1: kontrolli kogutud teabe kvaliteedi tööühma ja partneritega. Tee järgmist.

1. Kui tehnilist infot jääb puudu, aga on põhjust arvata, et see peaks olemas olema (nt süsteem on valmis ehitatud), siis soovitame pöörduda jaotises 4 loetletud kontaktide poole ja paluda neil puuduvaid lahtreid täita.
2. Kui süsteemi pole testitud, aga testimine oleks võimalik, siis soovitame tutvuda juhendi alajaotistega 5.2 ja 5.3 ja korraldada kallutatuse testimine kas musta kasti või valge kasti meetodil.
3. Kui tööühma kõik liikmed ei saanud töös osaleda, siis pärast töölehe täitmist on soovitatav teha lühikese tähtajaga kooskõlastusring. Eesmärk on materjalis oleva lünkade täitmine ning kõigile järgmiseks tegevuseks (ohtude kirjeldamiseks ja analüüsiks) ühtse inforuumi loomine.

5 Kallutatuse ohtude hindamine

5.1 Kes, millal ja milleks?

Teostaja. Tehisintellektisüsteemi kallutatuse ohtude kirjeldamise eest vastutab süsteemi omanik, kes teeb koostööd õiguseksperdi, riskijuhi ja/või turbeeksperdiga. Kaitsemeetmete kirjeldamiseks on otstarbekas kaasata süsteemi talitlust ja tehnilist teostust tundvaid tehnilisi isikuid.

Aeg. Kallutatuse ohtude analüüsiks kõige efektiivsem aeg on hetk, kui esmane tehniline lahendus on teada, kuid see pole veel lõplikult teostatud. Sellises olukorras saab süsteemi leitud ohtude järgi veel kuluefektiivselt muuta. Tarkvaraarenduse elutsükli mõistes on hea hetk siis, kui valminud on kas arhitektuur või prototüüp, kuid arendustöö ei ole veel alanud. Olukord erineb toote ostmise puhul. Sellisel juhul tuleb riskianalüüs teha siis, kui toode või selle kandidaadid on teada, aga enne tootesse suuremate investeeringute tegemist.

Eesmärk. Ohtude kirjeldamise ja hindamise käigus saab selgeks, millised tehisintellekti kallutatusega seotud riskid vajavad meie edasist tähelepanu. Siin etapis tehtud hinnangute põhjal on võimalik jätkata riskikäsitlemisega.

5.2 Ettevalmistus

Samm 1: tutvu tehisintellekti kallutatusega seotud ohtudega. Tee järgmist.

1. Tutvu juhendis toodud näidisjuhtumitega, kus kirjeldatakse tehisintellektisüsteemide kallutatuse kontseptsiooni ja sellega seotud ohte (alajaotis 4.3).
2. Võimaluse korral tutvu varasemate (pole vahet, kas selle või teiste metoodikate järgi tehtud) kallutatuse riskianalüüsides ja enda organisatsiooni riskitaluvust kirjeldavate dokumentide ja poliitikatega.
3. vajaduse korral jaga enda leide ja sobivaid abimaterjale töörühmaga, et ka nemad oleks tööks ette valmistatud.

Samm 2: määra tingimused, millest alates ohtu hindama hakatakse. Tee järgmist.

1. Riskianalüüsi efektiivsuse huvides on otstarbekas paika panna, millisest tasemest alates riske hindama ja käsitlema hakatakse.
2. Hindamisele tuleks saata ohud, mille realiseerumine on süsteemi eluea ajal võimalik ning millel on reaalsed võimalikud kahjud. Vaata üle alamlehe VÕIMALIKUD OHUD (7) jaotis ABISTAVAD KÜSIMUSED ja seal all olevad tingimused ja lävendid. Vali lävendi näidiste alt enda organisatsiooni jaoks sobivad ning kirjuta need helehallidesse valitud lävendite lahtrisse.
3. Tähelepanu! Kui tehisintellektisüsteemi omaniku organisatsioonis on varasemalt kallutatuse riske analüüsitud (käesoleva või mõne muu metoodika järgi), siis kaalu lävendite taaskasutamist.

Samm 3: määra mõju ja võimalikkuse hinnete skaalad. Tee järgmist.

1. Riskide ühtlaseks hindamiseks on vaja paika panna hindamisskaalad, mis peaksid käsitlemise alla võetud riskide seas kehtestama võrdluskala.
2. Vaata alamlehte OHU HINDAMISE PROTOKOLLI NÄIDIS ning täida ära mõju ja võimalikkuse hindamise skaalade tabelite helehallid lahtrid. Võid kasutada paremal pool antud näidiseid.
3. Tähelepanu! Kui tehisintellektisüsteemi omaniku organisatsioonis on varasemalt kallutatuse

riske analüüsitud (käesoleva või mõne muu metoodika järgi), siis kaalu hindamisskaalade taaskasutamist.

4. Tähelepanu! Kui varem hinnatud süsteemile tehakse uut hindamist, tuleb uute skaaladega ümber hinnata kõik riskid. Kõik riskid peavad võrreldavuse huvides olema hinnatud sama skaalaga.

5.3 Töölehe täitmine

Samm 1: koosta võimalike ohustsenaariumite loetelu. Täida alamlehel VÕIMALIKUD OHUD (7) lahtrid 7.1, 7.2 jne.

1. Vaata järjest kõiki kirja pandud kasutuslugusid. Jäta käsitlusest välja need kasutuslood, kus tehisintellektikomponenti väljund ei mõjuta isikute, asutuste või nende gruppide kohta tehtavaid otsuseid.
2. Allesjäänud kasutuslugude kohta hinda, kas sellega võib olla seotud sellega seotud mõni ohustsenaarium. Ohustsenaariumite näited leiad juhendi alapeatükist 4.3.
3. Kui näed ohustsenaariume, pane need alamlehele järjest kirja, lisa vajaduse korral ridu.
4. Väga tähtis on, et ohustsenaariumi kirjeldus oleks sõnastatud lühidalt ja konkreetset kahju välja tuues. Soovitame kasutada vormelit "konkreetne sündmus või nende lühike jada põhjustab konkreetse kahju". Vt näiteid töölehel VÕIMALIKUD OHUD (7).
5. Tähelepanu! Ohtude kandidaatide loetlemisel ei pea muretsema selle pärast, kas oht on piisavalt tõsine. See on järgmine etapp.

Samm 2: vali, millised ohud lähevad hindamisele. Täida alamlehel HINNATAVAD OHUD (O) kollased lahtrid O1.1, O1.2, O2.1, O2.2, O3.1, O3.2 jne iga ohustsenaariumi kohta, mille te hindamisele saadate. Toetu järgmistele soovitudele.

1. Tee igale ohustsenaariumi kandidaadile alamlehel VÕIMALIKUD OHUD (7) eelhindamine selleks, et otsustada, kas ta jääb käsitluse. Kasuta tegevuse ette valmistamisel paika pandud lävendeid töölehe jaotises ABISTAVAD KÜSIMUSED.
2. Kui ohustsenaarium vastab mõnele tingimusele ja lävendile, siis kanna ohukirjeldus alamlehele HINNATAVAD OHUD (O) ning täida esimesed kolm lahtrit, kus kirjeldad ohustsenaariumit.

Samm 3: leia ohustsenaariumi võimalikkuse hindamiseks vajalik teave. Täida alamlehel HINNATAVAD OHUD (O) kollased lahtrid O1.3, O1.4, O1.5, O1.6, O1.7, O2.3, O2.4, O2.5, O2.6 jne iga tuvastatud ohustsenaariumi kohta, arvestades järgmiseid soovitusi.

1. Eeldused ohustsenaariumi realiseerumiseks (lahter O1.3) peaksid sisaldama vähimat hulka samme, mis peavad korraga või teineteise järel juhtuma, et ohustsenaarium tõeks saaks ning kahjud tekiksid.
2. Ohustsenaariumi realistlikkust (O1.4) on kõige parem hinnata, kui maailmas on juba teada analoogiline realiseerunud ohustsenaarium (vt juhendi erinevaid peatükke). Kui seda ei ole, tuleb tugineda eksperthinnangule.
3. Mõne ohusündmuse puhul on võimalik ka öelda, et ohusündmuse realiseerumisel tekib mõne seaduse või normi rikkumine. Lahtrisse O1.5 tuleb kirja panna sellistest rikkumistest mõned kõige tõsisemad.
4. Kui süsteemil on olemas kaitsemeetmed selle ohustsenaariumi vältimiseks, siis need tuleb kirja panna lahtrisse O1.6. Tähelepanu! Ära siin veel meetmete majandusliku otstarbekuse pärast muretse. Pane kirja, kas ja mida üldse saab teha riski kahandamiseks.

5. Kui kaitsemeetmeid veel ei ole, aga need on hindamise tööühmal kohe võimalik välja pakuda, siis tasub need kirja panna lahtrisse O1.7.
6. Jätka sama AI kasutusloo järgmise ohuga või järgmise AI kasutuslooga, lisades töölehele uusi ohte. Grupeeri sarnased ohud üheks kokku. Välti ohtude arvu kasvamist. Ühe keskmise süsteemi kohta on efektiivsuse säilitamiseks mõistlik käsitleda 10–30 ohustsenaariumit.

Samm 4: korralda ohustsenaariumi hindamine. Täida alamlehel HINNATAVAD OHUD (O) kollased lahtrid O1.8, O1.9, O2.7, O2.8 jne iga tuvastatud ohustsenaariumi kohta, arvestades järgmisi soovitusi.

1. Iga ohu kohta annab iga ohtude hindamisel osaleja mõju ja võimalikkuse hinde, tuginedes ohu kirjeldusele alamlehel HINNATAVAD OHUD (O) ning sellega seotud kasutusloo kirjeldusele vastava kasutusloo alamlehel.
2. Iga hinnatava ohu jaoks soovitage kopeerida alamlehte OHU HINDAMISE PROTOKOLLI NÄIDIS ja selle täita. Alamlehel on käepäraselt näidatud hindamisskaalad ning on antud lahtrid mitme hinde aritmeetilise keskmise ja levinuma hinde arvutamiseks ja hinnete jaotuse kujutamiseks.
3. Tähelepanu! Ühe süsteemi hindamisel peab kasutama ühte ja sama skaalade kirjeldust.
4. Soovitame, et iga hindaja peaks võimalusel kirja panna ka lause või paar enda hinnangu põhjendamiseks.
5. Kui hindajaid on mitu, soovitage esmased hinded anda teineteisest sõltumatult, et vältida olukorda, kus järgmised hindajad on kallutatud esimese hindaja valikust. Kui hinnete salajane kogumine pole võimalik või efektiivne, siis alustatagu iga üksikut hinnete kogumist erinevast isikust, kuid märgime ära, et ka see tekitab riskianalüüsi kallutatust.
6. Kui hindajaid oli mitu, tuleb kokku leppida lõplikus hinded, mis kantakse töölehele ohu hinnangu lahtrisse. Selle metoodika peab valima lähtuvalt skaala iseloomust. Kui skaala on lineaarne, sobib aritmeetiline keskmine. Teiste skaalade puhul tasub kaaluda näiteks kõige populaarsema hinde kasutamist.

5.4 Järeltegevused

Samm 1: vajaduse korral saada hinded kvaliteedikontrolli. Tee järgmist.

1. Kui mõne ohu hindamisel jäi mõju või võimalikkuse hinnete andmisel lähteandmetest puudu ja mõni oht jäi hindamata, saada järelepärimised vastava ohustsenaariumi detailidega kursis olevatele isikutele. Need võivad olla nii andmeteadlased, juristid kui ka turbespetsialistid. Nende teadmised aitavad täpsustada hindamise lähteandmeid pärast mida tuleb korraldada uus hindamine, kus kõigil on samad lähteandmed.

6 Kallutatuse riskide käsitlemine

6.1 Kes, millal ja milleks?

Teostaja. Riskide käsitlemise peaks teostama sama rühm inimesi, kes kirjeldas ja hindas ohustsenaariumid. Neil on riskikäsitlemiseks parim taustateadmus. Kaasata tuleb ka eksperdid, kes oskavad öelda, kuidas saab ohte praktiliselt kahandada. Eriti selles etapis on ülikasulik, kui hindamises osaleb tehisintellektisüsteemi käitava organisatsiooni juht, kellel on volitus võtta organisatsiooni nimel vastu riske. Kui see isik ei saa olla kaasatud riskikäsitlemise, peab ta olema kaasatud otsuse kinnitamisel.

Aeg. Riskikäsitlemine peaks toimuma võimalikult kiiresti pärast riskide hindamist.

Eesmärk. Riskikäsitlemine on tehisintellektisüsteemi kallutatuse hindamise viimane etapp, mille lõpus on olemas otsused.

6.2 Ettevalmistus

Samm 1: otsusta, millised on tähelepanuväärse kallutatuse riski tunnused. Tee järgmist.

1. Kõikide hinnatud riskide jaoks ei tarvitse olla majanduslikult otstarbekas või võimalik leeven-dusmeetmeid rakendada. Seda tasub teha riskide puhul, mis on tähelepanuväärsed.
2. Vaata üle alamlehe RISKIKÄSITLUSE OTSUSED (R) jaotis ABISTAVAD KÜSIMUSED ja seal all olevad tingimused ja lävendid. Vali lävendi näidiste alt enda organisatsiooni jaoks sobivad ning kirjuta need valitud lävendi alla helehalli lahtrisse.
3. Tähelepanu! Kui tehisintellektisüsteemi omaniku organisatsioonis on varasemalt kallutatuse riske analüüsitud (siinse või mõne muu metoodika järgi), siis kaalu tähelepanuväärsuse lävendite taaskasutamist.

Samm 2: selgita välja, millised riskide kahandamise meetmed on võimalikud. Tee järgmist.

1. Selgita välja, millised on süsteemi (tehisintellektikomponendi, treeningandmestiku või üldise teostuse) muutmise või väljavahetamise võimalused ning kas nendeks tegevusteks on olemas rahalised vahendid või muud kokkulepped.
2. Kui süsteemi muutmise ei ole võimalik, siis tuleb selgeks teha, millises ulatuses saab tehisintellektisüsteemi seadistada ning kuidas need muudatused selle väljundite kallutatust võivad mõjutada.
3. Viimaks tasub selgeks teha, millises ulatuses on võimalik muuta tehisintellektisüsteemi väljundite kasutamise protsessi, nt sinna täiendavate kontrollide lisamisega.

6.3 Töölehe täitmine

Samm 1: vali välja tähelepanuväärsed ohustsenaariumid. Täida alamlehel RISKIKÄSITLUSE OTSUSED (R) kollased lahtrid R1.1, R1.2, R2.1, R2.2, R3.1, R3.2 jne iga tähelepanuväärse ohustsenaariumi kohta (lisa või eemalda neid töölehel vastavalt vajadusele). Toetu järgmistele soovitudele.

1. Vaata järjest kõiki kirja pandud ohustsenaariumeid alamlehel HINNATAVAD OHUD (O). Rakenda igaühele tähelepanuväärsuse kriteeriumeid alamlehe RISKIKÄSITLUSE OTSUSED (R) jaotises ABISTAVAD KÜSIMUSED ning kopeeri lävendi ületavate ohtude nimed riskikäsitlemise tabelisse (lahter R1.1).

2. Lisa lahtrisse R1.2 selgitus, millise lävendi oht ületas.

Samm 2: leia kallutatuse ohtudele võimalikud leevendusmeetmed. Täida alamlehel RISKIKÄSITLUSE OTSUSED (R) kollased lahtrid R1.3, R1.4, R2.3, R2.4 jne iga tähelepanuväärse ohustsenaariumi kohta, arvestades järgmiseid soovitusi.

1. Kui kallutatuse riski võimalikud leevendusmeetmed said kirja juba ohu kirjelduse juurde (alamlehele HINNATAVAD OHUD (O)), siis vali nende seast sellised, mis on olemasoleva vahenditega teostatavad või mille jaoks on võimalik vajalikke vahendeid või tööaega tõenäoliselt saada.
2. Tehnilised meetmed võivad nõuda terve süsteemi või selle tehisintellektikomponendi väljavahetamist, masinõppemudeli väljavahetamist või treenimist täiendavate andmetega. Vaata inspiratsiooniks juhendi alajaotisi 4.5 ja 5.4.
3. Organisatoorsed meetmed tähendavad üldiselt täiendava inimliku kontrolli lisamist, aga võivad tähendada ka piiranguid süsteemi kasutajatele (nt riskigruppidel kasutust ei võimaldata) või täiendavate läbipaistvuse meetmete kasutuselevõtt. Vaata lisa juhendi alajaotisest 4.5.

Samm 3: hinda ohustsenaariumi uuesti nagu oleks leevendusmeetmed rakendatud. Täida alamlehel RISKIKÄSITLUSE OTSUSED (R) kollased lahtrid R1.5, R1.6, R1.7, R2.5, R2.6, R2.7 jne iga tähelepanuväärse ohustsenaariumi kohta, arvestades järgmiseid soovitusi.

1. Käitu samamoodi nagu eelmisel hindamisel, aga nüüd kujutle ennast olukorda, kus kallutatuse riski jaoks kavandatud leevendusmeetmed on juba rakendatud.
2. Oluline on mõista, et realistlikud leevendusmeetmed võivad, aga ei pea, hindeid muutma. On võimalik, et näiteks ohu mõju kahaneb, aga võimalikkus jääb samaks. Aga on ka võimalik, et kumbki ei kahane. Järgi samu soovitusi hindamise korralduseks, et hindajate kallutatust kahandada.
3. Pane lahtrisse R1.7 kirja, millises ulatuses jääb oht alles ka siis, kui leevendusmeetmed kasutusele võetakse (kõige tõenäolisem stsenaarium kallutatusega seotud ohu realiseerumiseks, mis ka pärast süsteemi täiendamist alles jääb).
4. Kindlasti kasuta sama skaalade kirjeldust ning mitme hinde koondamise viisi.

Samm 4: tee iga tähelepanuväärse ohu jaoks riskikäsitluse otsus. Täida alamlehel RISKIKÄSITLUSE OTSUSED (R) kollased lahtrid R1.8, R2.8 jne iga tähelepanuväärse ohustsenaariumi kohta, arvestades järgmiseid soovitusi.

1. Otsusta, kas kallutatuse ohu jääkrisk pärast leevendusmeetmete rakendamist on vastuvõetav, arvestades teenuse omaniku riskipoliitikat.
2. Kui risk on vastuvõetav, siis märgi lahtrisse R1.8 "Vastuvõetav" ning selgita, kuidas saab teenuse omanik võimalike kahjudega toime tulla.
3. Kui risk ei ole vastuvõetav, siis märgi lahtrisse R1.8 "Vastuvõetamatu" ning selgita, miks organisatsioon valis sellest riskist keeldumise.

6.4 Järeltegevused

Samm 1: vormista koos tehisintellektisüsteemi omaniku volitatud otsustajaga lõplik kallutatuse riskianalüüsi otsus. Tee järgmist.

1. Leia isik, kelle volituste seas on organisatsiooni nimel riskide võtmine.
2. Tutvusta talle tähelepanuväärseid riske ning nende riskikäsitluse otsuseid.
3. Palu teha otsus süsteemi üldise kasutuselevõtu kohta (võtta kasutusele, võtta kasutusele

muudatustega, eemaldada kasutusest). Dokumenteereri otsus töölehe alamlehel Otsus (S) lahtris S.1.

4. Dokumenteereri töölehe lahtrites S.2–S.6 muudatused, ilma milleta süsteemi töösse anda ei tohi. Saa volitatud isikult kinnitus muudatuste tegemiseks vajalike vahendite kohta.
5. Dokumenteereri töölehe lahtrites S.7–S.9 muudatused, mille tegemine on soovitatav, kuid vajaduse korral edasi lükatav.
6. Märki lahtrisse S.X võimalusel otsuse alusdokument.
7. Lase volitatud isikul otsus kinnitada (nt digitaalse allkirjaga).

Samm 2: vii kallutatuse riskianalüüsi otsus ellu. Tee järgmist.

1. Levita otsust kõigi riskianalüüsi kaasatud isikute seas.
2. Kui tehisintellektisüsteem jääb käiku, aga muudatustega, algata kokku lepitud kohustuslike ja võimalusel ka soovituslike muudatuste läbiviimine selleks ette nähtud vahenditega.
3. Kui tehisintellektisüsteem eemaldatakse kasutuselt, siis alusta sellega seotud tegevusi.