Neatkarīga

AI augsta līmeņa ekspertu grupa,

ko 2018. gada jūnijā izveidojusi Eiropas Komisija

Ētikas vadlīnijas uzticamam AI

ĒTIKAS VADLĪNIJAS UZTICAMAM MĀKSLĪGAJAM INTELEKTAM

AI augsta līmeņa ekspertu grupa

Šo dokumentu izstrādājusi AI augsta līmeņa ekspertu grupa (AI ALEG). Tajā minētie AI ALEG locekļi atbalsta šajās vadlīnijās ieteikto kopējo uzticama mākslīgā intelekta (AI) satvaru, taču var nepiekrist katram atsevišķam dokumentā iekļautam apgalvojumam.

III nodaļā publicēto uzticama AI novērtēšanas sarakstu vispirms izmēģinās ieinteresētās personas, kas sniegs praktiskas atsauksmes. Pārskatīta novērtēšanas saraksta redakcija, kurā būs ņemtas vērā izmēģinājuma posmā saņemtās atsauksmes, tiks iesniegta Eiropas Komisijai 2020. gada sākumā.

AI ALEG ir Eiropas Komisijas 2018. gada jūnijā izveidota neatkarīgu ekspertu grupa.

Kontaktpersona: *Nathalie Smuha* – AI ALEG koordinatore
E-pasts: CNECT-HLG-AI@ec.europa.eu

Eiropas Komisija
B-1049 Brisele

Dokuments publiskots 2019. gada X. aprīlī.

Pirmais dokumenta projekts tika publicēts 2018. gada 18. decembrī, un tika sarīkota tā publiska apspriešana, kurā tika saņemtas atsauksmes no vairāk nekā 500 respondentiem. Vēlamies atklāti un sirsnīgi pateikties ikvienam, kas iesnieguši atsauksmes par dokumenta sākotnējo projektu. Saņemtās atsauksmes tika ievērotas, gatavojot šo pārstrādāto redakciju.

Ne Eiropas Komisija, ne personas, kas rīkojas Komisijas vārdā, neatbild par turpmākās informācijas izmantošanu. Par šā darba dokumenta saturu atbild tikai AI augsta līmeņa ekspertu grupa (AI ALEG). Lai gan Komisijas darbinieki palīdzēja sagatavot šīs vadlīnijas, šajā dokumentā paustie uzskati atspoguļo AI ALEG viedokli un to nekādā gadījumā nevar uzskatīt par Eiropas Komisijas oficiālas nostājas atspoguļojumu.

Sīkākas ziņas par AI augsta līmeņa ekspertu grupu pieejamas tiešsaistē (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).
Eiropas Komisijas dokumentu atkalizmantošanas politiku nosaka Lēmums 2011/833/ES (OV L 330, 14.12.2011., 39. lpp.). Lai varētu izmantot vai reproducēt fotoattēlus vai citus materiālus, uz ko ES nav autortiesību, jāsaņem atļauja tieši no autortiesību subjekta.

SATURA RĀDĪTĀJS

KOPSAVILKUMS **2**

A. IEVADS **4**

B. UZTICAMA AI SATVARS **6**

I nodaļa. Uzticama AI pamati **9**

1. Pamattiesības kā morālas un juridiskas tiesības 9
2. No pamattiesībām līdz ētikas principiem 10

II nodaļa. Uzticama AI īstenošana **14**

1. Prasības uzticamam AI 14
2. Tehniski un netehniski paņēmieni uzticama AI īstenošanai 21

III nodaļa. Uzticama AI novērtēšana **24**

C. AI RADĪTO IESPĒJU UN KRITISKO PROBLĒMJOMU PIEMĒRI **35**

D. NOBEIGUMS **38**

GLOSĀRIJS **40**

KOPSAVILKUMS

(1) Šo vadlīniju mērķis ir veicināt uzticamu AI. Uzticamam AI ir **trīs iezīmes**, kas ir jāievēro visā sistēmas dzīves ciklā: a) tam jābūt **likumīgam** un jāievēro visi piemērojamie normatīvie akti; b) tam jābūt **ētiskam** – jānodrošina ētikas principu un vērtību ievērošana; c) tam jābūt **noturīgam** no tehniskā un sociālā viedokļa, jo AI sistēmas var nodarīt netīšu kaitējumu pat tad, ja nodomi ir bijuši labi. Katra no iezīmēm ir nepieciešama uzticama AI sasniegšanai, bet pati par sevi nav pietiekama. Ideālā gadījumā visas trīs iezīmes darbojas kā saskaņoti komponenti, kuru darbība pārklājas. Ja praksē starp iezīmēm rodas pretrunas, sabiedrībai ir jācenšas tās saskaņot.

(2) Šajās vadlīnijās ir izklāstīts **uzticama AI sasniegšanas satvars**. Satvars tieši neskar pirmo uzticama AI iezīmi (likumīgs AI)[1]. Tā mērķis ir dot norādījumus par ētiska un noturīga AI (otrā un trešā iezīme) veicināšanu un nodrošināšanu. Šīs vadlīnijas, kas ir adresētas visām ieinteresētajām personām, ir paredzētas kā tālejošākas par ētikas principu sarakstu – tās dod norādījumus, kā šādus principus iedzīvināt praksē sociāltehniskajās sistēmās. Norādījumi ir trijos abstrakcijas līmeņos – no visabstraktākā I nodaļā līdz viskonkrētākajam III nodaļā, un nobeigumā doti AI sistēmu sniegto iespēju un kritisko problēmjomu piemēri.

I. Pamatojoties uz pamattiesībās balstītu pieeju, I nodaļā ir noteikti **ētikas principi** un ar tiem saistītās vērtības, kas jāievēro AI sistēmu izstrādē, izvietošanā un izmantošanā.

Galvenie no I nodaļas izrietošie norādījumi

- ✓ Izstrādāt, izvietot un izmantot AI sistēmas tā, lai ievērotu ētikas principus: *cilvēka patstāvības ievērošanu, kaitējuma novēršanu, taisnīgumu un izskaidrojamību.* Atzīt un risināt iespējamās pretrunas šo principu starpā.
- ✓ Pievērst īpašu uzmanību situācijām, kurās ir iesaistītas mazāk aizsargātas grupas, piemēram, bērni, invalīdi un citas personas, kas vēsturiski atradušās nelabvēlīgākā situācijā vai ir pakļautas atstumšanas riskam, kā arī situācijām, kuras raksturo varas vai informācijas asimetrija, piemēram, attiecībās starp darba devēju un darbinieku vai uzņēmumu un patērētāju[2].
- ✓ Atzīt un ņemt vērā, ka AI sistēmas gan sniedz būtiskus ieguvumus indivīdiem un sabiedrībai, taču arī rada zināmu risku un tām var būt negatīva ietekme, ieskaitot grūti paredzamu, nosakāmu un izmērāmu ietekmi (piemēram, uz demokrātiju, tiesiskumu un taisnīgu sadali, kā arī uz pašu cilvēka prātu). Attiecīgos gadījumos veikt pienācīgus pasākumus šā riska mazināšanai samērā ar riska apmēru.

II. Turpinot I nodaļu, II nodaļā ir doti norādījumi par to, kā var realizēt uzticamu AI, un norādītas **septiņas prasības** AI sistēmām. To īstenošanā var izmantot gan tehniskus, gan netehniskus paņēmienus.

Galvenie no II nodaļas izrietošie norādījumi

- ✓ Nodrošināt AI sistēmu izstrādes, izvietošanas un lietošanas atbilstību uzticama AI prasībām, kuras ir šādas: 1) cilvēka subjektība un virsvadība; 2) tehniskā noturība un drošums; 3) privātums un datu pārvaldīšana; 4) pārredzamība; 5) daudzveidība, nediskriminēšana un taisnīgums; 6) vides un sabiedrības labklājība; 7) atbildīgums.
- ✓ Apsvērt tehniskus un netehniskus paņēmienus, ar ko nodrošināt šo prasību īstenošanu.

---

[1] Visu šajā dokumentā ietverto normatīvo apgalvojumu mērķis ir atspoguļot norādījumus, kā panākt uzticama AI otro un trešo iezīmi (ētisku un noturīgu AI). Tādēļ šie apgalvojumi nav paredzēti kā juridiski ieteikumi vai vadlīnijas par to, kā nodrošināt atbilstību piemērojamajām tiesību normām, tomēr tiek atzīts, ka daudzi no šiem apgalvojumiem zināmā mērā jau ir atspoguļoti esošajos tiesību aktos. Šajā sakarā sk. 21. un turpmākos punktus.

[2] Sk. Eiropas Savienības Pamattiesību hartas (ES harta) 24.–27. pantu, kuros minētas bērnu un vecu cilvēku tiesības, invalīdu integrācija un darba ņēmēju tiesības. Sk. arī 38. pantu, kas attiecas uz patērētāju tiesību aizsardzību.

- ✓ Stimulēt pētniecību un inovāciju, kas palīdzētu izvērtēt AI sistēmas un veicinātu prasību izpildi; ar rezultātiem un neatbildētajiem jautājumiem iepazīstināt plašāku sabiedrību, sistēmiski gatavot jaunu mākslīgā intelekta ētikas ekspertu paaudzi.
- ✓ Skaidri un proaktīvi ieinteresētajām personām darīt zināmu informāciju par AI sistēmu spējām un robežām, kas ļautu noteikt reāli īstenojamo, kā arī informāciju par prasību īstenošanas veidu. Skaidri norādīt, ka notiek saskarsme ar AI sistēmu.
- ✓ Atvieglot AI sistēmu trasējamību un pārbaudāmību, it īpaši kritiskos apstākļos vai situācijās.
- ✓ Iesaistīt ieinteresētās personas visā AI sistēmas dzīves ciklā. Veicināt apmācību un izglītību, lai visas ieinteresētās personas būtu informētas un apmācītas attiecībā uz uzticamu AI.
- ✓ Apzināties, ka starp dažādiem principiem un prasībām pašos pamatos var pastāvēt pretrunas. Pastāvīgi identificēt, izvērtēt un dokumentēt šos kompromisus un to risinājumus un informēt par tiem.

III. III nodaļā dots konkrēts un neizsmeļošs uzticama AI novērtēšanas saraksts, kas ir vērsts uz II nodaļā izklāstīto prasību iedzīvināšanu. Šis **novērtēšanas saraksts** ir jāpielāgo AI sistēmas konkrētajam lietojumam[3].

Galvenie no III nodaļas izrietošie norādījumi

- ✓ Izstrādājot, izvietojot vai izmantojot AI sistēmas, veikt uzticama AI novērtēšanu un pielāgot to konkrētajam sistēmas lietošanas gadījumam.
- ✓ Ņemt vērā, ka šāds novērtēšanas saraksts nekad nebūs izsmeļošs. Uzticama AI nodrošināšana nav ķeksīšu ievilkšana, bet gan pastāvīga prasību identificēšana un īstenošana, risinājumu izvērtēšana un uzlabotu iznākumu nodrošināšana visā AI sistēmas dzīves ciklā, kā arī ieinteresēto personu iesaiste šajās darbībās.

(3) Pēdējās dokumenta sadaļas mērķis ir konkretizēt dažas satvarā skartās problēmas, dodot piemērus par izmantojamām izdevīgām iespējām un par AI sistēmu kritiskajām problēmjomām, kas rūpīgi jāizvērtē.

(4) Lai gan šo vadlīniju mērķis ir dot norādījumus par AI lietojumiem kopumā, tā veidojot horizontālu pamatu uzticama AI sasniegšanai, atšķirīgās situācijās rodas atšķirīgas problēmas. Tādēļ ir jāizvērtē, vai papildus šim horizontālajam satvaram nepieciešama nozaru pieeja, ņemot vērā AI sistēmu atkarību no konteksta.

(5) Šo vadlīniju mērķis nav ne aizstāt esošo vai nākotnes politiku vai noteikumus, ne atturēt no to ieviešanas. Tās ir jāuzskata par dinamisku dokumentu, kas laika gaitā jāpārskata un jāatjaunina, lai nodrošinātu, ka tas nezaudē nozīmību, attīstoties tehnoloģijām, sociālajai videi un zināšanām. Šis dokuments ir radīts kā sākumpunkts diskusijai par uzticamu mākslīgo intelektu Eiropai[4]. Ārpus Eiropas šo vadlīniju mērķis ir veicināt pētniecību, apceri un diskusijas par ētisku regulējumu AI sistēmu jomā pasaules mērogā.

[3] Atbilstīgi 2. punktā izklāstītā satvara tvērumam šajā novērtēšanas sarakstā nav doti ieteikumi, kā nodrošināt juridisko atbilstību (likumīgs AI), bet vienīgi doti norādījumi, kā ievērot otro un trešo uzticama AI iezīmi (ētisks un noturīgs AI).

[4] Šo ideālu ir paredzēts attiecināt uz AI sistēmām, ko izstrādā, izvieto un izmanto ES dalībvalstīs, kā arī sistēmām, kas izstrādātas vai ražotas citur, bet tiek izvietotas un izmantotas ES. Atsauces uz "Eiropu" šajā dokumentā attiecas uz visām ES dalībvalstīm. Tomēr šīm vadlīnijām būtu jābūt nozīmīgām arī ārpus ES. Šajā ziņā ir jānorāda, ka gan Norvēģija, gan Šveice piedalās koordinētajā AI plānā, par kuru 2018. gada decembrī vienojās un kuru publicēja Komisija un dalībvalstis.

A. IEVADS

(6) Eiropas Komisija (turpmāk "Komisija") 2018. gada 25. aprīļa paziņojumā un 2018. gada 7. decembra paziņojumā izklāstīja savu ieceri par mākslīgo intelektu (AI), paužot atbalstu "ētiska, droša un progresīva mākslīgā intelekta izstrādei Eiropā"[5]. Komisijas iecere balstās uz trim pīlāriem: i) palielināt publiskos un privātos ieguldījumus AI, lai palielinātu tā izmantošanu; ii) gatavoties sociālekonomiskajām pārmaiņām; iii) nodrošināt atbilstīgu ētisko un juridisko satvaru Eiropas vērtību stiprināšanai.

(7) Lai atbalstītu šā ieceres īstenošanu, Komisija izveidoja AI augsta līmeņa ekspertu grupu (AI ALEG) – neatkarīgu grupu, kurai tika uzdots izstrādāt divus dokumentus: 1) AI ētikas vadlīnijas, 2) ieteikumus par politiku un ieguldījumiem.

(8) Šajā dokumentā ir ietvertas AI ētikas vadlīnijas, kas pēc tālākas apspriešanās mūsu grupā ir pārskatītas, ņemot vērā atsauksmes, kuras tika saņemtas pēc 2018. gada 18. decembrī publicētā projekta publiskās apspriešanas. Tās ir Eiropas Dabaszinātņu un jauno tehnoloģiju ētikas grupas veiktā darba[6] turpinājums, un to izstrādē iedvesma gūta arī no citiem līdzīgiem centieniem[7].

(9) Pēdējo mēnešu laikā mēs 52 cilvēku sastāvā tikāmies, apspriedāmies un kontaktējāmies, apņēmušies ievērot Eiropas devīzi: "Vienoti dažādībā". Mēs esam pārliecināti, ka AI var būtiski pārveidot sabiedrību. AI nav pašmērķis, bet gan daudzsološs veids, kā veicināt cilvēces uzplaukumu, tādējādi vairojot indivīdu un sabiedrības labklājību un kopējo labumu, kā arī radot progresu un inovāciju. AI sistēmas var īpaši palīdzēt sasniegt ANO ilgtspējīgas attīstības mērķus, piemēram, veicināt dzimumu līdzsvaru un cīņu pret klimata pārmaiņām, racionalizēt dabas resursu izmantošanu, uzlabot mūsu veselību, mobilitāti un ražošanas procesus, kā arī atbalstīt uzraudzību attiecībā uz ilgtspējas un sociālās kohēzijas rādītāju progresu.

(10) Lai to panāktu, AI sistēmām[8] jābūt **antropocentriskām** un balstītām uz apņemšanos tās izmantot cilvēces un koplabuma interesēs ar mērķi uzlabot cilvēku labklājību un brīvību. Lai gan AI sistēmas sniedz lieliskas iespējas, tās arī rada noteiktu risku, attiecībā uz kuru jārīkojas pienācīgi un samērīgi. Tagad mums ir pavērusies svarīga iespēja ietekmēt to attīstību. Mēs vēlamies panākt, lai mēs varētu uzticēties sociāltehniskajai videi, kurā ir iekļautas AI sistēmas, un mēs vēlamies, lai AI sistēmu ražotāji gūtu priekšrocību konkurences cīņā, iekļaujot uzticamu AI savās precēs un pakalpojumos. Šim nolūkam ir jācenšas **maksimāli palielināt ieguvumus no AI sistēmām**, vienlaikus **novēršot un mazinot to radīto risku**.

(11) Mēs uzskatām, ka strauju tehnoloģisko pārmaiņu kontekstā ir ārkārtīgi svarīgi, lai sabiedrību, kopienu, ekonomikas un ilgtspējīgas attīstības pamats arī turpmāk ir uzticēšanās. Tādēļ mēs **par pamatmērķi nosakām uzticamu AI**, jo cilvēki un kopienas varēs uzticēties tehnoloģijas attīstībai un tās lietojumiem tikai tad, kad būs ieviests skaidrs un visaptverošs satvars tās uzticamības panākšanai.

(12) Mūsuprāt, Eiropai ir jāizvēlas šis ceļš, lai pozicionētu sevi kā augsto un ētisko tehnoloģiju mājvietu un līderi. Mēs, Eiropas pilsoņi, izmantojot uzticamu AI, centīsimies gūt no tā labumu tādā veidā, kas atbilst mūsu pamatvērtībām – cilvēktiesību, demokrātijas un tiesiskuma ievērošanai.

Uzticams AI

(13) Uzticamība ir priekšnoteikums, kas jāizpilda, lai cilvēki un sabiedrība izstrādātu, izvietotu un izmantotu AI sistēmas. Ja netiks acīmredzami apliecināta AI sistēmu un aiz tām stāvošo cilvēku uzticamība, var iestāties

---

[5] COM(2018) 237 un COM(2018) 795. Komisijas paziņojumā konsekventi lietots termins "Eiropā radīts AI". Taču šo vadlīniju mērķis ir aptvert ne tikai Eiropā izstrādātas AI sistēmas, bet arī sistēmas, kuras izstrādātas citur un kuras izvieto vai izmanto Eiropā. Tādēļ šajā dokumentā mūsu nolūks ir popularizēt "uzticamu AI Eiropai".

[6] Eiropas Dabaszinātņu un jauno tehnoloģiju ētikas grupa (*EGE*) ir Komisijas padomdevēja grupa.

[7] Sk. COM(2018) 237 3.3. iedaļu.

[8] Šā dokumenta beigās pievienotajā glosārijā ir sniegta AI sistēmu definīcija šā dokumenta vajadzībām. Definīcija sīkāk izstrādāta atsevišķā dokumentā, kuru sagatavojusi AI ALEG un kurš pievienots šīm vadlīnijām ar nosaukumu "AI definīcija: galvenās spējas un zinātnes disciplīnas".

nevēlamas sekas un to izmantošana var tikt kavēta, tādējādi liedzot realizēt varbūtējos plašos sociālos un ekonomiskos ieguvumus, ko rada AI sistēmas. Lai palīdzētu Eiropai sasniegt šos ieguvumus, mūsu skatījumā ētika ir jāizmanto kā pamatpīlārs uzticama AI nodrošināšanai un izvēršanai.

(14) Uzticēšanās AI sistēmu izstrādei, izvietošanai un izmantošanai ir atkarīga ne tikai no tehnoloģijai piemītošajām īpašībām, bet arī no to sociāltehnisko sistēmu īpašībām, kuras ir saistītas ar AI lietojumiem[9]. Līdzīgi kā jautājumos par uzticēšanos aviācijai, kodolenerģijai vai pārtikas nekaitīgumam (vai uzticības zaudēšanu), uzticēšanos rada nevis tikai AI sistēmas sastāvdaļas, bet gan sistēma tās kopējā kontekstā. Tādēļ tiekšanās uz uzticamu AI ne vien attiecas uz pašas AI sistēmas uzticamību, bet arī prasa holistisku un sistēmisku pieeju, kas aptver visu sistēmas sociāltehniskajā kontekstā tās dzīves ciklā iesaistīto personu un procesu uzticamību.

(15) Uzticamam AI ir **trīs iezīmes**, kas ir jāievēro visā sistēmas dzīves ciklā:

1. tam jābūt **likumīgam** – jānodrošina, ka tiek ievēroti visi piemērojamie normatīvie akti;
2. tam jābūt **ētiskam** – jānodrošina atbilstība ētikas principiem un vērtībām; un
3. tam jābūt **noturīgam** gan no tehniskā, gan sociālā viedokļa, jo AI sistēmas var nodarīt negribētu kaitējumu pat tad, ja nodomi ir bijuši labi.

(16) Katra no šīm trim iezīmēm ir nepieciešama, bet pati par sevi nav pietiekama uzticama AI sasniegšanai[10]. Ideālā gadījumā visas trīs sadarbojas saskaņoti un to darbība pārklājas. Tomēr praksē šo elementu starpā var pastāvēt pretrunas (piemēram, dažos gadījumos esošo tiesību aktu darbības joma un saturs var nesakrist ar ētikas normām). Mūsu kā sabiedrības atsevišķā un kolektīvā atbildība ir strādāt, lai visas trīs iezīmes palīdzētu nodrošināt uzticamu AI[11].

(17) Pieejai, kas balstās uz uzticēšanos, ir liela nozīme "atbildīgas konkurētspējas" iespēju radīšanā, jo tā nodrošina pamatu tam, ka visas AI sistēmu skartās personas paļaujas, ka sistēmu izstrāde, attīstīšana un izmantošana ir likumīga, ētiska un noturīga. Šo vadlīniju mērķis ir veicināt atbildīgu un ilgtspējīgu AI inovāciju Eiropā. To nolūks ir panākt, ka ētika ir pamatpīlārs, kad tiek izstrādāta tāda unikāla pieeja AI, kuras mērķis ir radīt ieguvumus, sniegt iespējas un aizsargāt gan atsevišķu cilvēku izaugsmi, gan sabiedrības kopējo labumu. Mēs uzskatām, ka tas ļaus Eiropai uzstāties kā pasaules līderei augsto tehnoloģiju AI, kas pelna mūsu individuālu un kolektīvu uzticēšanos. Tikai tad, ja tiks nodrošināta uzticamība, eiropieši varēs pilnībā izmantot AI sistēmu sniegtos ieguvumus, droši apzinoties, ka ir izveidoti pasākumi aizsardzībai pret varbūtēju risku.

(18) Tā kā AI sistēmu izmantošana neapstājas pie valstu robežām, tur neapstājas arī to ietekme. Tādēļ AI radītajām pasaules mēroga iespējām un problēmām ir nepieciešami pasaules mēroga risinājumi. Līdz ar to mēs mudinām visas ieinteresētās personas strādāt pie pasaules mēroga regulējuma uzticamam AI, veidojot starptautisko vienprātību un vienlaikus popularizējot un ievērojot mūsu pieeju, kas ir balstīta pamattiesībās.

Adresāti un darbības joma

(19) Šīs vadlīnijas ir adresētas visām AI jomas ieinteresētajām personām, kuras projektē, izstrādā, izvieto, īsteno, izmanto AI vai kuras tas skar, tostarp (bet ne tikai) uzņēmumiem, organizācijām, pētniekiem, sabiedrisko pakalpojumu sniedzējiem, valsts aģentūrām, iestādēm, pilsoniskās sabiedrības organizācijām, indivīdiem, darba ņēmējiem un patērētājiem. Ieinteresētās personas, kas ir apņēmušās panākt AI uzticamību, var brīvprātīgi izvēlēties izmantot šīs vadlīnijas savas apņemšanās iedzīvināšanai praksē, jo īpaši – izmantojot III nodaļā sniegto praktisko novērtēšanas sarakstu AI sistēmu izstrādes un izvietošanas procesos. Šis novērtēšanas saraksts var arī papildināt esošos vērtēšanas procesus un tādējādi tikt iekļauts tajos.

---

[9] Šajās sistēmās ietilpst cilvēki, valsts struktūras, uzņēmumi, infrastruktūra, programmatūra, protokoli, standarti, pārvaldība, esošie tiesību akti, pārraudzības mehānismi, stimulu struktūras, revīzijas procedūras, ziņošana par labāko pieredzi u. c.

[10] Tas neizslēdz, ka var būt (vai kļūt) nepieciešami papildu nosacījumi.

[11] Tas nozīmē arī to, ka likumdevējiem vai politikas veidotājiem var būt nepieciešams pārskatīt esošo tiesību aktu piemērotību, ja tie, iespējams, nesakrīt ar ētikas principiem.

(20) Šo vadlīniju mērķis ir sniegt norādījumus par AI lietojumiem kopumā, tā veidojot horizontālu pamatu uzticama AI sasniegšanai. Tomēr **dažādās situācijās rodas atšķirīgas problēmas**. AI sistēmas, kas iesaka mūziku, nerada tādas pašas ētiskas problēmas kā AI sistēmas, kas iesaka svarīgu medicīnisko aprūpi. Tāpat atšķirīgas iespējas un problēmas ir saistītas ar AI sistēmām, ko izmanto sakarā ar uzņēmumu darījumiem ar patērētājiem, uzņēmumu darījumiem ar uzņēmumiem, darba devēju un darbinieku, kā arī sabiedrības un indivīda attiecībām, vai vispārīgi – dažādās nozarēs vai lietojumos. Tādēļ, ņemot vērā AI sistēmu atkarību no konteksta, tiek atzīts, ka šo vadlīniju īstenošana ir jāpielāgo konkrētajam AI lietojumam. Turklāt būtu jāizvērtē nepieciešamība pēc papildu nozaru pieejas, kas papildinātu šajā dokumentā ierosināto vispārīgāko horizontālo satvaru.

Lai labāk izprastu iespējas šīs vadlīnijas īstenot horizontālā līmenī un jautājumus, kam nepieciešama nozaru pieeja, aicinām visas ieinteresētās personas izmēģināt uzticama AI novērtēšanas sarakstu (III nodaļa), kas paredzēts šā satvara praktiskai īstenošanai, un sniegt mums atsauksmes. Balstoties uz šajā izmēģinājuma posmā iegūtajām atsauksmēm, mēs līdz 2020. gada sākumam pārskatīsim šajās vadlīnijās iekļauto novērtēšanas sarakstu. Izmēģinājuma posms sāksies 2019. gada vasarā un turpināsies līdz gada beigām. Visas ieinteresētās personas varēs piedalīties, izsakot interesi ar Eiropas AI alianses starpniecību.

B. UZTICAMA AI SATVARS

(21) Šajās vadlīnijās ir formulēts uzticama AI sasniegšanas satvars, kas balstās Eiropas Savienības Pamattiesību hartā (ES harta) un attiecīgajās starptautisko cilvēktiesību normās nostiprinātajās pamattiesībās. Tālāk īsi aprakstītas uzticama AI trīs iezīmes.

Likumīgs AI

(22) AI sistēmas nedarbojas bezlikumu pasaulē. Jau šobrīd Eiropas, valstu un starptautiskā līmenī ir vairāki juridiski saistoši noteikumi, ko piemēro vai kas attiecas uz AI sistēmu izstrādi, izvietošanu un izmantošanu. Šajā jomā cita starpā ir jāņem vērā tādi tiesību avoti kā ES primārās tiesību normas (Eiropas Savienības līgumi un tās Pamattiesību harta), ES sekundārās tiesību normas (piemēram, Vispārīgā datu aizsardzības regula, diskriminācijas novēršanas direktīvas, Mašīnu direktīva, Produktatbildības direktīva, Nepersondatu brīvas aprites regula, patērētāju tiesības un darba drošības un arodveselības direktīvas), kā arī ANO cilvēktiesību līgumi un Eiropas Padomes konvencijas (piemēram, Eiropas Cilvēktiesību konvencija), un dažādi ES dalībvalstu tiesību akti. Papildus horizontāli piemērojamajiem noteikumiem pastāv dažādi nozarēm paredzēti noteikumi, ko piemēro konkrētiem AI lietojumiem (piemēram, veselības aprūpes jomā – Medicīnisko ierīču regula).

(23) Tiesību akti paredz gan pozitīvos pienākumus, gan pienākumus atturēties no rīcības, kas nozīmē, ka juridiski ir ne tikai jāsaprot, kas ir *aizliegts*, bet arī jāatsaucas uz to, kā ir *jārīkojas*. Tiesību akti ne tikai aizliedz konkrētas darbības, bet arī ļauj veikt citas. Te jānorāda, ka ES harta satur pantus par uzņēmējdarbības brīvību un humanitāro un eksakto zinātņu brīvību līdztekus pantiem, kas attiecas uz jomām, par kurām AI uzticamības nodrošināšanas kontekstā esam labāk informēti, piemēram, datu aizsardzību un nediskriminēšanu.

(24) Vadlīnijas tieši nepievēršas pirmajai uzticama AI iezīmei (likumīgam AI), bet to mērķis ir dod norādījumus par to, kā veicināt un nodrošināt otro un trešo iezīmi (ētisku un noturīgu AI). Lai gan abi pēdējie aspekti savā ziņā jau bieži ir atspoguļoti esošajos tiesību aktos, to pilnīga īstenošana varētu prasīt vairāk par esošajiem juridiskajiem pienākumiem.

(25) Nekas šajā dokumentā nav saprotams vai interpretējams kā juridisku ieteikumu vai norādījumu sniegšana attiecībā uz to, kā panākt atbilstību piemērojamajām spēkā esošajām tiesību normām un prasībām. Nekas šajā dokumentā nerada juridiskas tiesības un neparedz juridiskus pienākumus trešām personām. Tomēr mēs atgādinām, ka ikvienai fiziskai un juridiskai personai ir pienākums ievērot gan spēkā esošos tiesību aktus, gan tiesību aktus, kas tiks pieņemti nākotnē atbilstoši AI attīstībai. Šajās vadlīnijās tiek pieņemts, ka **ir saistošas un ir pilnīgi jāievēro visas juridiskās tiesības un pienākumi, kas attiecas uz AI izstrādes, izvietošanas un izmantošanas procesiem un darbībām.**

Ētisks AI

(26) Uzticama AI sasniegšana neprasa tikai ievērot tiesību aktus, jo likumība ir tikai viena no trim tā iezīmēm. Tiesību akti dažkārt atpaliek no tehnoloģiju attīstības vai neatbilst ētikas normām, vai vienkārši nav īsti piemēroti konkrētu problēmu risināšanai. Tādēļ, lai AI sistēmas būtu uzticamas, tām jābūt arī ētiskām, proti, jānodrošina saskaņotība ar ētikas normām.

Noturīgs AI

(27) Pat ja ir nodrošināts ētisks mērķis, indivīdiem un sabiedrībai arī jāgūst pārliecība, ka AI sistēmas neizraisīs netīšu kaitējumu. Šādām sistēmām ir jādarbojas droši, aizsargāti un neatteicīgi, un ir jāparedz aizsardzības pasākumi, kas novērstu netīšu nelabvēlīgu ietekmi. Tādēļ ir svarīgi nodrošināt, lai AI sistēmas būtu noturīgas. Tas ir nepieciešams gan no tehniskā viedokļa (sistēmas tehniskās noturības nodrošināšana, kā nepieciešams attiecīgā kontekstā, piemēram, lietojuma jomā vai dzīves cikla posmā), gan no sociālā viedokļa (pienācīgi ņemot vērā sistēmas darbības kontekstu un vidi). Līdz ar to ētisks un noturīgs AI ir cieši saistīti aspekti, kas viens otru papildina. I nodaļā izklāstītie principi un no tiem II nodaļā atvasinātās prasības skar abas minētās iezīmes.

Satvars

(28) Šajā dokumentā iekļautie norādījumi tiek sniegti trijos abstrakcijas līmeņos – no visabstraktākā I nodaļā līdz viskonkrētākajam III nodaļā:

I) Uzticama AI pamati. I nodaļā ir aprakstīti uzticama AI pamati, izklāstot tā pamatā esošo pamattiesībās[12] balstīto pieeju. Tajā ir norādīti un aprakstīti ētikas principi, kas jāievēro, lai nodrošinātu ētisku un noturīgu AI.

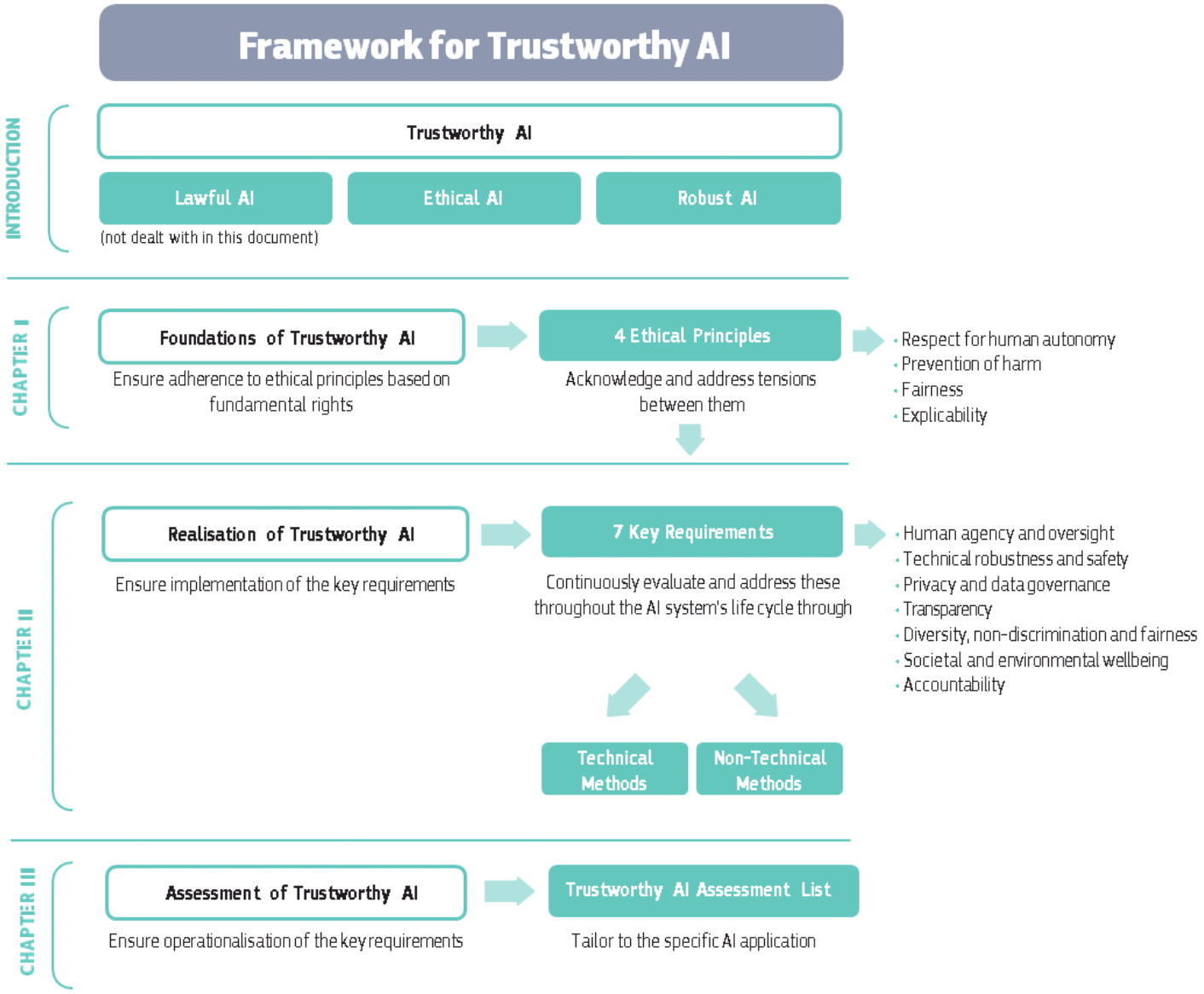
II) Uzticama AI īstenošana. II nodaļā iepriekš minētie ētikas principi ir izteikti septiņās prasībās, kuras jāīsteno AI sistēmās un kuras tām jāievēro visā dzīves ciklā. Tajā arī piedāvāti gan tehniski, gan netehniski paņēmieni šo prasību izpildei.

III) Uzticama AI novērtēšana. AI jomas praktiķi gaida konkrētus norādījumus. Tādēļ III nodaļā ir sniegts provizorisks neizsmeļošs uzticama AI novērtēšanas saraksts, kas ļauj praktiski piemērot II nodaļas prasības. Šis novērtējums ir jāpielāgo konkrētās sistēmas lietojumam.

(29) Dokumenta pēdējā sadaļā ir sniegti AI sistēmu radītu izdevīgu iespēju un kritisku problēmjomu piemēri, kuri jāapsver un par kuriem mēs vēlamies rosināt tālākas diskusijas.

(30) Vadlīniju struktūra ir atspoguļota turpmāk *1. attēlā*.

[12] Pamattiesības ir gan starptautisko, gan ES cilvēktiesību centrā un ir pamats tiesībām, ko garantē ES līgumi un ES Pamattiesību harta un kam ir juridisks spēks. Juridiski saistošā rakstura dēļ pamattiesību ievērošana ir pakļauta uzticama AI pirmajai iezīmei "Likumīgs AI". Tomēr pamattiesības var saprast arī kā visu indivīdu tādu konkrētu morālu tiesību atspoguļojumu, ko tiem dod piederība cilvēcei, neatkarīgi no šo tiesību juridiski saistošā statusa. Šādā ziņā tās ir daļa arī no uzticama AI otrās iezīmes "Ētisks AI".



1. attēls. Vadlīnijas kā uzticama AI satvars

I. I nodaļa. Uzticama AI pamati

(31) Šajā nodaļā ir izklāstīti uzticama AI pamati, kas balstās pamattiesībās un ko atspoguļo četri ētikas principi, kuri jāievēro, lai nodrošinātu ētisku un noturīgu AI. Šī nodaļa lielā mērā balstās uz ētikas nozari.

(32) AI ētika ir lietišķās ētikas apakšnozare, kas pievēršas AI izstrādes, izvietošanas un izmantošanas radītajām ētiskajām problēmām. Tās galvenais uzdevums ir noteikt, kā AI var uzlabot indivīdu dzīves līmeni vai radīt bažas par to no dzīves kvalitātes vai demokrātiskā sabiedrībā nepieciešamās cilvēku patstāvības un brīvības viedokļa.

(33) Ētiskas pārdomas par AI tehnoloģiju var kalpot vairākiem mērķiem. Pirmkārt, tās var stimulēt pārdomas par nepieciešamību aizsargāt indivīdus un grupas pašā zemākajā līmenī. Otrkārt, tās var stimulēt jaunu inovāciju, kas cenšas veicināt ētiskas vērtības, piemēram, tādas, kas palīdzētu sasniegt ANO ilgtspējīgas attīstības mērķus[13], kuri ir cieši iekļauti gaidāmajā ES ilgtspējīgas attīstības programmā 2030. gadam[14]. Lai gan šis dokuments lielākoties attiecas uz pirmo minēto mērķi, nedrīkst nenovērtēt ietekmi, kāda ētikai varētu būt uz otro. Uzticams AI var vairot indivīdu uzplaukumu un sabiedrības labklājību, radot pārticību, veidojot vērtību un palielinot bagātību. Tas var palīdzēt izveidot taisnīgu sabiedrību, palīdzot uzlabot iedzīvotāju veselību un labklājību tā, lai tiktu veicināta ekonomisko, sociālo un politisko iespēju vienlīdzīga sadale.

(34) Tādēļ ir ļoti svarīgi, lai mēs saprastu, kā vislabāk atbalstīt AI izstrādi, izvietošanu un izmantošanu, lai nodrošinātu ikvienam sekmīgu attīstību uz AI balstītā pasaulē un radītu labāku nākotni, vienlaikus esot konkurētspējīgi pasaules mērogā. Kā jebkuras jaudīgas tehnoloģijas gadījumā, AI sistēmu izmantošana mūsu sabiedrībā rada vairākus ētikas jautājumus, piemēram, saistībā ar to ietekmi uz cilvēkiem un sabiedrību, lēmumu pieņemšanas iespējām un drošību. Ja mēs arvien vairāk izmantosim AI sistēmu palīdzību vai uzticēsim tām lēmumu pieņemšanu, mums ir jāpārliecinās, ka šo sistēmu ietekme uz cilvēku dzīvi ir taisnīga, tās atbilst neapspriežamām vērtībām un spēj attiecīgi rīkoties, un ka to nodrošina piemēroti atbildīguma procesi.

(35) Eiropai ir jānosaka, kādu normatīvu ieceri par nākotni ar plašu AI izplatību tā vēlas īstenot, un attiecīgi jāsaprot, kāds AI būtu jāpēta, jāizstrādā, jāizvērš un jāizmanto Eiropā šās ieceres sasniegšanai. Ar šo dokumentu mēs vēlamies dot ieguldījumu šajos centienos, ieviešot uzticama AI jēdzienu, kas, mūsuprāt, ir pareizais veids, kā veidot nākotni ar AI. Nākotne, kurā AI sistēmu pamatā ir demokrātija, tiesiskums un pamattiesības un kur šīs sistēmas pastāvīgi uzlabo un aizsargā demokrātijas kultūru, arī ļaus veidot vidi, kurā var plaukt inovācija un atbildīga konkurētspēja.

(36) Nozares ētikas kodekss, lai cik konsekventas, attīstītas un precizētas būtu tā nākotnes redakcijas, nekādā gadījumā nevar aizstāt ētisku spriešanu kā tādu, jo tajā vienmēr jāpatur prātā konteksta nianses, kuras nevar ietvert vispārīgās vadlīnijās. Lai tiktu nodrošināts uzticams AI, papildus noteikumu izstrādei mums ir arī jāveido un jāuztur ētiska kultūra un domāšanas veids. To var panākt ar publiskām diskusijām, izglītošanu un praktisku apmācību.

1. Pamattiesības kā morālas un juridiskas tiesības

(37) Mēs ticam tādai pieejai AI ētikai, kuras pamatā ir pamattiesības, kas nostiprinātas ES līgumos[15], ES Pamattiesību hartā (ES harta) un starptautisko cilvēktiesību normās[16]. Pamattiesību ievērošana demokrātijas un tiesiskuma satvarā ir visdrošākais pamats to abstrakto ētikas principu un vērtību noteikšanai, ko var iedzīvināt praksē AI kontekstā.

---

13 https://ec.europa.eu/commission/publications/reflection-paper-towards-sustainable-europe-2030_lv.

14 https://sustainabledevelopment.un.org/?menu=1300.

15 ES balstās uz konstitucionālu apņemšanos aizsargāt cilvēku pamattiesības un nedalāmās tiesības, lai nodrošinātu tiesiskumu, veicinātu demokrātisko brīvību un atbalstītu kopējo labumu. Šīs tiesības ir atspoguļotas Līguma par Eiropas Savienību 2. un 3. pantā, kā arī ES Pamattiesību hartā.

16 Šīs saistības atspoguļo un sīkāk nosaka citi tiesību instrumenti, piemēram, Eiropas Padomes Eiropas Sociālā harta, vai konkrēti tiesību akti, piemēram, ES Vispārīgā datu aizsardzības regula.

(38) ES līgumos un ES Pamattiesību hartā ir paredzēta virkne pamattiesību, kuras ES dalībvalstīm un ES iestādēm ir juridisks pienākums ievērot, īstenojot ES tiesību aktus. Šīs tiesības ES hartā ir aprakstītas, atsaucoties uz cieņu, brīvībām, vienlīdzību un solidaritāti, pilsoņu tiesībām un tiesiskumu. Šīs tiesības vieno kopīgs pamats, ko var saprast kā cilvēka cieņas ievērošanu un kas tādējādi atspoguļo "antropocentrisku pieeju", kurā cilvēkam ir unikāls un neatņemams morālas prioritātes statuss pilsoniskajā, politiskajā, saimnieciskajā un sociālajā jomā[17].

(39) Lai gan ES hartā izklāstītās tiesības ir juridiski saistošas[18], ir svarīgi apzināties, ka pamattiesības ne vienmēr nodrošina visaptverošu juridisku aizsardzību. Piemēram, ir svarīgi uzsvērt, ka ES Pamattiesību hartas piemērošanas joma aptver tikai ES tiesību jomas. Starptautiskās cilvēktiesības un sevišķi Eiropas Cilvēktiesību konvencija juridiski saista ES dalībvalstis, arī jomās, ko neskar ES tiesību normas. Vienlaikus jāuzsver, ka pamattiesības indivīdiem un (ierobežotā apmērā) grupām piemīt arī, pamatojoties uz to morālo statusu kā cilvēciskām būtnēm, neatkarīgi no juridiskā spēka. Tādēļ pamattiesības, ja tās izprot kā tiesības ar juridisku spēku, ir attiecināmas uz pirmo uzticama AI iezīmi (likumīgs AI), kas aizsargā tiesību aktu ievērošanu. Ja tās saprot kā ikviena tiesības, kas balstās cilvēkiem iedzimtajā morālajā statusā, tās ir pamatā arī otrajai uzticama AI iezīmei (ētiskam AI) un skar ētikas normas, kas var nebūt juridiski saistošas, taču ir ļoti svarīgas uzticēšanās nodrošināšanai. Tā kā šā dokumenta mērķis nav dot norādījumus par pirmo iezīmi, nesaistošajās vadlīnijās atsauces uz pamattiesībām ir saistītas ar otro iezīmi.

2. No pamattiesībām līdz ētikas principiem

2.1 Pamattiesības kā uzticama AI pamats

(40) No visaptverošā nedalāmo tiesību kopuma, kas noteikts starptautiskajās cilvēktiesībās, ES līgumos un ES hartā, uz AI sistēmām īpaši attiecas tālāk minētās pamattiesību grupas. Daudzām no šīm tiesībām konkrētos apstākļos ES ir juridisks spēks, tādēļ to noteikumu ievērošana ir juridisks pienākums. Tomēr arī tad, kad ir panākta atbilstība pamattiesībām ar juridisku spēku, ētiskas pārdomas mums var palīdzēt saprast, kā AI izstrāde, izvietošana un izmantošana varētu ietekmēt pamattiesības un to pamatā esošās vērtības, kā arī var palīdzēt dod sīkāk izstrādātus norādījumus saistībā ar centieniem noteikt, kas mums ar tehnoloģiju *jādara*, nevis ko mēs (patlaban) *varam* ar to izdarīt.

(41) ***Cilvēka cieņas respektēšana***. Cilvēka cieņas jēdziens ietver domu, ka ikvienam ir "iekšēja vērtība", ko nekad nedrīkst mazināt, apdraudēt vai apspiest citas personas – vai jaunas tehnoloģijas, piemēram, AI sistēmas[19]. Saistībā ar AI cilvēka cieņas ievērošana nozīmē, ka pret visiem cilvēkiem jāattiecas ar cieņu, kāda tiem pienākas kā morāliem *subjektiem*, nevis tikai *objektiem*, kurus var sijāt, šķirot, vētīt, ievirzīt, ietekmēt vai izmantot. Tādēļ AI sistēmas ir jāizstrādā tā, lai ievērotu, atbalstītu un aizsargātu cilvēku fizisko un garīgo neaizskaramību, personīgo un kultūras identitātes apziņu, kā arī pamatvajadzību apmierināšanu.[20]

(42) ***Indivīda brīvība***. Cilvēkiem jāsaglabā brīvība pašiem pieņemt lēmumus par savu dzīvi. Tā ietver brīvību no suverēna iejaukšanās un arī prasa valdības un nevalstisko organizāciju iesaistīšanos ar mērķi nodrošināt, ka indivīdi vai atstumšanai pakļautas personas var vienlīdz piekļūt AI sagādātajiem ieguvumiem un iespējām. AI sakarā indivīda brīvība paredz pienākumu mazināt (ne)tiešu prettiesisku piespiešanu, draudus garīgajai patstāvībai un garīgajai veselībai, nepamatotu novērošanu, maldināšanu un netaisnīgu izmantošanu. Faktiski indivīda brīvība nozīmē apņemšanos nodrošināt cilvēkiem iespēju īstenot vēl plašāku kontroli pār savu dzīvi, ieskaitot (citu starpā) arī uzņēmējdarbības brīvības, humanitāro un eksakto zinātņu brīvības, vārda brīvības,

---

[17] Jānorāda, ka apņemšanās veidot antropocentrisku AI un tās nostiprināšana pamattiesībās prasa kopīgus sociālos un konstitucionālos pamatus, kuros indivīda brīvība un cilvēka cieņas ievērošana ir gan praktiski iespējama, gan svarīga, nevis liek domāt par nepamatoti individualizētu priekšstatu par cilvēku.

[18] Saskaņā ar hartas 51. pantu to piemēro ES iestādēm un ES dalībvalstīm ES tiesību aktu īstenošanā.

[19] *C. McCrudden*, "*Human Dignity and Judicial Interpretation of Human Rights*", *EJIL*, 19(4), 2008. g.

[20] Par šādu "cilvēka cieņas" izpratni sk.: *E. Hilgendorf*, "*Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity*", iekļauts *D. Grimm, A. Kemmerer, C. Möllers* (eds.), "*Human Dignity in Context. Explorations of a Contested Concept*", 2018. g., 325. un turpmākās lpp.

tiesību uz privāto dzīvi un privātumu, pulcēšanās un biedrošanās brīvības aizsardzību.

(43) ***Cieņa pret demokrātiju, taisnīgumu un tiesiskumu***. Konstitucionālā demokrātijā visai valsts varai jābūt juridiski pilnvarotai un ierobežotai ar likumu. AI sistēmām jākalpo demokrātisko procesu uzturēšanai un veicināšanai, kā arī jārespektē vērtību daudzveidība un indivīdu dzīvē izdarītās izvēles. AI sistēmas nedrīkst kaitēt demokrātiskajiem procesiem, cilvēku apspriedēm vai demokrātiskām balsošanas sistēmām. AI sistēmām ir jāietver arī apņemšanās nodrošināt, la tās darbojoties nekaitē pamata saistībām, uz kurām balstās tiesiskums, un obligātajiem normatīvajiem aktiem, kā arī nodrošināt pienācīgu procesu un vienlīdzību likuma priekšā.

(44) ***Vienlīdzība, nediskriminēšana un solidaritāte*** – *ieskaitot to personu tiesības, kam draud atstumšana.* Ir jānodrošina, ka tiek vienlīdz ievērota visu cilvēku morālā vērtība un cieņa. Tas ir kas vairāk par nediskriminēšanu, kas pieļauj nošķirt savstarpēji atšķirīgas situācijas, balstoties uz objektīvu pamatojumu. AI kontekstā vienlīdzība prasa, lai sistēmas darbība neradītu rezultātus ar netaisnīgu neobjektivitāti (piem., datiem, ko izmanto AI sistēmu apmācībai, jābūt iespējami iekļaujošiem un jāatspoguļo dažādas iedzīvotāju grupas). Tam nepieciešams arī pienācīgi ņemt vērā varbūtēji neaizsargātas personas un grupas[21], kā darba ņēmējus, sievietes, invalīdus, mazākumtautības, bērnus, patērētājus un citas personas, kam draud atstumšana.

(45) ***Pilsoņu tiesības***. Pilsoņi var izmantot plašu tiesību klāstu, ieskaitot tiesības balsot, tiesības uz labu pārvaldi vai piekļuvi publiskiem dokumentiem, kā arī tiesības vērsties pie pārvaldes ar lūgumiem. AI sistēmas sniedz ievērojamu potenciālu, kas ļauj uzlabot vērienu un efektivitāti, ar kādu valdība sniedz sabiedrībai publiskas preces un pakalpojumus. Vienlaikus AI lietojumi var negatīvi ietekmēt pilsoņu tiesības, tāpēc tās ir jāaizsargā. Apzīmējuma "pilsoņu tiesības" izmantošana šajā dokumentā nav trešo valstu piederīgo un nelikumīgi (vai nelegāli) ES atrodošos personu tiesību noliegšana vai neievērošana, jo arī viņām ir tiesības starptautisko tiesību izpratnē un tādēļ arī AI jomā.

2.2 Ētikas principi AI sistēmu kontekstā[22]

(46) Daudzas publiskās, privātās un pilsoniskās organizācijas ir iedvesmojušās no pamattiesībām, izstrādājot AI ētiskos satvarus[23]. Eiropas Savienībā Eiropas Dabaszinātņu un jauno tehnoloģiju ētikas grupa (*EGE*) ir ierosinājusi deviņus pamatprincipus, kas ir balstīti uz ES līgumos un ES Pamattiesību hartā paredzētajām pamatvērtībām[24]. Mēs šo darbu attīstām tālāk, atzīstot lielāko daļu principu, kurus līdz šim piedāvājušas dažādas grupas, vienlaikus precizējot mērķus, kurus attīstīt un atbalstīt cenšas visi principi. Šie ētikas principi var iedvesmot jaunus un konkrētus regulatīvos instrumentus, palīdzēt interpretēt pamattiesības līdz ar mūsu sociāltehniskās vides attīstību laika gaitā, kā arī noteikt AI sistēmu izstrādes, izmantošanas un īstenošanas pamatojumu, dinamiski pielāgojoties pārmaiņām pašā sabiedrībā.

(47) AI sistēmām ir jāuzlabo indivīdu un kolektīva labklājība. Šajā sadaļā ir minēti **četri ētikas principi**, kuri sakņojas pamattiesībās un kuri ir jāievēro, lai nodrošinātu AI sistēmu uzticamu izstrādi, izvietošanu un izmantošanu. Tie ir formulēti kā **ētiski pienākumi**, kurus AI jomas praktiķiem ir jācenšas ievērot vienmēr. Nenosakot to kā hierarhiju, mēs turpmāk uzskaitām principus, ievērojot secību, kādā ES Pamattiesību hartā parādās

---

[21] Sk. glosārijā šā termina aprakstu atbilstīgi tā lietojumam dokumentā.

[22] Šie principi attiecas arī uz citu tehnoloģiju izstrādi, izvietošanu un izmantošanu, tādēļ tie nav paredzēti AI sistēmām vien. Tālāk tekstā mūsu mērķis ir izklāstīt to nozīmi konkrēti ar AI saistītā kontekstā.

[23] Paļaušanās uz pamattiesībām arī palīdz ierobežot regulatīvo nenoteiktību, jo tā var balstīties uz desmitiem gadu ilgo pamattiesību aizsardzības praksi ES, tādējādi sniedzot skaidrību, uztveramību un paredzamību.

[24] Nesen *AI4People* darba grupa ir apsekojusi minētos *EGE* principus un 36 citus līdz šim brīdim ieteiktos ētikas principus un apkopojusi tos četros virsprincipos: *L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena* (2018. g.), "*AI4People –An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*", *Minds and Machines* 28(4): 689.-707. lpp.

pamattiesības, uz kurām tie ir balstīti[25].

(48) Tie ir šādi principi:

(i) Cilvēka patstāvības ievērošana
(ii) Kaitējuma novēršana
(iii) Taisnīgums
(iv) Izskaidrojamība

(49) Daudzi no tiem jau lielā mērā ir atspoguļoti esošajās juridiskajās prasībās, kuru ievērošana ir obligāta, tādēļ tie ir saistīti arī ar "likumīgu AI", kas ir uzticama AI pirmā iezīme[26]. Tomēr, kā norādīts iepriekš, lai gan daudzi juridiskie pienākumi atspoguļo ētikas principus, ētikas principu ievērošana ir kas vairāk nekā formāla spēkā esošo tiesību aktu ievērošana[27].

- Cilvēka patstāvības ievērošanas princips

(50) Pamattiesības, uz kuru pamata ir veidota ES, ir vērstas uz to, lai nodrošinātu cieņu pret cilvēka brīvību un patstāvību. Cilvēkiem, kas saskaras ar AI sistēmām, jāspēj saglabāt pilnīga un efektīva pašnoteikšanās, kā arī jāspēj piedalīties demokrātiskajos procesos. AI sistēmas nedrīkst cilvēkus nepamatoti pakļaut, piespiest, maldināt, ietekmēt, ievirzīt vai iegrožot. Tā vietā AI sistēmas jāveido tā, lai tās uzlabotu, papildinātu un sniegtu iespējas cilvēku kognitīvajām, sociālajām un kultūras prasmēm. Cilvēku un AI sistēmu savstarpējā funkciju sadalē ir jāievēro antropocentriski izstrādes principi un pēc būtības jāparedz iespēja cilvēkiem izdarīt izvēli. Tas nozīmē, ka ir jānodrošina, lai AI sistēmu darba procesus pārraudzītu[28] un kontrolētu cilvēks. AI sistēmas, iespējams, arī būtiski mainīs situāciju darba jomā. Tām jāsniedz atbalsts cilvēkiem darba vidē un jābūt vērstām uz lietderīga darba radīšanu.

- Kaitējuma novēršanas princips

(51) AI sistēmas nedrīkst ne izraisīt, ne palielināt kaitējumu[29] vai citādi nelabvēlīgi ietekmēt cilvēkus[30]. Šā iemesla dēļ ir jāaizsargā cilvēka cieņa, kā arī garīgā un fiziskā integritāte. AI sistēmām un vidēm, kurās tās darbojas, ir jābūt drošām un neapdraudētām. Tām jābūt tehniski noturīgām, turklāt ir jānodrošina, lai tās nevarētu izmantot ļaunprātīgi. Neaizsargātām personām ir jāvelta vairāk uzmanības, un tās ir jāņem vērā, izstrādājot un izvietojot AI sistēmas. Īpaša uzmanība ir jāpievērš arī situācijām, kurās AI sistēmas var izraisīt vai palielināt nelabvēlīgu ietekmi varas vai informācijas asimetrijas dēļ, piemēram, attiecībās starp darba devējiem un darba ņēmējiem, uzņēmumiem un patērētājiem vai valdībām un iedzīvotājiem. Kaitējuma novēršana ietver arī apsvērumus attiecībā uz dabu un visām dzīvajām būtnēm.

- Taisnīguma princips

(52) AI sistēmu izstrādei, izvietošanai un izmantošanai jābūt taisnīgai. Lai gan mēs atzīstam, ka taisnīgumu var izprast daudzos dažādos veidos, mēs uzskatām, ka taisnīgumam ir gan materiāli, gan procesuāli aspekti. Materiālie aspekti ietver apņemšanos nodrošināt vienlīdzīgu un taisnīgu ieguvumu un izmaksu sadali un nodrošināt, ka indivīdi un grupas nesaskaras ar neobjektivitāti, diskrimināciju un stigmatizāciju. Ja ir iespējams

---

[25] Cilvēka patstāvības ievērošana ir cieši saistīta ar tiesībām uz cilvēka cieņu un brīvību (atspoguļotas hartas 1. un 6. pantā). Kaitējuma novēršana ir cieši saistīta ar fiziskās un garīgās neaizskaramības aizsardzību (atspoguļota 3. pantā). Taisnīgums ir cieši saistīts ar tiesībām uz nediskriminēšanu, solidaritāti un tiesiskumu (atspoguļotas 21. un turpmākajos pantos). Izskaidrojamība un atbildība ir cieši saistītas ar tiesībām uz tiesiskumu (kā atspoguļots 47. pantā).

[26] Sk., piemēram, VDAR vai ES patērētāju tiesību aizsardzības noteikumus.

[27] Vairāk par šo tēmu sk., piemēram: *L. Floridi*, "*Soft Ethics and the Governance of the Digital*", *Philosophy & Technology,* 2018. gada marts, 31. sējums, 1. izdevums, 1.–8. lpp.

[28] Cilvēka veiktas pārraudzības jēdziens ir plašāk raksturots turpmāk 65. punktā.

[29] Kaitējums var būt individuāls vai kolektīvs un var ietvert nemantisku kaitējumu sociālajai, kultūras un politiskajai videi.

[30] Kaitējuma novēršana ietver arī personu un sociālo grupu dzīvesveidu, tā izvairoties, piemēram, no kaitējuma kultūrai.

izvairīties no neobjektivitātes, AI sistēmas var pat uzlabot sociālo taisnīgumu. Ir jāveicina arī vienlīdzīgas iespējas piekļuvē izglītībai, precēm, pakalpojumiem un tehnoloģijām. Turklāt AI sistēmu izmantošana nekādā gadījumā nedrīkst novest pie (gala)lietotāju maldināšanas vai viņu izvēles brīvības ierobežošanas. Taisnīgums arī nozīmē, ka AI jomas praktiķiem ir jāievēro līdzekļu un mērķa proporcionalitātes princips, kā arī rūpīgi jāapsver, kā saskaņot konkurējošas intereses un mērķus[31]. Taisnīguma procesuālais aspekts ietver iespēju apstrīdēt AI sistēmu un cilvēku, kuri tās vada, pieņemtos lēmumus un panākt efektīvu tiesisko aizsardzību[32]. Šim nolūkam ir jāspēj identificēt par lēmumu atbildīgā struktūra, turklāt lēmumu pieņemšanas procesam jābūt izskaidrojamam.

- Izskaidrojamības princips

(53) Veidojot un uzturot lietotāju uzticēšanos AI sistēmām, ļoti svarīga loma ir izskaidrojamībai. Tas nozīmē, ka procesiem ir jābūt pārredzamiem, AI sistēmu spējas un mērķi ir skaidri jāpaziņo un lēmumiem, ciktāl tas iespējams, ir jābūt izskaidrojamiem personām, kuras tie tieši un netieši skar. Bez šādas informācijas lēmumu nevar pienācīgi apstrīdēt. Ne vienmēr ir iespējams izskaidrot, kādēļ modelis ir radījis konkrētus izvaddatus vai pieņēmis konkrētu lēmumu (un kāda ievaddatu faktoru kombinācija to noteikusi). Šie gadījumi ir tā dēvētie "melnās kastes" algoritmi, kam jāpievērš īpaša uzmanība. Šādos apstākļos var būt nepieciešami citi izskaidrojamības pasākumi (piem., trasējamība, pārbaudāmība un skaidra informēšana par sistēmas spējām), ar noteikumu, ka sistēma kopumā ievēro pamattiesības. Nepieciešamā izskaidrojamības pakāpe ir ļoti atkarīga no konteksta un tā, cik smagas sekas var radīt kļūdaini vai citādi neprecīzi izvaddati[33].

2.3 Pretrunas principu starpā

(54) Starp minētajiem principiem var rasties pretrunas, kuru atrisināšanai nav viennozīmīga risinājuma. Ievērojot ES būtisko apņemšanos nodrošināt demokrātisku iesaisti, pienācīgu procesu un atklātu politisko līdzdalību, ir jāizveido atbildīgas apspriešanās paņēmieni šādu pretrunu risināšanai. Piemēram, dažādās lietojuma jomās var pastāvēt pretruna starp *kaitējuma novēršanas principu* un *cilvēka patstāvības principu*. Kā piemēru var minēt AI sistēmu izmantošanu uz prognozēm balstītam policijas darbam, kas var palīdzēt samazināt noziedzību, taču ietver darbības, kuras aizskar indivīdu brīvību un privātumu. Turklāt AI sistēmu kopējiem ieguvumiem ir būtiski jāpārsniedz prognozējamie atsevišķie riski. Lai gan šie principi nepārprotami sniedz norādījumus ceļā uz risinājumiem, tie joprojām ir abstraktas ētikas normas. Tādēļ nevar gaidīt, ka AI jomas praktiķi spēs rast pareizo risinājumu, balstoties uz minētajiem principiem, tomēr tiem ētikas dilemmas un kompromisi ir jāvērtē pamatotās, uz pierādījumiem balstītās pārdomās, nevis intuitīvi vai nejauši un patvaļīgi. Tomēr var būt situācijas, kad ētiski pieņemamus kompromisus rast nav iespējams. Dažas pamattiesības un ar tām saistītie principi (piem., cilvēka cieņa) ir absolūti un tiem nevar piemērot līdzsvarošanu.

Galvenie no I nodaļas izrietošie norādījumi

- ✓ Izstrādāt, izvietot un izmantot AI sistēmas tā, lai ievērotu ētikas principus: *cilvēka patstāvības ievērošanu, kaitējuma novēršanu, taisnīgumu un izskaidrojamību.* Atzīt un risināt iespējamās pretrunas šo principu starpā.
- ✓ Pievērst īpašu uzmanību situācijām, kurās ir iesaistītas mazāk aizsargātas grupas, piemēram, bērni, personas ar invaliditāti un citas grupas, kas vēsturiski atradušās nelabvēlīgākā situācijā vai ir pakļautas

[31] Pastāv saikne ar proporcionalitātes principu (atspoguļojas principā, ka "ar lielgabalu nav jāšauj pa zvirbuļiem"). Mērķa sasniegšanas līdzekļiem (piem., datu izguves pasākumiem, kas ieviesti, lai īstenotu AI optimizācijas funkciju) ir jāaprobežojas vienīgi ar noteikti nepieciešamo. Minētais paredz arī, ka tad, ja mērķa sasniegšanai ir pieejami vairāki atšķirīgi pasākumi, priekšroka jādod tam, kas ir vismazāk pretrunā pamattiesībām un ētikas normām (piem., AI izstrādātājiem vienmēr jādod priekšroka publiskā sektora datiem, nevis persondatiem). Jāmin arī samērīgums lietotāja un izvietotāja attiecībās – ir jāņem vērā uzņēmumu tiesības (tostarp uz intelektuālo īpašumu un konfidencialitāti), no vienas puses, un lietotāja tiesības, no otras puses.

[32] Tostarp – izmantojot tiesības uz biedrošanos un tiesības iestāties arodbiedrībā darba vidē, kā paredz Eiropas Savienības Pamattiesību hartas 12. pants.

[33] Piemēram, AI sistēmas ģenerēti neprecīzi iepirkšanās ieteikumi nerada lielas ētiskas problēmas, taču citādi ir ar AI sistēmām, kas izvērtē, vai par noziedzīgu nodarījumu notiesātais būtu nosacīti atbrīvojams pirms termiņa.

atstumšanas riskam, un/vai situācijām, kuras raksturo varas vai informācijas asimetrija, piemēram, attiecībās starp darba devēju un darbinieku vai uzņēmumu un patērētāju[34].

✓ Atzīt un ņemt vērā, ka dažām AI sistēmām, lai gan tās var sniegt daudzus būtiskus ieguvumus indivīdiem un sabiedrībai, var būt arī negatīva ietekme, ieskaitot grūti paredzamu, nosakāmu vai izmērāmu ietekmi (piemēram, uz demokrātiju, tiesiskumu un taisnīgu sadali, kā arī uz cilvēka prātu kā tādu). Attiecīgos gadījumos veikt pienācīgus pasākumus šā riska mazināšanai samērīgi ar riska apmēru.

II. II nodaļa. Uzticama AI īstenošana

(55) Šajā nodaļā doti norādījumi par uzticama AI īstenošanu un realizāciju – attīstot I nodaļā izklāstītos principus, ir norādītas septiņas izpildāmas prasības. Iepazīstinām arī ar pašlaik pieejamajiem tehniskajiem un netehniskajiem paņēmieniem šo prasību īstenošanai visā AI sistēmas dzīves ciklā.

1. Prasības uzticamam AI

(56) Lai sasniegtu uzticamu AI, I nodaļā izklāstītie principi ir jāizsaka konkrētās prasībās. Šīs prasības piemēro dažādām ieinteresētajām personām, kuras piedalās AI sistēmu dzīves ciklā: izstrādātājiem, izvietotājiem un galalietotājiem, kā arī plašākai sabiedrībai. Ar izstrādātājiem saprot personas, kas pēta, projektē un/vai izstrādā AI sistēmas. Ar izvietotājiem saprot publiskas vai privātas organizācijas, kas izmanto AI sistēmas savos darbības procesos, kā arī piedāvājot citiem preces un pakalpojumus. Galalietotāji ir personas, kas tieši vai netieši saskaras ar AI sistēmu. Visbeidzot, plašāka sabiedrība ir visas pārējās personas, kuras tieši vai netieši skar AI sistēmas.

(57) Dažādām ieinteresēto personu grupām ir dažādas lomas prasību izpildes nodrošināšanā:

a. izstrādātājiem ir jāīsteno un jāpiemēro prasības projektēšanas un izstrādes procesos;
b. izvietotājiem ir jānodrošina, ka to izmantotās sistēmas un piedāvātās preces un pakalpojumi atbilst prasībām;
c. galalietotājiem un plašākai sabiedrībai ir jābūt informētai par šīm prasībām un jāspēj pieprasīt to izpildi.

(58) Turpmāk sniegtais prasību saraksts nav izsmeļošs[35]. Tas ietver sistēmiskos, individuālos un sabiedrības aspektus:

1 Cilvēka subjektība un virsvadība
Ieskaitot pamattiesības, cilvēka subjektību un cilvēka virsvadību

2 Tehniskā noturība un drošums
Ieskaitot noturību pret uzbrukumiem un drošību, rezerves plānus un vispārējo drošumu, precizitāti, bezatteiksmi un reproducējamību

3 Privātums un datu pārvaldība
Ieskaitot cieņu pret privātumu, datu kvalitāti un integritāti, kā arī piekļuvi datiem

4 Pārredzamība
Ieskaitot trasējamību, izskaidrojamību un paziņošanu

5 Daudzveidība, nediskriminēšana un taisnīgums
Ieskaitot izvairīšanos no neobjektivitātes, pieejamību un universālo dizainu, kā arī ieinteresēto personu

[34] Sk. ES hartas 24.–27. pantu, kuros minētas bērnu un vecāka gadagājuma cilvēku tiesības, personu ar invaliditāti integrācija un darba ņēmēju tiesības. Sk. arī 38. pantu, kas attiecas uz patērētāju tiesību aizsardzību.

[35] Nenosakot to kā hierarhiju, mēs turpmāk uzskaitām principus, ievērojot secību, kādā ES hartā parādās ar tiem saistītie principi un tiesības.

dalību

6 Sabiedrības un vides labklājība

Ieskaitot ilgtspēju un draudzīgumu videi, sociālo ietekmi, sabiedrību un demokrātiju

7 Atbildīgums

Ieskaitot pārbaudāmību, negatīvās ietekmes mazināšanu un ziņošanu par to, kompromisus un tiesisko aizsardzību

2. attēls. Septiņu prasību savstarpējā saikne: visas ir vienlīdz svarīgas un atbalsta cita citu, un tās ir jāīsteno un jāizvērtē visā AI sistēmas dzīves ciklā

(59) Lai gan visas prasības ir vienlīdz svarīgas, piemērojot tās dažādās nozarēs un jomās, ir jāņem vērā konteksts un iespējamās pretrunas to starpā. Šīs prasības ir jāpiemēro visā AI sistēmas dzīves ciklā, un piemērošana ir atkarīga no konkrētā lietojuma. Lai gan vairums prasību attiecas uz visām AI sistēmām, īpaša uzmanība ir jāpievērš sistēmām, kuras tieši vai netieši skar fiziskas personas. Tādēļ dažos lietojumos (piemēram, rūpniecības vidē) šīm prasībām var būt mazāka nozīme.

(60) Minētās prasības ietver elementus, kas dažos gadījumos jau ir atspoguļoti spēkā esošajos tiesību aktos. Mēs uzsveram, ka atbilstīgi uzticama AI pirmajai iezīmei AI sistēmu izstrādātāji un izvietotāji ir atbildīgi par to, lai tiktu ievēroti juridiskie pienākumi, kas noteikti gan horizontāli piemērojamos noteikumos, gan nozaru īpašajos noteikumos.

(61) Turpmākajos punktos katra no prasībām ir izklāstīta sīkāk.

1. <u>Cilvēka subjektība un virsvadība</u>

(62) AI sistēmām ir jāatbalsta cilvēku patstāvība un lēmumu pieņemšana, kā paredz *cilvēka patstāvības ievērošanas princips*. Tas paredz, ka AI sistēmām gan jābūt demokrātiskas, plaukstošas un taisnīgas sabiedrības nodrošinātājām, atbalstot lietotāja subjektību un veicinot pamattiesības, gan jānodrošina, ka cilvēks var tās pārraudzīt.

(63) ***Pamattiesības***. AI sistēmas, līdzīgi kā daudzas citas tehnoloģijas, var gan atbalstīt, gan apgrūtināt pamattiesību ievērošanu. Tās var sniegt cilvēkiem labumu, piemēram, palīdzot tiem izsekot saviem personas datiem vai palielinot izglītības pieejamību, tādējādi atbalstot tiesības uz izglītību. Tomēr, ņemot vērā AI sistēmu tvērumu un spējas, tās var arī negatīvi ietekmēt pamattiesības. Situācijās, kurās pastāv šāds risks, ir jānovērtē ietekme uz pamattiesībām. Tam ir jānotiek pirms izstrādes un jāietver izvērtējums, vai šo risku var samazināt vai pamatot kā nepieciešamu demokrātiskā sabiedrībā, lai ievērotu citu personu tiesības un brīvības. Turklāt ir jāievieš mehānismi, ar kuriem saņemt trešo personu atsauksmes par AI sistēmām, kas var aizskart pamattiesības.

(64) ***Cilvēka subjektība.*** Lietotājiem ir jāspēj pieņemt informētus patstāvīgus lēmumus attiecībā uz AI sistēmām. Tiem ir jānodrošina zināšanas un rīki, lai pienācīgā līmenī izprastu AI sistēmas un mijdarbotos ar tām, un, ja tas iespējams, jāļauj pašiem veikt pamatotu novērtējumu vai apstrīdēt sistēmu. AI sistēmām jāatbalsta indivīdi labākas un informētākas izvēles izdarīšanā atbilstīgi viņu mērķiem. AI sistēmas dažkārt var tikt izvietotas, lai veidotu un ietekmētu cilvēku uzvedību, izmantojot mehānismus, kurus var būt grūti atklāt, jo tie var iedarboties uz zemapziņas procesiem, tostarp dažādos negodīgas manipulācijas, maldināšanas, rīcības ievirzīšanas un ietekmēšanas veidos, no kuriem ikviens var apdraudēt indivīdu patstāvību. Sistēmas funkcionalitātē vislielākā nozīme jāpiešķir kopējam lietotāja patstāvības principam. Vissvarīgākās šajā ziņā ir tiesības netikt pakļautam lēmumiem, kas ir balstīti tikai uz datu automatizētu apstrādi, ja tie rada tiesiskas sekas lietotājiem vai citādi būtiski tos ietekmē[36].

(65) ***Cilvēka virsvadība***. Cilvēka virsvadība palīdz nodrošināt, ka AI sistēma nekaitē cilvēka patstāvībai un nerada citu nelabvēlīgu ietekmi. Virsvadību var panākt, izmantojot pārvaldes mehānismus, piemēram, pieeju, kas paredz cilvēka iesaistīšanu ciklā (*human-in-the-loop*, *HITL*), cilvēka pārraudzību pār ciklu (*human-on-the-loop*, *HOTL*) vai cilvēka vadību (*human-in-command*, *HIC*). *HITL* ir iespēja cilvēkam iejaukties ikvienā sistēmas lēmumu pieņemšanas ciklā, kas daudzos gadījumos nav ne iespējams, ne vēlams. *HOTL* ir iespēja cilvēkam iejaukties sistēmas projektēšanas ciklā un pārraudzīt sistēmas darbību. *HIC* ir iespēja pārraudzīt AI sistēmas kopējo darbību (arī tās plašāko ekonomisko, sociālo, juridisko un ētisko ietekmi) un spēja izlemt, vai un kā izmantot sistēmu konkrētā situācijā. Tā var ietvert lēmumu konkrētā situācijā neizmantot AI sistēmu, paredzēt cilvēka rīcības brīvības pakāpi sistēmas izmantošanas laikā vai nodrošināt iespēju atcelt sistēmas pieņemtu lēmumu. Turklāt ir jānodrošina, ka publiskie uzraugi spēj īstenot pārraudzību saskaņā ar to pilnvarām. Atkarā no AI sistēmas pielietošanas jomas un iespējamā riska var būt nepieciešams, lai uzraudzības mehānismi dažādā apmērā atbalstītu citus drošības un kontroles pasākumus. Pastāvot vienādiem apstākļiem – jo mazāk cilvēka pārraudzības pār AI sistēmu iespējams, jo plašāka testēšana un stingrāka pārvaldīšana nepieciešama.

2. Tehniskā noturība un drošums

(66) Uzticama AI panākšanā ļoti liela nozīme ir tehniskajai noturībai, kas ir cieši saistīta ar *kaitējuma novēršanas principu*. Tehniskā noturība prasa, lai AI sistēmas tiktu izstrādātas, preventīvi pieejot riskam un tā, lai tās stabili darbotos paredzētajā veidā, vienlaikus mazinot netīšu un neparedzētu kaitējumu, kā arī novēršot nepieņemamu kaitējumu. Minētajam jāattiecas arī uz iespējamām to darbības vides izmaiņām vai citu tādu (cilvēka un mākslīgu) faktoru klātbūtni, kas var nelabvēlīgi mijiedarboties ar sistēmu. Tāpat ir jānodrošina arī cilvēku fiziskā un garīgā neaizskaramība.

(67) ***Noturība pret uzbrukumu un drošība***. Kā jebkura programmatūras sistēma, AI sistēmas ir jāaizsargā pret

[36] Sk. VDAR 22. pantu, kurā šīs tiesības jau ir nostiprinātas.

ievainojamību, kuras dēļ tās varētu ļaunprātīgi izmantot ienaidnieki, piemēram, tās uzlaužot. Uzbrukumi var būt vērsti pret datiem (datu "saindēšana"), modeli (modeļa noplūde) vai pamatā esošo infrastruktūru – kā programmatūru, tā aparatūru. Ja AI sistēmai tiek uzbrukts, piemēram, pretinieka uzbrukumā, var tikt izmainīti gan dati, gan sistēmas uzvedība, kā rezultātā sistēma var pieņemt atšķirīgus lēmumus vai atslēgties pavisam. Sistēmas un datus var sabojāt arī ar ļaunprātīgu nodomu vai pakļaujot negaidītām situācijām. Nepietiekami drošības procesi var arī izraisīt kļūdainus lēmumus vai pat fizisku kaitējumu. Lai AI sistēmas uzskatītu par drošām[37], ir jāņem vērā iespējamie neparedzētie AI lietojumi (piemēram, divējādās lietošanas gadījumi) un iespēja, ka ļaunprātīgas personas izmanto AI sistēmu negodprātīgi, kā arī jāveic pasākumi šo gadījumu novēršanai un riska mazināšanai.[38]

(68) ***Rezerves plāns un vispārējais drošums***. AI sistēmām ir nepieciešami aizsardzības pasākumi, kas problēmu gadījumā ļauj izmantot rezerves plānu. Tas var nozīmēt, ka AI sistēmām ir jāpāriet no statistikā balstītas procedūras uz kārtulbalstītu procedūru vai ka tām pirms darba turpināšanas ir jālūdz cilvēka operatora iesaistīšanās[39]. Ir jānodrošina, ka sistēma dara tai paredzēto darbu, nekaitējot dzīvām būtnēm un videi. Tas ietver neparedzētu seku un kļūdu mazināšanu. Turklāt ir jāparedz procesi ar AI sistēmu izmantošanu saistītā varbūtējā riska precizēšanai un novērtēšanai dažādās izmantošanas jomās. Nepieciešamo drošības pasākumu līmenis ir atkarīgs no tā, cik augstu risku rada AI sistēma, kas savukārt ir atkarīgs no sistēmas spējām. Ja var paredzēt, ka izstrādes process vai pati sistēma radīs īpaši augstu risku, ir ārkārtīgi svarīgi proaktīvi izstrādāt un testēt drošības pasākumus.

(69) ***Precizitāte.*** Precizitāte attiecas uz AI sistēmas spēju pieņemt pareizus spriedumus, piemēram, lai pareizi klasificētu informāciju pienācīgajās kategorijās, vai tās spēju paust pareizas prognozes, ieteikumus vai lēmumus, balstoties uz datiem vai modeļiem. Skaidrs un labi izveidots izstrādes un vērtēšanas process var atbalstīt, mazināt un labot neparedzētu risku, ko radījušas nepareizas prognozes. Ja no atsevišķām nepareizām prognozēm nav iespējams izvairīties, ir svarīgi, lai sistēma spētu norādīt, kāda ir šo kļūdu varbūtība. Augsta precizitāte ir īpaši svarīga situācijās, kad AI sistēma tieši ietekmē cilvēku dzīvi.

(70) ***Bezatteiksme un reproducējamība.*** Ir ļoti svarīgi, lai AI sistēmu rezultāti būtu reproducējami un nekļūdīgi. Bezatteiksmes AI sistēma ir tāda, kas pienācīgi darbojas ar dažādiem ievaddatiem un dažādās situācijās. Tas ir nepieciešams, lai rūpīgi pārbaudītu AI sistēmu un novērstu negribētu kaitējumu. Reproducējamība attiecas uz to, vai AI eksperimenta iznākums ir tāds pats, ja to atkārto ar tiem pašiem nosacījumiem. Tā ļauj zinātniekiem un politikas veidotājiem precīzi aprakstīt, ko dara AI sistēmas. Atkārtošanas datnes[40] var atvieglot testēšanas un uzvedības reproducēšanas procesu.

3. <u>Privātums un datu pārvaldība</u>

(71) Ar *kaitējuma novēršanas principu* cieši ir saistīts privātums – pamattiesības, ko jo īpaši skar AI sistēmas. Pret privātumu vērstā kaitējuma novēršana arī prasa pienācīgu datu pārvaldību, kas aptver izmantoto datu kvalitāti un integritāti, to atbilstību jomai, kurā tiks izvietota AI sistēma, to piekļuves protokolus un spēju apstrādāt datus tā, lai tiktu aizsargāts privātums.

(72) ***Privātums un datu aizsardzība***. AI sistēmām ir jāgarantē privātuma un datu aizsardzība visā sistēmas dzīves ciklā[41]. Tas ietver lietotāja sākotnēji sniegto informāciju, kā arī informāciju, kas par lietotāju tiek radīta, tam lietojot sistēmu (piem., izvaddatus, kurus AI sistēma ģenerē konkrētiem lietotājiem, vai to, kā lietotāji reaģējuši

---

37 Sk., piemēram, Eiropas Savienības koordinētā mākslīgā intelekta plāna 2.7. punktu.

38 AI sistēmu drošības vajadzībām varētu pastāvēt stingrs pienākums spēt izveidot pētniecības un izstrādes pozitīvās mijiedarbības ciklu, kas ietver uzbrukumu izpratni, pienācīgas aizsardzības izstrādi un vērtēšanas metodikas uzlabošanu. Lai to panāktu, ir jāveicina AI un drošības speciālistu konverģence. Turklāt visas iesaistītās personas ir atbildīgas par vienotu pārrobežu drošuma un drošības standartu izstrādi, kā arī par savstarpējas uzticēšanās apstākļu veidošanu, tā veicinot starptautisko sadarbību. Attiecībā uz iespējamiem pasākumiem sk.: *Malicious Use of AI* (*Avin S., Brundage M., et. al.*, 2018. g.).

39 Ir jāapsver arī scenāriji, kuros cilvēka iejaukšanās uzreiz nebūs iespējama.

40 Tās ir datnes, kas atkārto katru AI sistēmas izstrādes procesa daļu – no pētniecības un sākotnējās datu vākšanas līdz rezultātiem.

41 Te minami esošie privātuma tiesību akti, piemēram, VDAR un gaidāmā E-privātuma regula.

uz konkrētiem ieteikumiem). Digitāli ieraksti par cilvēku uzvedību var ļaut AI sistēmām izsecināt ne tikai indivīdu vēlmes, bet arī viņu dzimumorientāciju, vecumu, dzimumu, reliģiskos vai politiskos uzskatus. Lai cilvēki varētu uzticēties datu vākšanas procesam, ir jānodrošina, ka par viņiem savāktie dati netiek izmantoti, lai viņus nelikumīgi vai netaisnīgi diskriminētu.

(73) ***Datu kvalitāte un integritāte***. AI sistēmu rezultātus stipri ietekmē izmantoto datu kopu kvalitāte. Dati, kas tiek apkopoti, var ietvert sociāli radītu neobjektivitāti, neprecizitātes, kļūdas un aplamību. Šī situācija ir jārisina, pirms tiek veikta apmācība ar konkrēto datu kopu. Turklāt ir jānodrošina datu integritāte. Kaitīgu datu ievadīšana AI sistēmā var mainīt tās uzvedību, jo īpaši tādu sistēmu gadījumā, kas mācās pašas. Izmantotie procesi un datu kopas katrā solī, piemēram, plānošanā, apmācībā, testēšanā un izvietošanā, ir jātestē un jādokumentē. Minētajam jāattiecas arī uz AI sistēmām, kas nav pašu izstrādātas, bet iegūtas citur.

(74) ***Piekļuve datiem***. Jebkurā organizācijā, kura rīkojas ar fizisku personu datiem (neatkarīgi no tā, vai persona ir vai nav sistēmas lietotājs), ir jāievieš datu protokoli, kas regulē piekļuvi datiem. Šajos protokolos ir jānosaka, kas un kādās situācijās var piekļūt datiem. Atļauja to darīt būtu jāpiešķir tikai pienācīgi kvalificētām personām, kas ir kompetentas un kam ir nepieciešams piekļūt fizisko personu datiem.

4. Pārredzamība

(75) Šī prasība ir cieši saistīta ar *izskaidrojamības principu* un ietver AI sistēmai svarīgu elementu – datu, sistēmas un darbības modeļu – pārredzamību.

(76) ***Trasējamība***. Datu kopas un procesus, kuros iegūst AI sistēmas lēmumu, ieskaitot datu vākšanas un datu marķēšanas procesus un izmantotos algoritmus, ir jādokumentē, ciktāl vien iespējams, lai garantētu trasējamību un uzlabotu pārredzamību. Teiktais attiecas arī uz AI sistēmas pieņemtajiem lēmumiem. Tā tiek radīta iespēja identificēt iemeslus, kādēļ AI lēmums bijis kļūdains, un tas savukārt var palīdzēt novērst kļūdas nākotnē. Tādējādi trasējamība veicina pārbaudāmību un izskaidrojamību.

(77) ***Izskaidrojamība.*** Izskaidrojamība attiecas uz spēju izskaidrot gan AI sistēmas tehniskos procesus, gan ar tiem saistītos cilvēku lēmumus (piemēram, AI sistēmas piemērošanas jomas). Tehniskā izskaidrojamība paredz, ka cilvēkiem jāspēj izprast AI sistēmas pieņemtos lēmumus un izsekot tiem. Turklāt var nākties veidot kompromisu starp sistēmas izskaidrojamības uzlabošanu (kas var mazināt tās precizitāti) un tās precizitātes palielināšanu (uz izskaidrojamības rēķina). Vienmēr, kad AI sistēma būtiski ietekmē cilvēku dzīvi, ir jābūt iespējai pieprasīt pienācīgu skaidrojumu par lēmumu pieņemšanas procesu AI sistēmā. Šim skaidrojumam ir jābūt savlaicīgam un pielāgotam attiecīgās ieinteresētās personas (piem., nespeciālista, regulatora, pētnieka) zināšanu līmenim. Turklāt ir jābūt pieejamiem paskaidrojumiem par to, cik lielā mērā AI sistēma ietekmē un veido lēmumu pieņemšanas procesu organizācijā, kā arī par sistēmas projektēšanas gaitā izdarītajām izvēlēm un tās izvietošanas pamatojumu (tādējādi nodrošinot darbības modeļa pārredzamību).

(78) ***Saziņa.*** AI sistēmas nedrīkst lietotājiem uzdot sevi par cilvēkiem; cilvēkiem ir tiesības būt informētiem, ka to saskarsmes partneris ir AI sistēma. Tas nozīmē, ka AI sistēmām jābūt attiecīgi identificējamām. Turklāt, ja tas nepieciešams pamattiesību ievērošanas nodrošināšanai, ir jāpiedāvā iespēja atteikties no šādas saskarsmes par labu saskarsmei ar cilvēku. Papildus minētajam AI sistēmas spējas un ierobežojumi būtu attiecīgajam lietojumam piemērotā veidā jāpaziņo AI jomas praktiķiem vai galalietotājiem. Tas varētu ietvert informēšanu par AI sistēmas precizitātes līmeni, kā arī tās ierobežojumiem.

5. Daudzveidība, nediskriminēšana un taisnīgums

(79) Lai sasniegtu uzticamu AI, ir jānodrošina iekļaušana un daudzveidība visā AI sistēmas dzīves ciklā. Papildus tam, ka procesa gaitā tiek ņemtas vērā un iesaistītas visas skartās ieinteresētās personas, minētais ietver arī vienlīdzīgas piekļuves nodrošināšanu ar iekļaujošu projektēšanas procesu un vienlīdzīgu attieksmi. Šī prasība ir cieši saistīta ar *taisnīguma principu*.

(80) ***Izvairīšanās no neobjektivitātes***. Datu kopām, ko izmanto AI sistēmas (gan apmācībai, gan darbībai), var piemist tādi trūkumi kā netīšas vēsturiskas neobjektivitātes iekļaušana, nepilnīgums un sliktas pārvaldes modeļi. Šādu aizspriedumu turpināšana var novest pie neparedzētiem (ne)tiešiem aizspriedumiem un diskriminācijas[42] pret noteiktām grupām vai personām, un tas var padziļināt aizspriedumus un atstumšanu. Kaitējumu var radīt arī apzināta (patērētāju) aizspriedumu ļaunprātīga izmantošana vai negodīga konkurēšana, piemēram, cenu homogenizācija uz slepenas norunas pamata vai nepārredzamā tirgū[43]. Identificējama un diskriminējoša neobjektivitāte, ja iespējams, jānovērš vākšanas stadijā. Neobjektivitātei var būt pakļauta arī AI sistēmu izstrāde (piem., algoritmu programmēšana). Tam var pretoties, ieviešot pārraudzības procesus, lai skaidri un pārredzami analizētu un risinātu sistēmas mērķi, robežas, prasības un lēmumus. Arī dažādas izcelsmes, kultūru un nozaru darbinieku piesaistīšana var nodrošināt viedokļu daudzveidību un būtu atbalstāma.

(81) ***Pieejamība un universālais dizains***. Īpaši uzņēmumu darījumos ar patērētājiem sistēmām ir jābūt orientētām uz lietotāju un projektētām tā, lai AI preces vai pakalpojumus varētu izmantot ikviens neatkarīgi no vecuma, dzimuma, spējām vai īpašībām. Īpaši svarīgi ir, lai šai tehnoloģijai varētu piekļūt personas ar invaliditāti, kas ir sastopamas visās sabiedrības grupās. AI sistēmām nedrīkst piemērot pieeju "viens risinājums visiem", un ir jāapsver universālā dizaina[44] principu izmantošana, orientējoties uz iespējami plašu lietotāju loku un ievērojot attiecīgos pieejamības standartus[45]. Šādi visiem cilvēkiem tiks nodrošināta taisnīga piekļuve un aktīva līdzdalība esošajās un jaunajās ar datora starpniecību veiktajās cilvēka darbībās, kā arī piekļuve palīgtehnoloģijām.[46]

(82) ***Ieinteresēto personu dalība***. Lai izstrādātu uzticamas AI sistēmas, ir ieteicams apspriesties ar ieinteresētajām personām, kuras sistēma var tieši vai netieši ietekmēt visā tās dzīves ciklā. Ir lietderīgi lūgt regulāras atsauksmes arī pēc sistēmas izvietošanas un izveidot ilgtermiņa mehānismus ieinteresēto personu līdzdalībai, piemēram, nodrošinot darbinieku informēšanu, apspriešanos ar tiem un to līdzdalību visā organizācijas AI sistēmas īstenošanas procesā.

6. <u>Sabiedrības un vides labklājība</u>

(83) Atbilstīgi *taisnīguma principam* un *kaitējuma novēršanas principam* arī sabiedrība kopumā, citas jūtīgas būtnes un vide ir jāuzskata par ieinteresētajām personām visā AI dzīves ciklā. Ir jāatbalsta AI sistēmu ilgtspēja un ekoloģiskais atbildīgums, tāpat ir jāveicina tādu AI risinājumu pētniecība, kas ir vērsti uz pasaules līmenī svarīgām jomām, piemēram, ilgtspējīgas attīstības mērķiem. Būtu vēlams, lai AI tiktu izmantots visu cilvēku, arī nākamo paaudžu, labā.

(84) ***Ilgtspējīgs un videstīgs AI***. AI sistēmas var palīdzēt risināt dažas no visaktuālākajām sabiedrības problēmām, tomēr ir jānodrošina, lai tas notiktu iespējami draudzīgi videi. Šajā sakarā ir jāvērtē sistēmas izstrādes, izvietošanas un izmantošanas process, kā arī visa piegādes ķēde, piemēram, kritiski izvērtējot resursu un enerģijas patēriņu apmācības laikā un priekšroku dodot mazāk kaitīgām iespējām. Ir jāstimulē pasākumi, ar ko tiek nodrošināta visas AI sistēmas piegādes ķēdes videstība.

(85) ***Sociālā ietekme***. Bieža saskarsme ar sociālām AI sistēmām[47] visās mūsu dzīves jomās (izglītībā, darbā, aprūpē, izklaidē) var mainīt mūsu priekšstatus par sociālo darbību vai ietekmēt mūsu sociālās attiecības un saiknes. Lai

---

[42] Tiešās un netiešās diskriminācijas definīciju sk., piemēram, Padomes Direktīvas 2000/78/EK (2000. gada 27. novembris), ar ko nosaka kopēju sistēmu vienlīdzīgai attieksmei pret nodarbinātību un profesiju, 2. pantā. Sk. arī ES Pamattiesību hartas 21. pantu.

[43] Sk. Eiropas Savienības Pamattiesību aģentūras dokumentu "*BigData: Discrimination in data-supported decision making (2018)*": http://fra.europa.eu/en/publication/2018/big-data-discrimination.

[44] Publiskā iepirkuma direktīvas 42. pantā ir noteikts, ka tehniskajās specifikācijās ir jāņem vērā pieejamība un projektēšana visu lietotāju ērtībām.

[45] Piemēram, EN 301 549.

[46] Šī prasība ir saistīta ar Apvienoto Nāciju Organizācijas Konvenciju par personu ar invaliditāti tiesībām.

[47] Tā apzīmē AI sistēmas, kas sazinās un mijdarbojas ar cilvēkiem, simulējot sabiedriskumu cilvēku un robotu (ko iemieso AI) mijdarbībā, vai kas virtuālajā realitātē darbojas kā avatāri. Minētajām sistēmām šādas rīcības dēļ piemīt potenciāls mainīt mūsu sociālkulturālos ieradumus un mūsu sabiedriskās dzīves ritējumu.

gan AI sistēmas var tikt izmantotas sociālo prasmju uzlabošanai[48], tās tikpat viegli var veicināt to pasliktināšanos. Tās var ietekmēt arī cilvēku fizisko un garīgo labsajūtu. Tādēļ šo sistēmu ietekme ir rūpīgi jāuzrauga un jāapsver.

(86) ***Sabiedrība un demokrātija.*** Papildus tam, ka tiek vērtēta AI sistēmas izstrādes, izvietošanas un izmantošanas ietekme uz indivīdiem, šī ietekme būtu jāvērtē arī no sabiedrības viedokļa, ņemot vērā uz iestādēm, demokrātiju un plašāku sabiedrību atstātās sekas. AI sistēmu izmantošana ir rūpīgi jāapsver, jo īpaši ar demokrātiskajiem procesiem saistītās situācijās, tostarp ne tikai politiskajā lēmumu pieņemšanā, bet arī saistībā ar vēlēšanām.

7. Atbildība

(87) Atbildības prasība papildina minētās prasības un ir cieši saistīta ar *taisnīguma principu*. Tā paredz, ka ir jāievieš mehānismi AI sistēmu un to rezultātu atbildīguma un pārskata nodrošināšanai gan pirms, gan pēc to ieviešanas.

(88) ***Pārbaudāmība.*** Pārbaudāmība ietver iespēju izvērtēt algoritmus, datus un projektēšanas procesus. Minētais nenozīmē, ka ar AI sistēmu saistītajai informācijai par darbības modeļiem un intelektuālo īpašumu vienmēr jābūt publiski pieejamai. Tehnoloģijas uzticamību var veicināt iekšējo un ārējo revidentu veikti izvērtējumi un izvērtēšanas ziņojumu pieejamība. AI sistēmām, kuru lietojumi skar pamattiesības, ieskaitot drošību noteicošus lietojumus, jābūt neatkarīgi revidējamām.

(89) ***Negatīvās ietekmes mazināšana un ziņošana par to***. Ir jānodrošina iespējas gan ziņot par rīcību vai lēmumiem, kas noved pie noteikta sistēmas rezultāta, gan reaģēt uz šāda rezultāta sekām. Attiecībā uz (ne)tieši skartajām personām ir īpaši svarīgi identificēt, izvērtēt, paziņot un mazināt AI sistēmu iespējamo negatīvo ietekmi. Trauksmes cēlējiem, NVO, arodbiedrībām vai citām struktūrām, kad tās ziņo par būtiskām bažām attiecībā uz AI balstītu sistēmu, ir jābūt pieejamai pienācīgai aizsardzībai. Negatīvo ietekmi var palīdzēt mazināt ietekmes novērtējumu izmantošana (piemēram, "sarkanās komandas" veikta pārbaude vai dažādi algoritmiskie ietekmes novērtējumi) gan pirms AI sistēmu izstrādes, izvietošanas un izmantošanas, gan to laikā. Šiem novērtējumiem ir jābūt samērīgiem ar AI sistēmu radīto risku.

(90) ***Kompromisi.*** Īstenojot minētās prasības, to starpā var rasties pretrunas, kas neizbēgami var radīt nepieciešamību pēc kompromisiem. Šie kompromisi ir jārisina racionāli un metodiski, ievērojot sasniegto tehnisko līmeni. Tas nozīmē, ka ir jānosaka attiecīgās intereses un vērtības, ko skar AI sistēma, un, ja radies konflikts, savstarpējie kompromisi ir skaidri jāatzīst un jāvērtē, ņemot vērā risku, ko tie rada ētikas principiem, tostarp pamattiesībām. Situācijās, kurās nav iespējams rast ētiski pieņemamus kompromisus, AI sistēmas izstrādi, izvietošanu un izmantošanu šādā veidā nedrīkst turpināt. Lēmums par to, kādu kompromisu pieņemt, ir jāpamato un pienācīgi jādokumentē. Lēmumu pieņēmējam jāatbild par attiecīgā kompromisa īstenošanas veidu un periodiski jāpārskata pieņemtā lēmuma atbilstība, lai nodrošinātu, ka nepieciešamības gadījumā sistēmā var tikt veiktas vajadzīgās izmaiņas. [49]

(91) ***Tiesiskā aizsardzība***. Ir jāparedz pieejami mehānismi, kas nodrošina pietiekamu tiesisko aizsardzību, ja notikusi netaisnīga nelabvēlīga ietekme[50]. Lai nodrošinātu uzticēšanos, ir svarīgi zināt, ka sarežģījumu gadījumā būs

---

[48] Sk., piemēram, ES finansētu projektu tādas uz AI balstītas programmatūras izstrādei, kas ļauj robotiem efektīvāk kontaktēties ar autiskiem bērniem, cilvēka vadītās terapijas sesijās, kas tiem palīdz uzlabot sociālās un komunikatīvās prasmes: http://ec.europa.eu/research/infocentre/article_en.cfm?id=/research/headlines/news/article_19_03_12_en.html?infocentre&item=Infocentre&artid=49968.

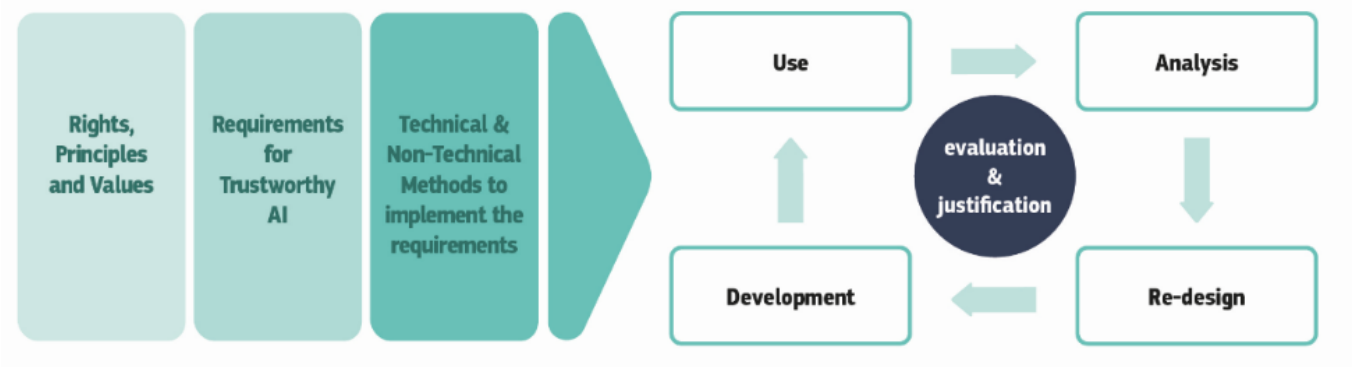
[49] To var palīdzēt sasniegt dažādi pārvaldības modeļi. Piemēram, iekšējā un/vai ārējā ētikas (un konkrētās nozares) eksperta vai komisijas iesaiste var noderēt, lai konstatētu iespējamo konfliktu jomas un ieteiktu, kā šos konfliktus vislabāk atrisināt. Noderīga ir arī saturīga apspriešanās un diskusijas ar ieinteresētajām personām, tostarp ar personām, kuras varētu nelabvēlīgi ietekmēt AI sistēma. Eiropas universitātēm būtu jāuzņemas vadošā loma nepieciešamo ētikas ekspertu apmācībā.

[50] Sk. arī Eiropas Savienības Pamattiesību aģentūras atzinumu "Uzlabota tiesiskās aizsardzības līdzekļu pieejamība uzņēmējdarbības un cilvēktiesību jomā ES līmenī" (2017. g.): https://fra.europa.eu/en/opinion/2017/business-human-rights.

iespējama tiesiskā aizsardzība. Īpaša uzmanība ir jāpievērš neaizsargātām personām vai grupām.

2. Tehniski un netehniski paņēmieni uzticama AI īstenošanai

(92) Lai īstenotu iepriekš minētās prasības, var izmantot gan tehniskus, gan netehniskus paņēmienus. Tie aptver visus AI sistēmas dzīves cikla posmus. Regulāri ir jāizvērtē prasību īstenošanai izmantotie paņēmieni, kā arī jāpaziņo un jāpamato[51] īstenošanas procesos izdarītās izmaiņas. Tā kā AI sistēmas nepārtraukti attīstās un darbojas dinamiskā vidē, uzticama AI īstenošana ir nepārtraukts process, kas attēlots turpmāk 5. attēlā.



3. attēls. Uzticama AI īstenošana visā sistēmas dzīves ciklā

(93) Turpmāk minētos paņēmienus var uzskatīt par savstarpēji papildinošiem vai alternatīviem, jo atšķirīgas prasības un atšķirīgi jutīgie jautājumi var radīt nepieciešamību izmantot dažādus īstenošanas paņēmienus. Šis pārskats nav jāuztver ne kā visaptverošs vai izsmeļošs, ne kā obligāts. Tā mērķis ir vienīgi piedāvāt sarakstu ar ieteicamajiem paņēmieniem, kas var palīdzēt īstenot uzticamu AI.

1. <u>Tehniskie paņēmieni</u>

(94) Šajā sadaļā ir aprakstīti tehniskie paņēmieni uzticama AI nodrošināšanai, kurus var iekļaut AI sistēmas projektēšanas, izstrādes un izmantošanas stadijās. Turpmāk uzskaitītie paņēmieni atšķiras gatavības ziņā[52].

- *Uzticama AI arhitektūra*

(95) Atbilstoši uzticama AI prasībām ir jānosaka procedūras un/vai procedūru ierobežojumi, kuri ir jānostiprina AI sistēmas arhitektūrā. To var panākt, izmantojot "baltā saraksta" noteikumus (uzvedību vai stāvokļus), kas sistēmai vienmēr jāievēro, "melnā saraksta" ierobežojumus uzvedībai vai stāvokļiem, ko sistēma nekādos apstākļos nedrīkst pārkāpt, kā arī minēto faktoru kombinācijas vai sarežģītākas pierādāmas garantijas par sistēmas uzvedību. Sistēmas atbilstību šiem ierobežojumiem tās darbības laikā var uzraudzīt, izmantojot atsevišķu procesu.

(96) AI sistēmas, kas spēj mācīties un prot dinamiski pielāgot savu uzvedību, var uzskatīt par nedeterministiskām sistēmām, kuru uzvedība var būt neparedzēta. Tās nereti tiek aplūkotas caur teorētisku cikla "uztvert–plānot–rīkoties" prizmu. Pielāgojot šo arhitektūru uzticama AI nodrošināšanai, prasības ir nepieciešams integrēt visos trijos cikla posmos: i) uztveršanas posmā sistēmai jābūt izstrādātai tā, lai tā atpazītu visus vides elementus, kas nepieciešami, lai nodrošinātu atbilstību prasībām; ii) plānošanas posmā sistēmai jāapsver tikai prasībām atbilstoši plāni; iii) rīcības posmā sistēmas darbībām jāaprobežojas tikai ar tādu uzvedību, kas īsteno prasības.

(97) Iepriekš aprakstītā arhitektūra ir vispārēja un lielāko daļu AI sistēmu apraksta neprecīzi. Tomēr tā ir atskaites

[51] Tas ietver, piemēram, pamatojumu, kādēļ izdarīta noteikta izvēle sistēmas projektēšanā, izstrādē un izvietošanā, lai iekļautu minētās prasības.

[52] Daži no šiem paņēmieniem jau ir pieejami, bet citiem vēl nepieciešami papildu pētījumi. Jomas, kurās nepieciešami tālāki pētījumi, tiks aplūkotas arī AI ALEG otrajā iesniedzamajā dokumentā, proti, ieteikumos par politiku un ieguldījumiem.

punkts ierobežojumiem un politikai, kas jāatspoguļo konkrētajos moduļos, lai iegūtu kopējo sistēmu, kas ir uzticama un kā tāda tiek uztverta.

- *Integrēta ētika un tiesiskums ("integrēts X")*

(98) Paņēmieni integrētu vērtību nodrošināšanai sniedz precīzu un tiešu saikni starp abstraktajiem principiem, kuri sistēmai jāievēro, un konkrētiem īstenošanas lēmumiem. Šā paņēmiena pamatā ir doma, ka atbilstību standartiem var īstenot jau AI sistēmas projektā. Uzņēmumi atbild par to, lai jau sākotnēji tiktu noteikta to AI sistēmu ietekme, kā arī standarti, kuri būtu jāievēro to AI sistēmām, lai izvairītos no negatīvas ietekmes. Dažādas integrēšanas koncepcijas, piemēram, *integrēta privātuma aizsardzība* un *integrēta drošība*, tiek plaši izmantotas jau tagad. Kā norādīts iepriekš, lai nopelnītu uzticēšanos, AI procesiem, datiem un rezultātiem jābūt drošiem, kā arī tam jābūt projektētam tā, lai tas būtu noturīgs pret kaitīgiem datiem un uzbrukumiem. Tam ir jāīsteno mehānisms, kā kļūdas gadījumā droši atslēgties, un jāpieļauj darbības atjaunošana pēc piespiedu atslēgšanas (piemēram, uzbrukuma).

- *Skaidrošanas paņēmieni*

(99) Lai sistēma būtu uzticama, mums ir jāspēj saprast, kādēļ tā rīkojusies noteiktā veidā un kādēļ sniegusi konkrētu interpretāciju. Ir vesela pētniecības joma – izskaidrojama AI (*XAI*) pētījumi, kas cenšas risināt šo jautājumu, lai labāk saprastu sistēmas pamata mehānismus un rastu risinājumus. Pašlaik uz neirālajiem tīkliem balstītām AI sistēmām tā joprojām ir neatrisināta problēma. Apmācības procesos ar neironu tīkliem var tikt iegūti skaitliskām vērtībām atbilstoši tīkla parametri, kuru korelācija ar rezultātiem ir grūti nosakāma. Turklāt dažreiz nelielas izmaiņas datu vērtībās var novest pie radikālām interpretācijas atšķirībām, kuru rezultātā sistēma, piemēram, sajauc skolas autobusu ar strausu. Šī ievainojamība var tikt ļaunprātīgi izmantota arī pret sistēmu vērstu uzbrukumu laikā. Paņēmieni, kas ietver *XAI* pētniecību, ir ļoti svarīgi ne tikai, lai izskaidrotu sistēmas uzvedību lietotājiem, bet arī lai ieviestu bezatteiksmes tehnoloģiju.

- *Testēšana un validācija*

(100) AI sistēmu nedeterministiskās un no konteksta atkarīgās būtības dēļ ar ierasto testēšanu vien nepietiek. Sistēmas izmantoto koncepciju un atveides kļūdas var izpausties tikai tad, kad programmu piemēro pietiekami reāliem datiem. Tādēļ, lai pārbaudītu un validētu datu apstrādi, pamatā esošais modelis gan apmācības, gan izvietošanas gaitā ir rūpīgi jāuzrauga attiecībā uz stabilitāti, noturību un darbību labi izprastās un paredzamās robežās. Ir jānodrošina, lai plānošanas procesa rezultāti atbilstu ievaddatiem un lēmumi tiktu pieņemti tā, lai būtu iespējams validēt pamatā esošo procesu.

(101) Sistēmas testēšanai un validācijai ir jānotiek iespējami agri, nodrošinot, ka sistēma visā tās dzīves ciklā, jo īpaši pēc izvietošanas, darbojas, kā plānots. Tai ir jāaptver visas AI sistēmas sastāvdaļas, ieskaitot datus, iepriekš apmācītus modeļus, vidi un sistēmas uzvedību kopumā. Tā ir jāplāno un jāizpilda iespējami daudzveidīgai cilvēku grupai. Ir jāizstrādā vairāki rādītāji, lai testējamās kategorijas aptvertu no dažādām perspektīvām. Var apsvērt uzticamu un daudzveidīgu "sarkano komandu" veiktu "naidīgo testēšanu", kuras laikā tiek mēģināts apzināti "uzlauzt" sistēmu nolūkā atrast vājās vietas, un "atlīdzību par kļūdām", kas motivē trešās personas atklāt un atbildīgi paziņot par sistēmas kļūdām un nepilnībām. Visbeidzot, ir jānodrošina, ka rezultāti vai darbības atbilst iepriekšējo procesu rezultātiem, salīdzinot tos ar iepriekš noteiktu politiku, lai pārliecinātos, ka tā netiek pārkāpta.

- *Pakalpojumu kvalitātes rādītāji*

(102) AI sistēmām var definēt attiecīgus pakalpojumu kvalitātes rādītājus, lai nodrošinātu, ka pastāv pamata izpratne par to, vai tās ir testētas un izstrādātas, ņemot vērā drošuma un drošības apsvērumus. Šie rādītāji var ietvert pasākumus algoritmu testēšanas un apmācības izvērtēšanai, kā arī ierastos programmatūras mērījumus: funkcionalitāte; veiktspēja; izmantojamība; bezatteiksme; drošība; remontējamība.

2. <u>Netehniski paņēmieni</u>

(103) Šajā sadaļā ir aprakstīti dažādi netehniski paņēmieni, kam var būt svarīga loma uzticama AI nodrošināšanā un uzturēšanā. Arī tie ir **pastāvīgi** jāizvērtē.

- *Noteikumi*

(104) Kā minēts iepriekš, AI uzticamības nodrošināšanas regulējums jau pastāv – par tādu ir uzskatāms produktu drošības un atbildības juridiskais regulējums. Ciktāl, pēc mūsu domām, šis regulējums būtu jāpārskata, jāpielāgo vai jāievieš no jauna kā aizsardzības pasākums un iespēju nodrošināšanas instruments, par to mēs izteiksimies mūsu otrajā iesniedzamajā dokumentā par ieteikumiem attiecībā uz AI politiku un ieguldījumiem.

- *Rīcības kodeksi*

(105) Organizācijas un ieinteresētās personas var apņemties izmantot vadlīnijas un pielāgot savu korporatīvās atbildības nolikumu, galvenos darbības rādītājus (*KPI*), ētikas kodeksus vai iekšējos politikas dokumentus, lai tajos iekļautu centienus panākt uzticamu AI. Organizācija, kas strādā ar AI sistēmas ieviešanu, kopumā var dokumentēt savus nodomus, kā arī garantēt tos ar standartiem, kuri atbilst noteiktām vēlamām vērtībām, piemēram, pamattiesībām, pārredzamībai un kaitējuma novēršanai.

- *Standartizācija*

(106) Standarti, piemēram, projektēšanas, ražošanas un uzņēmējdarbības jomu praksē, var kalpot kā AI lietotāju, patērētāju, organizāciju, pētniecības iestāžu un valdību kvalitātes vadības sistēma, kas sniedz iespēju atpazīt un mudināt ētisku rīcību, pieņemot lēmumus par iepirkumu. Papildus parastajiem standartiem pastāv arī kopregulācijas pieeja: akreditācijas sistēmas, profesionālie ētikas kodeksi vai pamattiesībām atbilstoša dizaina standarti. Pašlaik kā piemērus var minēt *ISO* standartus vai *IEEE* P7000 standartu sēriju, taču nākotnē varētu būt lietderīgs "uzticama AI" marķējums, kas ar atsauci uz konkrētiem tehniskiem standartiem apliecinātu, ka sistēma, piemēram, atbilst drošuma, tehniskās noturības un izskaidrojamības prasībām.

- *Sertificēšana*

(107) Tā kā nevar prasīt, lai ikviens pilnībā izprastu AI sistēmu darbību un sekas, varētu apsvērt tādu organizāciju izveidi, kas var apliecināt plašākai sabiedrībai, ka AI sistēma ir pārredzama, atbildīga un taisnīga[53]. Šāda sertifikācija piemērotu dažādām lietojuma jomām un AI paņēmieniem izstrādātus standartus, kas attiecīgi saskaņoti ar atšķirīgā kontekstā pastāvošajiem rūpnieciskajiem un sabiedrības standartiem. Tomēr sertifikācija nekad nevar aizvietot atbildību. Tādēļ tā ir jāpapildina ar atbildīguma satvariem, ieskaitot atrunas, kā arī pārskatīšanas un novēršanas mehānismiem[54].

- *Atbildīguma nodrošināšana ar pārvaldības satvariem*

(108) Organizācijām būtu jāizveido iekšējie un ārējie pārvaldes satvari, tā nodrošinot atbildību AI izstrādes, izvietošanas un izmantošanas lēmumu ētiskajos aspektos. Šajā nolūkā var, piemēram, iecelt par AI ētikas jautājumiem atbildīgu personu vai iekšēju/ārēju ētikas padomi vai komisiju. Šādas personas, padomes vai komisijas iespējamās funkcijas ietver pārraudzību un ieteikumu došanu. Kā izklāstīts iepriekš, šajā ziņā var palīdzēt arī sertifikācijas specifikācijas un/vai struktūras. Ir jānodrošina sakaru kanāli saziņai ar nozari un/vai publiskās pārraudzības grupām, lai dalītos ar labāko pieredzi, apspriestu problēmjautājumus vai ziņotu par jaunām problēmjomām, kas raisa ētiskas bažas. Šādi mehānismi var papildināt, bet nevar aizstāt juridisko pārraudzību (piemēram, datu aizsardzības speciālista iecelšanu vai līdzīgus pasākumus, kas saskaņā ar datu aizsardzības tiesībām ir juridiski obligāti).

- *Izglītība un informētība ētiska domāšanas veida veicināšanai*

---

[53] Šādu nostāju aizstāv, piemēram, *IEEE* ētikai atbilstoša dizaina iniciatīva: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

[54] Vairāk par sertifikācijas trūkumiem sk.: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

(109) Uzticams AI mudina visas ieinteresētās personas informēti līdzdarboties. Svarīga loma ir paziņošanai, izglītošanai un apmācībai – lai gan plaši izplatītu zināšanas par AI sistēmu iespējamo ietekmi, gan informētu iedzīvotājus, ka tie var piedalīties sabiedrības attīstības veidošanā. Teiktais aptver visas ieinteresētās personas, proti, ražojumu gatavošanā iesaistītās personas (projektētājus, izstrādātājus), lietotājus (uzņēmumus vai fiziskas personas) un citas skartās grupas (personas, kuras varbūt nepērk vai neizmanto AI sistēmu, bet attiecībā uz kurām lēmumus pieņem AI sistēma, un sabiedrību kopumā). Visā sabiedrībā ir jāstimulē pamata izpratne par AI. Sabiedrības izglītošanas priekšnoteikums ir pienācīgu šīs jomas ētikas speciālistu prasmju un apmācību nodrošināšana.

- *Ieinteresēto personu līdzdalība un sociālais dialogs*

(110) AI sniedz daudz ieguvumu, un Eiropai ir jānodrošina, lai tie būtu pieejami ikvienam. Tam nepieciešama atvērta diskusija un sociālo partneru, ieinteresēto personu un plašākas sabiedrības iesaiste. Daudzas organizācijas jau iesaista ieinteresēto personu grupas, lai apspriestu AI sistēmu un datu analītikas izmantošanu. Šajās grupās ietilpst dažādas personas, piemēram, juridiskie eksperti, tehniskie eksperti, ētikas speciālisti, patērētāju pārstāvji un darba ņēmēji. Rezultātu un pieejas vērtēšanā palīdz aktīva līdzdalības un dialoga meklēšana attiecībā uz AI sistēmu izmantošanu un ietekmi; tas var būt īpaši noderīgi sarežģītās lietās.

- *Daudzveidība un iekļaujošas projektētāju komandas*

(111) Reālajā dzīvē izmantotu AI sistēmu izstrādē būtiska loma ir daudzveidībai un iekļaušanai. Tā kā AI sistēmas arvien vairāk uzdevumu paveic pašas, ir ļoti svarīgi, lai kolektīvi, kas projektē, izstrādā, testē un uztur, izvieto un/vai iepērk šīs sistēmas, atspoguļotu daudzveidību lietotāju vidū un sabiedrībā kopumā. Šādi tiek stiprināta objektivitāte un dažādu perspektīvu, vajadzību un mērķu izvērtēšana. Vislabāk ir, ja kolektīvi ir daudzveidīgi ne tikai pēc dzimuma, kultūras un vecuma, bet arī profesionālās pieredzes un prasmēm.

Galvenie no II nodaļas izrietošie norādījumi

- ✓ Nodrošināt, ka viss AI sistēmas dzīves cikls atbilst uzticama AI prasībām: 1) cilvēka subjektībai un virsvadībai; 2) tehniskajai noturībai un drošumam; 3) privātumam un datu pārvaldībai; 4) pārredzamībai; 5) daudzveidībai, nediskriminēšanai un taisnīgumam; 6) vides un sabiedrības labklājībai; 7) atbildīgumam.
- ✓ Apsvērt tehniskus un netehniskus paņēmienus, ar ko nodrošināt šo prasību īstenošanu.
- ✓ Stimulēt pētniecību un inovāciju, kas palīdzētu izvērtēt AI sistēmas un veicinātu prasību izpildi; ar rezultātiem un neatbildētajiem jautājumiem iepazīstināt plašāku sabiedrību, sistēmiski gatavot jaunu mākslīgā intelekta ētikas ekspertu paaudzi.
- ✓ Skaidri un proaktīvi paziņot ieinteresētajām personām informāciju par AI sistēmu spējām un robežām, kas ļautu noteikt reāli īstenojamas gaidas, kā arī informāciju par prasību īstenošanas veidu. Skaidri norādīt, ka notiek saskarsme ar AI sistēmu.
- ✓ Atvieglot AI sistēmu trasējamību un pārbaudāmību, jo īpaši kritiskos apstākļos un situācijās.
- ✓ Iesaistīt ieinteresētās personas visā AI sistēmas dzīves ciklā. Veicināt apmācību un izglītību, lai visas ieinteresētās personas būtu informētas un apmācītas attiecībā uz uzticamu AI.
- ✓ Apzināties, ka starp dažādiem principiem un prasībām pašos pamatos var pastāvēt pretrunas. Pastāvīgi identificēt, izvērtēt un dokumentēt šos kompromisus un to risinājumus un informēt par tiem.

III. III nodaļa. Uzticama AI novērtēšana

(112) Balstoties uz galvenajām II nodaļas prasībām, šajā nodaļā ir izklāstīts neizsmeļošs **uzticama AI novērtēšanas saraksts** (izmēģinājuma redakcijā), lai **praksē iedzīvinātu uzticamu AI**. Tas jo īpaši attiecas uz AI sistēmām, kas tieši mijiedarbojas ar lietotājiem, un galvenokārt ir adresēts AI sistēmu (pašizstrādātu vai no trešām personām iegūtu) izstrādātājiem un izvietotājiem. Šis novērtēšanas saraksts neskar pirmās

uzticama AI iezīmes (likumīgs AI) īstenošanu. Atbilstība šim novērtēšanas sarakstam nepierāda juridisku atbilstību, un tas nav paredzēts kā vadlīnijas, kā nodrošināt atbilstību piemērojamajām tiesību normām. Tā kā AI sistēmas ir specifiskas katram lietojumam, novērtēšanas saraksts būs jāpielāgo konkrētiem lietojumiem un apstākļiem, kādos sistēmas darbojas. Turklāt šajā nodaļā ir vispārīgi ieteikumi par to, kā izmantot uzticama AI novērtēšanas sarakstu pārvaldības struktūrā, kas aptver kā darbības, tā vadības līmeni.

(113) Šis novērtēšanas saraksts un pārvaldības struktūra tiks izstrādāta ciešā sadarbībā ar ieinteresētajām personām publiskajā un privātajā sektorā. Process tiks virzīts kā "izmēģinājuma process", kas ļaus saņemt plašas atsauksmes no diviem paralēliem procesiem:

a. kvalitatīva procesa, kurš nodrošinās reprezentativitāti un kurā neliels skaits uzņēmumu, organizāciju un iestāžu (dažādu nozaru pārstāvji un atšķirīga lieluma) pieteiksies izmēģināt praksē novērtēšanas sarakstu un pārvaldības struktūru un sniegt padziļinātas atsauksmes;

b. kvantitatīva procesa, kurā visas ieinteresētās personas varēs pieteikties novērtēšanas saraksta izmēģināšanai praksē un sniegt atsauksmes publiskas apspriešanās ietvaros.

(114) Pēc izmēģinājuma stadijas mēs atsauksmju procesā iegūtos rezultātus integrēsim novērtēšanas sarakstā un 2020. gada sākumā sagatavosim pārskatītu versiju. Mērķis ir izveidot satvaru, ko horizontāli var izmantot visos lietojumos, tādējādi piedāvājot pamatu uzticama AI nodrošināšanai visās nozarēs. Pēc šāda pamata izstrādes būs iespējams izstrādāt nozarei vai lietojumam specifisku satvaru.

Pārvalde

(115) Uzņēmumi, organizācijas un iestādes varētu apsvērt, kā organizācijā izmantot uzticama AI novērtēšanas sarakstu. To var darīt, iekļaujot novērtēšanas procesu esošajos pārvaldības mehānismos vai ieviešot jaunus procesus. Šo izvēli nosaka organizācijas iekšējā struktūra, kā arī tās lielums un pieejamie resursi.

(116) Pētījumi[55] liecina, ka pārmaiņu panākšanai ir svarīga visaugstākā līmeņa vadības uzmanība. Tie arī pierāda, ka jauna procesa (tehnoloģiska vai nē) ieviešana tiek labāk pieņemta un ir nozīmīgāka, ja tiek iesaistītas visas uzņēmuma, organizācijas vai iestādes ieinteresētās personas[56]. Tādēļ mēs iesakām īstenot procesu, kas paredz iesaisti gan darbības līmenī, gan augstākās vadības līmenī.

Līmenis	**Attiecīgās funkcijas (atkarā no organizācijas)**
Vadība un valde	Augstākā vadība apspriež un vērtē AI izstrādi, izvietošanu vai iepirkšanu, kā arī darbojas kā augstāka instance visu AI inovāciju un lietojumu izvērtēšanai, ja tiek atklātas kritiskas problēmas. Tā visā procesa gaitā iesaista ar iespējamo AI ieviešanu skartās personas (piemēram, darba ņēmējus) un to pārstāvjus, izmantojot informēšanas, apspriešanās un līdzdalības procedūras.
Atbilstības/juridiskā nodaļa / korporatīvās atbildības nodaļa	Atbildības nodaļa uzrauga novērtēšanas saraksta izmantošanu un saraksta attīstīšanu, kas nepieciešama, lai pielāgotos tehnoloģiju vai regulatīvajām izmaiņām. Tā atjaunina AI sistēmu standartus vai iekšējo politiku un nodrošina, ka šo sistēmu izmantošana atbilst spēkā esošajam juridiskajam un regulatīvajam satvaram, kā arī organizācijas vērtībām.

[55] https://www.mckinsey.com/business-functions/operations/our-insights/secrets-of-successful-change-implementation.

[56] Sk., piem.: *A. Bryson, E. Barth, H. Dale-Olsen*, "*The Effects of Organisational change on worker well-being and the moderating role of trade unions*", *ILRReview*, 66(4), 2013. gada jūlijs; *Jirjahn, U., Smith, S. C.* (2006. g.), "*What Factors Lead Management to Support or Oppose Employee Participation–With and Without Works Councils? Hypotheses and Evidence from Germany's Industrial Relations*", 45(4), 650.–680. lpp.; *Michie, J.*, *Sheehan, M.* (2003. g.), "*Labour market deregulation, "flexibility" and innovation*", *Cambridge Journal of Economics*, 27(1), 123.–143. lpp.

Preču un pakalpojumu izstrāde vai tās ekvivalents	Preču un pakalpojumu izstrādes nodaļa izmanto novērtēšanas sarakstu, lai izvērtētu AI balstītās preces un pakalpojumus, un reģistrē visus iegūtos rezultātus. Šie rezultāti tiek apspriesti vadības līmenī, kas visbeidzot apstiprina jaunos vai pārskatītos AI balstītos lietojumus.
Kvalitātes nodrošināšana	Kvalitātes nodrošināšanas nodaļa (vai tās ekvivalents) nodrošina un pārbauda novērtēšanas saraksta rezultātus un rīkojas, lai pārsūtītu problēmu augstākā līmenī, ja rezultāti ir neapmierinoši vai ja tiek atklāti neparedzēti rezultāti.
Cilvēkresursi	Cilvēkresursu nodaļa nodrošina, ka AI sistēmu izstrādātājiem ir pareizais kompetenču kopums un profilu daudzveidība. Tā nodrošina, ka organizācijas iekšienē tiek piedāvāta pienācīga līmeņa apmācība par uzticamu AI.
Iepirkumi	Iepirkumu nodaļa nodrošina, lai uz AI balstītu preču vai pakalpojumu iepirkšanas procesā būtu ietverta AI uzticamības pārbaude.
Ikdienas darbība	Izstrādātāji un projektu vadītāji iekļauj novērtēšanas sarakstu savā ikdienas darbā un dokumentē novērtējuma rezultātus un iznākumu.

Uzticama AI novērtēšanas saraksta izmantošana

(117) Izmantojot novērtēšanas sarakstu praksē, mēs iesakām pievērst uzmanību ne tikai jomām, kas rada bažas, bet arī jautājumiem, uz kuriem nav iespējams (viegli) atbildēt. Viena no varbūtējām problēmām ir AI sistēmu izstrādājošā un testējošā kolektīva prasmju un kompetenču daudzveidības trūkums, tādēļ var būt nepieciešams iesaistīt citas ieinteresētās personas organizācijā vai ārpus tās. Ir ļoti ieteicams reģistrēt visus rezultātus gan no tehniskā, gan vadības viedokļa, tā nodrošinot, ka problēmu risināšana ir saprotama visos pārvaldības struktūras līmeņos.

(118) Novērtēšanas saraksts ir paredzēts, lai AI jomas praktiķus vadītu uzticama AI izstrādē, izvietošanā un izmantošanā. Novērtējums ir proporcionāli jāpielāgo konkrētajam izmantošanas gadījumam. Izmēģinājuma posmā var atklāties konkrētas jutīgas jomas – nākamajā posmā tiks izvērtēta nepieciešamība šiem gadījumiem noteikt papildu specifikācijas. Lai gan šajā novērtēšanas sarakstā nav konkrētu atbilžu, ko sniegt uz uzdotajiem jautājumiem, tas mudina pārdomāt pasākumus, kuri var palīdzēt nodrošināt AI sistēmu uzticamību, un iespējamās šajā sakarā veicamās darbības.

Saistība ar esošajiem tiesību aktiem un procesiem

(119) Personām, kuras ir iesaistītas AI izstrādē, izvietošanā un lietošanā, ir svarīgi arī ņemt vērā, ka pastāv dažādi esošie tiesību akti, kas paredz pienākumu veikt noteiktus procesus un aizliedz noteiktus rezultātus, un tie var pārklāties un sakrist ar dažiem novērtēšanas sarakstā minētajiem pasākumiem. Piemēram, datu aizsardzības tiesības nosaka virkni juridisko prasību, kuras jāievēro personas datu vākšanā un apstrādē iesaistītajām personām. Tomēr, tā kā uzticams AI arī obligāti paredz ētisku rīcību ar datiem, iekšējās procedūras un politika, kuras mērķis ir nodrošināt atbilstību datu aizsardzības tiesībām, var arī palīdzēt atvieglot ētisku datu apstrādi un tādējādi papildināt esošos juridiskos procesus. Tomēr atbilstība šim novērtēšanas sarakstam *nepierāda* juridisku atbilstību, un tas nav paredzēts kā vadlīnijas, kā nodrošināt atbilstību piemērojamajām tiesību normām. Šā novērtēšanas saraksta mērķis ir sniegt tā adresātiem konkrētu jautājumu kopu nolūkā nodrošināt, ka to pieeja AI izstrādei un izvietošanai ir vērsta uz uzticamu AI un cenšas to sasniegt.

(120) Tāpat daudzi AI jomas praktiķi jau ir ieviesuši novērtēšanas rīkus un programmatūras izstrādes procesus, lai nodrošinātu atbilstību arī nejuridiskiem standartiem. Turpmāk izklāstītais novērtējums nav obligāti

jāveic atsevišķi – to var iestrādāt šādā pastāvošā praksē.

UZTICAMA AI NOVĒRTĒŠANAS SARAKSTS (IZMĒĢINĀJUMA REDAKCIJA)

1. Cilvēka subjektība un virsvadība

Pamattiesības

✓ Vai jūs veicāt novērtējumu par ietekmi uz pamattiesībām tiem lietojuma gadījumiem, kas varētu negatīvi ietekmēt pamattiesības? Vai jūs identificējāt un dokumentējāt varbūtējos kompromisus starp dažādiem principiem un tiesībām?

✓ Vai AI sistēma iesaistās galalietotāju cilvēku lēmumu pieņemšanā (piem., iesaka vēlamo rīcību vai lēmumu, piedāvā izvēli)?

- Vai šajos gadījumos pastāv risks, ka AI sistēma ietekmē cilvēka patstāvību, neparedzētā veidā iejaucoties galalietotāja lēmumu pieņemšanas procesā?
- Vai jūs apsvērāt, vai AI sistēmai ir jāinformē lietotāji, ka lēmums, saturs, ieteikums vai iznākums ir algoritmiski pieņemta lēmuma rezultāts?
- Ja AI sistēmā ietilpst tērzēšanas robots vai sarunu sistēma, vai galalietotāji cilvēki tiek informēti par to, ka sazinās ar sarunas biedru, kas nav cilvēks?

Cilvēka subjektība

✓ Ja AI sistēmu izmanto darba procesā – vai apsvērāt uzdevumu sadalījumu starp AI sistēmu un darbiniekiem, lai to saskarsme būtu saturīga un pastāvētu pienācīga cilvēka veikta pārraudzība un kontrole?

- Vai AI sistēma uzlabo vai paplašina cilvēku spējas?
- Vai esat veikuši aizsardzības pasākumus, lai novērstu pārmērīgu pārliecību vai paļaušanos uz AI sistēmu darba procesos?

Cilvēka virsvadība

✓ Vai apsvērāt, kāds cilvēka vadības līmenis pienāktos konkrētajai AI sistēmai un lietojumam?

- Vai varat aprakstīt attiecīgi cilvēka kontroles vai iesaistes līmeni? Kas ir "cilvēks, kas kontrolē"? Kuros brīžos vai ar kādiem rīkiem cilvēks var iejaukties?
- Vai jūs ieviesāt mehānismus un pasākumus, lai nodrošinātu šādu iespējamu cilvēka veiktu kontroli vai pārraudzību vai lai nodrošinātu, ka par lēmumu pieņemšanu kopumā ir atbildīgs cilvēks?
- Vai veicāt pasākumus, kas nodrošina iespēju revidēt un novērst ar AI patstāvības pārvaldīšanu saistītās problēmas?

✓ Ja AI sistēma vai lietojums ir saistīts ar pašmācību vai ir autonoms – vai ieviesāt konkrētākus kontroles un pārraudzības mehānismus?

- Kādus atklāšanas un reaģēšanas mehānismus jūs izveidojāt, lai izvērtētu, vai var notikt kaut kas

neparedzēts?

- Vai jūs nodrošinājāt apturēšanas iespēju vai procedūru, lai droši atceltu darbību, kad tas nepieciešams? Vai šī procedūra aptur procesu pavisam, daļēji vai nodod kontroli cilvēkam?

2. <u>Tehniskā noturība un drošums</u>

Noturība pret uzbrukumiem un drošība

✓ Vai izvērtējāt varbūtējās uzbrukuma formas, pret kurām AI sistēma varētu būt neaizsargāta?

- Pats svarīgākais – vai apsvērāt dažādus ievainojamības veidus un būtību, piemēram, datu piesārņošanu, fizisko infrastruktūru, kiberuzbrukumus?

✓ Vai ieviesāt pasākumus vai sistēmas, lai nodrošinātu, ka AI sistēma ir neskarta un noturīga pret iespējamiem uzbrukumiem?

✓ Vai izvērtējāt, kā jūsu sistēma rīkojas negaidītās situācijās un vidēs?

✓ Vai apsvērāt to, vai un cik lielā mērā jūsu sistēmai var būt divējāds lietojums? Vai šādā gadījumā veicāt pienācīgus profilakses pasākumus (tostarp, piemēram, nepublicējāt pētniecisko darbu vai neizvietojāt sistēmu)?

Rezerves plāns un vispārējais drošums

✓ Vai nodrošinājāt, ka jūsu sistēmai ir pienācīgs rezerves plāns gadījumam, ja tā saskartos ar naidīgiem uzbrukumiem vai citām neparedzētām situācijām (piem., tehniskas pārslēgšanās procedūras vai operatora–cilvēka izsaukšana pirms turpmākas rīcības)?

✓ Vai apsvērāt riska līmeni, ko konkrētajā lietojumā rada AI sistēma?

- Vai ieviesāt riska un drošības mērīšanas un vērtēšanas procesu?
- Vai sniedzāt nepieciešamo informāciju, ja pastāv risks cilvēku fiziskajai integritātei?
- Vai apsvērāt apdrošināšanas polises iegādi, lai nodrošinātos pret iespējamo AI sistēmas nodarīto kaitējumu?
- Vai identificējāt iespējamo drošības risku, ko rada (citi) paredzami tehnoloģijas izmantošanas veidi, ieskaitot tās nejaušu vai tīšu ļaunprātīgu izmantošanu? Vai pastāv plāns šā riska mazināšanai vai vadībai?

✓ Vai izvērtējāt, vai pastāv varbūtība, ka AI sistēma var nodarīt zaudējumus vai kaitējumu tās lietotājiem vai trešām personām? Vai šādā gadījumā izvērtējāt, kāda ir iespējamība, varbūtējais kaitējums, skartās personas un smagums?

- Ja pastāv risks, ka AI sistēma var nodarīt kaitējumu, vai jūs apsvērāt atbildības un patērētāju tiesību aizsardzības noteikumus un kā ņēmāt tos vērā?
- Vai apsvērāt iespējamo ietekmi vai drošības risku, kāds tiek radīts videi vai dzīvniekiem?

- Vai jūsu riska analīzē tika vērtēts, vai drošības vai tīkla problēmas (piemēram, kiberdrošības draudi) rada drošības vai kaitējuma risku AI sistēmas netīšas uzvedības dēļ?

✓ Vai aplēsāt tādas jūsu AI sistēmas kļūdas varbūtējo ietekmi, kuras dēļ tiek sniegti nepareizi rezultāti vai jūsu sistēma kļūst nepieejama, vai sistēma sniedz sociāli nepieņemamus (piemēram, diskriminējošus) rezultātus?

- Vai noteicāt minēto scenāriju robežvērtības un pārvaldību attiecībā uz to, kad iestājas alternatīvie vai rezerves plāni?
- Vai definējāt un testējāt rezerves plānus?

Precizitāte

✓ Vai izvērtējāt, kāds precizitātes līmenis un definīcija ir nepieciešama, ņemot vērā AI sistēmu un lietojumu?

- Vai izvērtējāt, kā tiek mērīta un nodrošināta precizitāte?
- Vai ieviesāt pasākumus, lai nodrošinātu, ka izmantotie dati ir visaptveroši un aktuāli?
- Vai ieviesāt pasākumus, lai izvērtētu, vai nepieciešami papildu dati, piemēram, lai uzlabotu precizitāti vai novērstu neobjektivitāti?

✓ Vai izvērtējāt kaitējumu, kas tiks nodarīts, ja AI sistēma sniegs neprecīzas prognozes?

✓ Vai ieviesāt mehānismus, ar ko mērīt, vai jūsu sistēma nepieņemami daudz reižu neprognozē kļūdaini?

✓ Ja tiktu sniegtas kļūdainas prognozes, vai jūs ieviesāt pasākumu kompleksu problēmas risināšanai?

Bezatteiksme un reproducējamība

✓ Vai ieviesāt stratēģiju, lai uzraudzītu un pārbaudītu, vai AI sistēma atbilst uzdevumiem, mērķiem un paredzētajiem lietojumiem?

- Vai pārbaudījāt, vai reproducējamības nodrošināšanai ir jāņem vērā konkrēts konteksts vai īpaši nosacījumi?
- Vai ieviesāt verifikācijas procesus vai metodes, lai mērītu un nodrošinātu dažādus bezatteiksmes un reproducējamības aspektus?
- Vai ieviesāt procesus, kā aprakstīt, ka AI sistēma dažos uzstādījumu apstākļos nedarbojas pareizi?
- Vai skaidri dokumentējāt un precizējāt šos procesus, kas paredzēti AI sistēmu bezatteiksmes testēšanai un verificēšanai?

  Vai ieviesāt mehānismus vai informācijas sniegšanu, lai pārliecinātu (gala) lietotājus par AI sistēmas bezatteiksmi?

3. <u>Privātums un datu pārvaldība</u>

Privātuma un datu aizsardzības ievērošana

✓ Vai atkarā no lietojuma izveidojāt mehānismu, kas ļauj citām personām atzīmēt ar privātumu vai datu

aizsardzību saistītas problēmas, kuras skar AI sistēmas datu vākšanas (apmācībām un darbībai) un datu apstrādes procesus?

- ✓ Vai vērtējāt jūsu datu kopās esošo datu veidu un apjomu (piem., vai tie satur personas datus)?
- ✓ Vai apsvērāt, kā izstrādāt AI sistēmu vai apmācīt modeli, neizmantojot vai iespējami maz izmantojot potenciāli sensitīvus datus vai personas datus?
- ✓ Vai ietvērāt mehānismus, kā paziņot un kontrolēt personas datus atkarā no lietojuma veida (piemēram, attiecīgā gadījumā – derīgu piekrišanu un iespēju to atsaukt)?
- ✓ Vai veicāt pasākumus privātuma uzlabošanai, piemēram, izmantojot šifrēšanu, anonimizāciju un apkopošanu?
- ✓ Ja jums ir datu aizsardzības speciālists (*DPO*), vai jūs šo personu iesaistījāt procesa agrīnā stadijā?

Datu kvalitāte un integritāte

- ✓ Vai pielāgojāt savu sistēmu iespējamajiem attiecīgajiem standartiem (piem., *ISO*, *IEEE*) vai plaši izmantotiem protokoliem datu ikdienas organizēšanas un pārvaldības jomā?
- ✓ Vai izveidojāt datu vākšanas, uzglabāšanas, apstrādes un izmantošanas pārraudzības mehānismus?
- ✓ Vai izvērtējāt, ciktāl jūs kontrolējat ārējo izmantoto datu avotu kvalitāti?
- ✓ Vai ieviesāt procesus jūsu datu kvalitātes un integritātes nodrošināšanai? Vai apsvērāt citus procesus? Kā pārbaudāt, vai jūsu datu kopas nav tikušas apdraudētas vai uzlauztas?

Piekļuve datiem

- ✓ Kādi protokoli, procesi un procedūras tika ievērotas, lai vadītu un nodrošinātu pienācīgu datu pārvaldību?
  - Vai izvērtējāt, kas un kādos gadījumos var piekļūt lietotāju datiem?
  - Vai nodrošinājāt, ka šīs personas ir kvalificētas un tām ir pienākums piekļūt datiem, un tām ir nepieciešamā kompetence, lai detalizēti saprastu datu aizsardzības politiku?
  - Vai nodrošinājāt pārraudzības mehānismu, kas reģistrē, kad, kur, kā, kurš un ar kādu mērķi piekļuvis datiem?

4. <u>Pārredzamība</u>

Trasējamība

- ✓ Vai ieviesāt pasākumus, kas var nodrošināt trasējamību? Tas var ietvert dokumentāciju attiecībā uz:
  - algoritmu sistēmas projektēšanas un izstrādes paņēmieniem:
    - kārtulbalstītai AI sistēmai ir jādokumentē programmēšanas paņēmieni vai modeļa veidošana;
    - uz mācīšanos balstītai AI sistēmai ir jādokumentē algoritma apmācību metode, tostarp tas, kuri ievaddati ir apkopoti un izvēlēti un kā tas noticis;

- algoritmu sistēmas testēšanas un validēšanas paņēmieniem:
  - kārtulbalstītai AI sistēmai ir jādokumentē testēšanā un validēšanā izmantotie scenāriji vai gadījumi;
  - uz mācīšanos balstītai AI sistēmai ir jādokumentē informācija par testēšanā un validēšanā izmantotajiem datiem;
- algoritmu sistēmas rezultātiem:
  - ir jādokumentē algoritma rezultāti vai tā pieņemtie lēmumi, kā arī citi varbūtējie lēmumi, kas tiktu pieņemti citos gadījumos (piemēram, attiecībā uz citām lietotāju apakšgrupām).

Izskaidrojamība

- ✓ Vai izvērtējāt, ciktāl ir iespējams saprast AI sistēmas lēmumus un attiecīgi – tās rezultātus?
- ✓ Vai nodrošinājāt, ka paskaidrojums par to, kādēļ sistēma pieņēmusi konkrētu lēmumu un iegūts konkrēts rezultāts, var tikt padarīts saprotams visiem lietotājiem, kas vēlētos to saņemt?
- ✓ Vai izvērtējāt, cik lielā mērā sistēmas lēmums ietekmē organizācijas lēmumu pieņemšanas procesus?
- ✓ Vai izvērtējāt, kādēļ šī konkrētā sistēma ieviesta šajā konkrētajā jomā?
- ✓ Vai izvērtējāt šīs sistēmas darbības modeli (piem., kā tā rada vērtību organizācijai)?
- ✓ Vai jau no paša sākuma projektējāt AI sistēmu, ņemot vērā interpretējamību?
  - Vai pētījāt un centāties izmantot vienkāršāko un vislabāk interpretējamo modeli, kas iespējams attiecīgajam lietojumam?
  - Vai izvērtējāt, vai jūsu apmācību un testa dati var tikt analizēti? Vai laika gaitā to varat mainīt un atjaunināt?
  - Vai izvērtējāt, vai pēc modeļa apmācības un izstrādes būs iespējams izvērtēt izskaidrojamību un vai jums ir piekļuve modeļa iekšējai darba plūsmai?

Paziņošana

- ✓ Vai jūs, izmantojot atrunu vai citādi, (gala) lietotājiem paziņojāt, ka tie sazinās ar AI sistēmu, nevis ar citu personu? Vai AI sistēmu attiecīgi apzīmējāt?
- ✓ Vai ieviesāt mehānismus lietotāju informēšanai par AI sistēmas rezultātu iemesliem un kritērijiem?
  - Vai tas attiecīgajiem lietotājiem ir paziņots skaidri un saprotami?
  - Vai izveidojāt procesus, kas paredz ņemt vērā lietotāju atsauksmes un izmantot tās, lai pielāgotu sistēmu?
  - Vai ziņojāt arī par iespējamu vai šķietamu risku, piemēram, neobjektivitāti?
  - Vai atkarā no lietojuma arī apsvērāt informēšanu un pārredzamību par labu citai auditorijai, trešām personām vai plašākai sabiedrībai?
- ✓ Vai skaidri izklāstījāt AI sistēmas mērķi un to, kas var gūt labumu no preces vai pakalpojuma?
  - Vai ražojuma lietošanas scenāriji ir noteikti un skaidri paziņoti, apsverot arī alternatīvas saziņas

formas, lai nodrošinātu, ka informācija ir saprotama un piemērota uzrunātajam lietotājam?

- Atkarā no lietojuma – vai ņēmāt vērā cilvēku psiholoģiju un iespējamos ierobežojumus, piemēram, sajaukšanas risku, tendenciozitāti vai kognitīvo nogurumu?

✓ Vai skaidri informējāt par AI sistēmas īpašībām, ierobežojumiem un varbūtējiem trūkumiem:

- izstrādes gadījumā: personu, kas sistēmu ievieš kā preci vai pakalpojumu?
- izvietošanas gadījumā: galalietotāju vai patērētāju?

5. Daudzveidība, nediskriminēšana un taisnīgums

Izvairīšanās no netaisnīgas neobjektivitātes

✓ Vai nodrošinājāt stratēģiju vai procedūru kopumu, lai izvairītos no netaisnīgas neobjektivitātes radīšanas vai nostiprināšanas AI sistēmā gan attiecībā uz ievaddatu izmantošanu, gan algoritmu izstrādi?

- Vai izvērtējāt un apzinājāties iespējamos ierobežojumus, kas izriet no izmantoto datu kopu satura?
- Vai apsvērāt datu daudzveidību un lietotāju pārstāvību? Vai veicāt testēšanu attiecībā uz konkrētām kopām vai problemātiskiem izmantošanas gadījumiem?
- Vai pētījāt un izmantojāt pieejamos tehniskos līdzekļus, lai uzlabotu jūsu izpratni par datiem, modeli un rezultātiem?
- Vai ieviesāt procesus, kā sistēmas izstrādes, izvietošanas un izmantošanas stadijā testēt un uzraudzīt varbūtējo neobjektivitāti?

✓ Vai atkarā no lietojuma nodrošinājāt mehānismu, kas ļauj citām personām atzīmēt ar AI sistēmas neobjektivitāti, diskrimināciju vai vāju veiktspēju saistītas problēmas?

- Vai apsvērāt skaidrus pasākumus un saziņas veidus attiecībā uz to, kā un kam adresēt sūdzības par šīm problēmām?
- Vai ņēmāt vērā ne tikai (gala)lietotājus, bet arī citas personas, ko netieši varētu skart AI sistēma?

✓ Vai izvērtējāt, vai vienos un tajos pašos apstākļos var rasties jebkādas iespējamas lēmumu atšķirības?

- Ja jā – vai izvērtējāt, kādi varētu būt iespējamie cēloņi?
- Ja pastāv nenoteiktība – vai izveidojāt mehānismu, kā mērīt vai novērtēt šādas nenoteiktības iespējamo ietekmi uz pamattiesībām?

✓ Vai noteicāt piemērotu "taisnīguma" praktisko definīciju, ko jūs piemērojat AI sistēmu projektēšanā?

- Vai jūsu definīcija ir vispārlietota? Vai apsvērāt citas definīcijas izvēli pirms šīs?
- Vai nodrošinājāt kvantitatīvo analīzi vai mērījumus, lai mērītu un testētu piemēroto taisnīguma definīciju?
- Vai izveidojāt mehānismus jūsu AI sistēmu taisnīguma nodrošināšanai? Vai apsvērāt citus

varbūtējos mehānismus?

Pieejamība un universālais dizains

- ✓ Vai nodrošinājāt, ka AI sistēma ir piemērota dažādām indivīdu vēlmēm un spējām?
  - Vai izvērtējāt, vai AI sistēmu var izmantot personas ar īpašām vajadzībām vai invaliditāti vai atstumtības riskam pakļautas personas? Kā tas tika ieprojektēts sistēmā, kā to verificē?
  - Vai nodrošinājāt, ka informācija par AI sistēmu ir pieejama arī palīgtehnoloģiju lietotājiem?
  - Vai iesaistījāt vai apspriedāties ar šo kopienu AI sistēmas izstrādes stadijā?
- ✓ Vai ņemāt vērā jūsu AI sistēmas ietekmi uz potenciālo lietotāju auditoriju?
  - Vai AI sistēmas veidošanā iesaistītais kolektīvs ir reprezentatīvs attiecībā uz mērķauditoriju? Vai tas ir reprezentatīvs attiecībā uz iedzīvotājiem kopumā, ņemot vērā arī citas grupas, kas var tikt skartas pastarpināti?
  - Vai izvērtējāt, vai nav tādu personu vai grupu, kuras nesamērīgi ietekmētu negatīvas sekas?
  - Vai saņēmāt atsauksmes no citiem kolektīviem vai grupām, kas pārstāv citu izcelsmi un dzīves pieredzi?

Ieinteresēto personu dalība

- ✓ Vai apsvērāt mehānismu, kā nodrošināt, lai AI sistēmas izstrādē un lietošanā piedalās dažādas ieinteresētās personas?
- ✓ Vai sagatavojāties AI sistēmas ieviešanai jūsu organizācijā, iepriekš par to informējot un iesaistot skartos darbiniekus un to pārstāvjus?

6. Sabiedrības un vides labklājība

Ilgtspējīgs un videstīgs AI

- ✓ Vai ieviesāt mehānismus, lai mērītu AI sistēmas izstrādes, izvietošanas un izmantošanas ietekmi uz vidi (piem., datu centra patērēto enerģiju, datu centra patērētās enerģijas veidu utt.)?
- ✓ Vai nodrošinājāt pasākumus, kas mazina jūsu AI sistēmas dzīves cikla ietekmi uz vidi?

Sociālā ietekme

- ✓ Ja ar AI sistēmu tieši mijdarbojas cilvēki:
  - Vai izvērtējāt, vai AI sistēma mudina cilvēkus attīstīt pieķeršanos un empātiju pret sistēmu?
  - Vai nodrošinājāt, ka AI sistēma nepārprotami signalizē, ka sociālā saskarsme ar to ir simulēta un ka tā nespēj "saprast" vai "just"?
- ✓ Vai nodrošinājāt, ka AI sistēmas sociālā ietekme ir labi izprasta? Piemēram, vai izvērtējāt, vai pastāv darba vietu zaudēšanas vai darbinieku prasmju pasliktināšanās risks? Kādi pasākumi ir veikti, lai cīnītos pret šo risku?

Sabiedrība un demokrātija

- ✓ Vai izvērtējāt AI sistēmas izmantošanas plašāko sociālo ietekmi papildus ietekmei uz atsevišķajiem (gala) lietotājiem, piemēram, iespējami netieši skartās ieinteresētās personas?

7. <u>Atbildīgums</u>

Pārbaudāmība

- ✓ Vai ieviesāt mehānismus, kas iekšējiem un/vai neatkarīgiem revidentiem atvieglo sistēmas pārbaudāmību, piemēram, nodrošinot AI sistēmas procesu un rezultātu trasējamību un žurnalēšanu?

Negatīvās ietekmes mazināšana un ziņošana par to

- ✓ Vai veicāt AI sistēmas riska vai ietekmes novērtējumu, ņemot vērā dažādās tieši un netieši skartās ieinteresētās personas?
- ✓ Vai ieviesāt apmācības un izglītošanas sistēmu, lai attīstītu atbildīguma praksi?
  - Kuri darbinieki vai kolektīva nodaļas tiek iesaistītas? Vai apmācība turpinās pēc izstrādes stadijas?
  - Vai šajās apmācībās tiek mācīts arī iespējamais AI sistēmai piemērojamais tiesiskais regulējums?
  - Vai apsvērāt "ētiska AI pārraudzības komisijas" izveidi vai līdzīgu mehānismu ar mērķi apspriest kopējo atbildīguma un ētikas praksi, ieskaitot iespējami neskaidrās "pelēkās zonas"?
- ✓ Vai papildus iekšējām iniciatīvām vai satvariem ētikas un atbildīguma pārraudzībai pastāv kādi ārējie norādījumi, vai ir izveidoti revīzijas procesi?
- ✓ Vai ir izveidoti procesi, kā trešās personas (piem., piegādātāji, patērētāji, izplatītāji/pārdevēji) vai darbinieki var ziņot par iespējamu ievainojamību, risku vai neobjektivitāti AI sistēmā vai lietojumā?

Kompromisu dokumentēšana

- ✓ Vai izveidojāt mehānismu, ar ko noteikt attiecīgās intereses un vērtības, ko skar AI sistēma, un to varbūtējos savstarpējos kompromisus?
- ✓ Kādu procesu jūs izmantojat, lai pieņemtu lēmumu par šādiem kompromisiem? Vai jūs nodrošinājāt, ka lēmums par kompromisu tiek dokumentēts?

Iespējas izmantot tiesisko aizsardzību

- ✓ Vai izveidojāt pienācīgu mehānismu kopumu, kas ļauj izmantot tiesisko aizsardzību, ja nodarīts jebkāds kaitējums vai radusies nelabvēlīga ietekme?
- ✓ Vai ieviesāt mehānismu, kā (gala)lietotājiem un trešām personām sniegt informāciju par tiesiskās aizsardzības iespējām?

Aicinām visas ieinteresētās personas praksē aprobēt šo novērtēšanas sarakstu un sniegt atsauksmes par tā īstenojamību, pilnīgumu, atbilstību konkrētam AI lietojumam vai jomai, kā arī pārklāšanos vai papildināmību ar esošiem atbilstības vai vērtējuma procesiem. Balstoties uz atsauksmēm, Komisijai 2020. gada sākumā tiks piedāvāta uzticama AI novērtēšanas saraksta pārskatīta redakcija.

Galvenie no III nodaļas izrietošie norādījumi

- ✓ Izstrādājot, izvietojot vai izmantojot AI, pieņemt uzticama AI **novērtēšanas sarakstu** un pielāgot to konkrētajam sistēmas izmantošanas gadījumam.
- ✓ Ņemt vērā, ka šāds novērtēšanas saraksts **nekad nebūs izsmeļošs**. Uzticama AI nodrošināšana nav ķeksīšu ievilkšana, bet gan pastāvīga prasību identificēšana, risinājumu izvērtēšana un uzlabotu iznākumu nodrošināšana visā AI sistēmas dzīves ciklā, kā arī ieinteresēto personu iesaiste šajās darbībās.

C. AI RADĪTO IESPĒJU UN KRITISKO PROBLĒMJOMU PIEMĒRI

(121) Šajā sadaļā esam snieguši AI izstrādes un izmantošanas piemērus, kuri būtu jāatbalsta, kā arī tādu gadījumu piemērus, kad AI izstrāde, izvietošana vai izmantošana var nonākt pretrunā mūsu vērtībām un var radīt konkrētas bažas. Ir jāpanāk līdzsvars attiecībā uz to, ko var un ko vajag darīt, izmantojot AI, un pienācīgi jāņem vērā, ko nedrīkstētu darīt ar AI.

1. Uzticama AI radītu iespēju piemēri

(122) Uzticams AI var sniegt lielisku iespēju palīdzēt mazināt tādas sabiedrību apdraudošas aktuālas problēmas kā sabiedrības novecošana, augoša sociālā nevienlīdzība un vides piesārņojums. Šis potenciāls tiek atspoguļots arī pasaules mērogā, piemēram, ANO ilgtspējīgas attīstības mērķos[57]. Nākamajā sadaļā tiek apsvērts, kā veicināt tādu Eiropas AI stratēģiju, kas risina dažus no šiem jautājumiem.

a. Ar klimatu saistītās darbības un ilgtspējīga infrastruktūra

(123) Lai gan cīņai pret klimata pārmaiņām būtu jābūt pasaules politikas veidotāju galvenajam prioritārajam uzdevumam, digitālajai pārveidei un uzticamam AI ir liels potenciāls samazināt cilvēku ietekmi uz vidi un ļaut efektīvi un lietderīgi izmantot enerģiju un dabas resursus[58]. Piemēram, uzticams AI var tikt sasaistīts ar lielo datu tehnoloģiju, lai precīzāk konstatētu enerģijas pieprasījumu, tā iegūstot efektīvāku energoinfrastruktūru un enerģijas patēriņu[59].

(124) Tādās nozarēs kā sabiedriskais transports intelektiskām transporta sistēmām paredzētas AI sistēmas[60] var izmantot, lai mazinātu rindas, uzlabotu maršrutēšanu, ļautu cilvēkiem ar redzes traucējumiem būt neatkarīgākiem[61], optimizētu energoefektīvus dzinējus un tādējādi stiprinātu dekarbonizācijas centienus un mazinātu vides pēdas nospiedumu zaļākas sabiedrības vārdā. Patlaban pasaulē ik pēc 23 sekundēm kāds cilvēks iet bojā autoavārijā[62]. AI sistēmas varētu palīdzēt būtiski samazināt nāves gadījumu skaitu, piemēram, nodrošinot labāku reakcijas laiku un precīzāku noteikumu ievērošanu[63].

b. Veselības aizsardzība un labklājība

---

57 https://sustainabledevelopment.un.org/?menu=1300.

58 Vairāku ES projektu mērķis ir attīstīt viedos energotīklus un enerģijas uzglabāšanu, kam ir potenciāls dot ieguldījumu sekmīgā digitāli atbalstītā enerģētikas pārkārtošanā, tostarp, izmantojot AI balstītus un citus digitālos risinājumus. Lai papildinātu šo atsevišķo projektu darbu, Komisija ir sākusi iniciatīvu *BRIDGE*, ļaujot pašreiz notiekošajiem programmas "Apvārsnis 2020" viedā energotīkla un enerģijas uzglabāšanas projektiem veidot kopīgu redzējumu par starpnozaru jautājumiem: https://www.h2020-bridge.eu/.

59 Sk., piemēram, projektu *Encompass*: http://www.encompass-project.eu/.

60 Jauni AI balstīti risinājumi palīdz pilsētām sagatavoties nākotnes mobilitātei. Sk., piemēram, ES finansēto projektu *Fabulos*: https://fabulos.eu/.

61 Sk., piemēram, projektu *PRO4VIP*, kas ir daļa no ES 2020. gada redzes stratēģijas cīņai ar novēršamu aklumu, jo īpaši vecuma izraisītu aklumu. Mobilitāte un orientēšanās bija starp projekta prioritārajām jomām.

62 https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

63 Piemēram, Eiropas projekta *UP-Drive* mērķis ir risināt minētās ar transportu saistītās problēmas, sniedzot ieguldījumu, kas ļautu pakāpeniski automatizēt transportlīdzekļus un tiem savstarpēji sadarboties, tādējādi veicinot drošāku, iekļaujošāku un pieejamāku transporta sistēmu: https://up-drive.eu/.

(125) Uzticama AI tehnoloģijas var izmantot (un tās jau izmanto), lai nodrošinātu ārstēšanu viedāk un mērķtiecīgāk un palīdzētu novērst slimības, kas apdraud dzīvību[64]. Ārsti un veselības aprūpes speciālisti varētu precīzāk un detalizētāk analizēt sarežģītus pacienta veselības datus vēl pirms tam, kad cilvēki ir saslimuši, un sniegt pielāgotu preventīvo aprūpi[65]. Eiropas novecojošās sabiedrības kontekstā AI un robotika var būt vērtīgi instrumenti, kas palīdzētu aprūpētājiem un sniegtu atbalstu senioru aprūpē[66], kā arī reāllaikā uzraudzītu pacientu stāvokli, tā glābjot dzīvības[67].

(126) Uzticams AI var palīdzēt arī plašākā mērogā. Tas, piemēram, var izvērtēt un identificēt vispārējās tendences veselības aprūpes un ārstniecības nozarē[68], kas ļautu agrāk atklāt slimības, efektīvāk izstrādāt zāles, vairāk pielāgot ārstēšanu[69] un galu galā izglābt vairāk dzīvību.

c. Pilnvērtīga izglītība un digitālā pārveide

(127) Jaunas tehnoloģiskās, ekonomiskās un vidiskās izmaiņas nozīmē, ka sabiedrībai ir jākļūst proaktīvākai. Valdības, nozaru līderi, izglītības iestādes un arodbiedrības saskaras ar pienākumu ievest iedzīvotājus jaunajā digitālajā laikmetā, nodrošinot, lai tiem būtu pareizās prasmes nākotnes darbvietu aizpildīšanai. Uzticama AI tehnoloģijas varētu palīdzēt, precīzāk prognozējot, kuras darbvietas un profesijas ietekmēs tehnoloģijas, kādas jaunas funkcijas tiks radītas un kādas prasmes būs nepieciešamas. Tas varētu palīdzēt valdībām, arodbiedrībām un nozarei plānot darbinieku (atkārtotu) kvalifikācijas celšanu. Arī cilvēkiem, kuri, iespējams, baidās zaudēt darbu, šādi varētu tikt piedāvāts ceļš uz jauna amata apgūšanu.

(128) AI var būt arī lielisks rīks cīņā pret nevienlīdzību izglītības jomā un izveidot personalizētas un pielāgojamas izglītības programmas, kas varētu palīdzēt ikvienam iegūt jaunu kvalifikāciju, prasmes un kompetences atbilstoši spējai mācīties[70]. Šādi tiktu uzlabots gan mācīšanās ātrums, gan izglītības kvalitāte no pamatskolas līdz universitātei.

2. AI kritisko problēmjomu piemēri

(129) Kritiskas bažas rodas, ja ir pārkāpta kāda no uzticama AI iezīmēm. Daudzas turpmāk uzskaitītās problēmjomas

---

[64] Sk., piemēram, projektu *REVOLVER* (atkārtota audzēja attīstība): https://www.healtheuropa.eu/personalised-cancer-treatment/87958/ vai projektu *Murab*, kurš veic precīzākas biopsijas un kura mērķis ir agrāk diagnosticēt audzējus un citas slimības: https://ec.europa.eu/digital-single-market/en/news/murab-eu-funded-project-success-story.

[65] Sk., piemēram, projektu *Live INCITE*: www.karolinska.se/en/live-incite. Šis veselības aprūpes nozares ražotāju konsorcijs aicina nozari izstrādāt viedo AI un citus IKT risinājumus, kas ļauj perioperatīvajā procesā ietekmēt dzīvesveidu. Mērķis ir saistīts ar jauniem novatoriskiem e-veselības risinājumiem, kas var personalizēti ietekmēt pacientus, lai tie savā dzīvesveidā ietvertu nepieciešamās darbības pirms un pēc operācijas, lai panāktu optimālu veselības aprūpes rezultātu.

[66] ES finansētais projekts *CARESSES* ir saistīts ar senioru aprūpes robotiem un īpaši pievēršas kultūrsensitīvajiem aspektiem – roboti pielāgo savu izturēšanos un runu, lai atbilstu aprūpējamā vecā cilvēka kultūrai un paradumiem: http://caressesrobot.org/en/project/. Sk. arī AI lietotni *Alfred* – virtuālu asistentu, kas palīdz veciem cilvēkiem saglabāt aktīvu dzīvesveidu: https://ec.europa.eu/digital-single-market/en/news/alfred-virtual-assistant-helping-older-people-stay-active. Turklāt projektā *EMPATTICS* (*EMpowering PAtients for a BeTTer Information and improvement of the Communication Systems*) tiks pētīts un definēts, kā veselības aprūpes speciālisti un pacienti izmanto IKT tehnoloģijas, tostarp AI sistēmas, lai plānotu darbības ar pacientiem un uzraudzītu to fiziskā un garīgā stāvokļa izmaiņas: www.empattics.eu.

[67] Sk., piemēram, *MyHealthAvatar* (www.myhealthavatar.eu), kas digitāli atveido pacienta veselības stāvokli. Pētniecības projektā darbu sāka lietotne un tiešsaistes platforma, kas apkopo datus un ļauj piekļūt jūsu digitalizētajai ilgtermiņa informācijai par veselības stāvokli. Tā izpaužas kā visu mūžu pastāvošs biedrs ("avatārs") veselības aprūpes jomā. *MyHealthAvatar* arī prognozē, kāds ir jūsu risks ciest no insulta, diabēta, sirds un asinsvadu slimībām un hipertensijas.

[68] Sk., piemēram, projektu *ENRICHME* (www.enrichme.eu), kas cīnās ar novecojošo iedzīvotāju pakāpenisko kognitīvo spēju zudumu. Integrēta interaktīvas automatizētas dzīvesvides (IAD) platforma un mobils apkalpošanas robots ilgtermiņa uzraudzībai un kontaktiem palīdzēs veciem cilvēkiem ilgāk saglabāt neatkarību un būt aktīviem.

[69] Sk., piemēram, kā mākslīgo intelektu izmanto *Sophia Genetics*, līdzsvarojot statistikas secinājumus, tendenču atpazīšanu un mašīnu mācīšanos, lai maksimāli palielinātu genomikas un radiomikas datu vērtību: https://www.sophiagenetics.com/home.html.

[70] Sk., piemēram, projektu *MaTHiSiS*, kura mērķis ir piedāvāt risinājumu normatīvajai apmācībai ērtā mācīšanās vidē, ko veido augsto tehnoloģiju ierīces un algoritmi: (http://mathisis-project.eu/). Sk. arī *IBM Watson Classroom* vai *Century Tech* platformu.

jau ietilpst tādu esošo juridisko prasību tvērumā, kas ir obligātas un tādēļ jāievēro. Tomēr arī gadījumos, kad ir pierādīta juridisko prasību ievērošana, tās var neatrisināt visas ētiskās bažas, kas varētu rasties. Tā kā mūsu izpratne par noteikumu un ētikas principu piemērotību vienmēr attīstās un laika gaitā var mainīties, tālāk izklāstītais neizsmeļošais problēmjomu saraksts nākotnē var tikt saīsināts, paplašināts, labots vai atjaunināts.

a. Fizisku personu identificēšana un izsekošana, izmantojot AI

(130) AI ļauj gan publiskām, gan privātām struktūrām arvien efektīvāk identificēt atsevišķas personas. Vērā ņemami mērogojamas AI identifikācijas piemēri ir sejas atpazīšana un citas piespiedu identifikācijas metodes, kas izmanto biometriskos datus (piemēram, melu atpazīšana, personības vērtēšana ar mikroizpausmju palīdzību, automātiska balss atpazīšana). Dažkārt fizisku personu identificēšana ir vēlama un atbilst ētikas principiem (piemēram, ja to izmanto, lai atklātu krāpšanu, nelikumīgi iegūtu līdzekļu legalizāciju vai terorisma finansēšanu). Tomēr automātiska identificēšana raisa būtiskas juridiskas un ētiskas bažas, jo tai var būt neparedzēta ietekme daudzos psiholoģiskos un sociālkulturālos līmeņos. Lai nosargātu Eiropas pilsoņu patstāvību, ir nepieciešams samērīgi izmantot AI kontroles paņēmienus. Uzticama AI sasniegšanai būs ļoti svarīgi skaidri definēt, vai, kad un kā AI var izmantot, lai automātiski identificētu cilvēkus, un nošķirt personas identificēšanu no tās izsekošanas, kā arī mērķtiecīgu uzraudzību no masveida novērošanas. Šādu tehnoloģiju izmantošana ir skaidri jāpamato saskaņā ar spēkā esošajiem tiesību aktiem[71]. Ja darbību pamats ir piekrišana, ir jāizstrādā praktiski līdzekļi[72], kas ļauj sniegt jēgpilnu un pārbaudītu piekrišanu automātiskai identifikācijai ar AI vai līdzvērtīgu tehnoloģiju. Tas attiecas arī uz "anonīmu" personas datu izmantošanu, ja šos datus var atkal sasaistīt ar personu.

b. Slēptas AI sistēmas

(131) Cilvēkiem vienmēr ir jāzina, vai tie tieši kontaktējas ar citu cilvēku vai ar automātu, un AI jomas praktiķi atbild par šā nosacījuma nemaldīgu izpildi. Tādēļ AI jomas praktiķiem ir jānodrošina, lai cilvēki būtu informēti vai varētu pieprasīt informāciju un pārliecināties par to, ka tie kontaktējas ar AI sistēmu (piemēram, lai tiktu sniegtas skaidras un pārredzamas atrunas). Jāņem vērā, ka ir robežgadījumi, kas sarežģī situāciju (piem., AI filtrē cilvēka ierunātu tekstu). Ir jāpatur prātā, ka cilvēku un automātu sajaukšanai var būt dažādas sekas, piemēram, pieķeršanās, ietekmēšanās vai cilvēciskās vērtības samazināšanās[73]. Tādēļ cilvēkam līdzīgu robotu izstrāde[74] ir rūpīgi jāizvērtē no ētikas viedokļa.

c. AI atbalstīta iedzīvotāju vērtēšana, kas pārkāpj pamattiesības

(132) Sabiedrībai jācenšas aizsargāt visu pilsoņu brīvību un patstāvību. Jebkāda iedzīvotāju vērtēšana var novest pie patstāvības zaudējuma un apdraudēt nediskriminēšanas principu. Vērtēšana būtu jāizmanto tikai tad, ja tā ir dibināta un ja pasākumi ir samērīgi un taisnīgi. Normatīva valsts iestāžu vai privātā sektora personu veikta iedzīvotāju vērtēšana (vispārējs "morālās personības" vai "ētiskās godprātības" vērtējums) *visos* aspektos un plašā mērogā apdraud šīs vērtības, jo īpaši, ja to neizmanto saskaņā ar pamattiesībām vai ja to izmanto nesamērīgi, bez skaidri noteikta un paziņota leģitīmā mērķa.

(133) Patlaban iedzīvotāju vērtēšana plašākā vai šaurākā mērogā jau tiek bieži izmantota tīri aprakstošos un konkrētai nozarei raksturīgos vērtējumos (piemēram, skolās, e-mācībās, vadītāja apliecību iegūšanā). Pat šādos šauros lietojumos iedzīvotājiem ir jādara zināma pilnībā pārredzama procedūra, ieskaitot informāciju par vērtēšanas procesu, mērķi un metodiku. Ir jāņem vērā, ka pārredzamība nevar novērst diskrimināciju vai nodrošināt taisnīgumu, kā arī nav brīnumlīdzeklis pret vērtēšanas problēmu. Vēlams, lai – kad vien tas

---

71 Šajā sakarā jāatgādina VDAR 6. pants, saskaņā ar kuru cita starpā datu apstrāde ir likumīga tikai tad, ja tai ir spēkā esošs juridiskais pamats.

72 Kā rāda pašreizējie mehānismi informētas piekrišanas došanai internetā, patērētāji parasti dod piekrišanu, to neizvērtējot pēc būtības. Tādēļ tos nevar uzskatīt par īsti praktiskiem.

73 *Madary & Metzinger* (2016. g.), "*Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology. Frontiers in Robotics and AI*", 3(3).

74 Tas attiecas arī uz AI darbinātiem avatāriem.

iespējams bez nelabvēlīgām sekām – tiktu piedāvāta iespēja izstāties no vērtēšanas mehānisma; citos gadījumos ir jāparedz vērtējuma apstrīdēšanas un labošanas mehānismi. Tas ir īpaši svarīgi situācijās, kad pušu starpā pastāv varas asimetrija. Šādas atteikšanās iespējas ir jānodrošina tehnoloģijas projektā, kad tas ir vajadzīgs, lai nodrošinātu atbilstību pamattiesībām, un ir nepieciešams demokrātiskā sabiedrībā.

d. Autonomas letālu ieroču sistēmas (*LAWS*)

(134) Pašreiz nenoteikts skaits valstu un nozaru pēta un izstrādā autonomas letālu ieroču sistēmas – no vadāmām raķetēm, kas spēj selektīvi izraudzīties mērķi, līdz mācīties spējīgām iekārtām ar kognitīvajām prasmēm bez cilvēka iejaukšanās izlemt, pret ko, kad un kur cīnīties. Tas raisa būtiskas ētiskas bažas, piemēram, ka šādi varētu izcelties vēl nekad nepieredzēta nekontrolējama bruņošanās sacensība un tiktu radīts militārais konteksts, kurā cilvēka kontroles gandrīz vairs nav un netiek risināts kļūdainu darbību risks. Eiropas Parlaments ir aicinājis steidzami izstrādāt vienotu juridiski saistošu nostāju, kas risinātu ētiskos un juridiskos jautājumus, kuri saistīti ar cilvēku kontroli, pārraudzību, atbildīgumu, kā arī starptautisko cilvēktiesību, starptautisko humanitāro tiesību un militāro stratēģiju īstenošanu[75]. Atsaucot atmiņā Eiropas Savienības mērķi – veicināt mieru, kā noteikts Līguma par Eiropas Savienību 3. pantā, mēs piekrītam un centīsimies atbalstīt Parlamenta 2018. gada 12. septembra rezolūciju un visus ar to saistītos centienus *LAWS* jomā.

e. Iespējamās ilgtermiņa problēmjomas

(135) AI izstrāde joprojām notiek pa nozarēm un prasa, lai labi apmācīti cilvēki – zinātnieki un inženieri – precīzi noteiktu tās mērķus. Tomēr ir iespējams hipotētiski noteikt atsevišķas kritiskas ilgtermiņa problēmjomas, veroties nākotnē un paturot prātā ilgāku laikposmu[76]. Uz risku balstīta pieeja liek domāt, ka šīs problēmjomas ir jāpatur prātā, ņemot vērā varbūtējos "nezināmus nezināmos" un negaidītos notikumus ("melnos gulbjus")[77]. Ņemot vērā šo problēmjomu lielo ietekmi un attiecīgo norišu pašreizējo nenoteiktību, šie jautājumi ir regulāri jāpārskata.

D. NOBEIGUMS

(136) Šajā dokumentā ir ietvertas mākslīgā intelekta ētikas vadlīnijas, ko izstrādājusi AI augsta līmeņa ekspertu grupa (AI ALEG).

(137) Mēs apzināmies komerciālo un sociālo pozitīvo ietekmi, ko jau radījušas un turpinās radīt AI sistēmas. Tomēr mēs arī bažījamies par to, kā nodrošināt, lai, ņemot vērā AI lietojumu, tiktu pienācīgi un samērīgi aptverts ar šīm tehnoloģijām saistītais risks un citāda nelabvēlīga ietekme. AI ir pārveidojoša un revolucionāra tehnoloģija, kuras attīstību dažu pēdējo gadu laikā ir sekmējusi liela daudzuma digitālo datu pieejamība, skaitļošanas jaudas un uzglabāšanas jaudas būtiskais tehnoloģiskais progress, kā arī svarīga zinātniskā un inženiertehniskā inovācija AI paņēmienu un instrumentu jomā. AI sistēmas turpinās ietekmēt sabiedrību un cilvēkus tādos veidos, kādus mēs vēl nespējam iedomāties.

(138) Šajā sakarā ir svarīgi veidot AI sistēmas, kurām var uzticēties, jo cilvēki to radītos ieguvumus spēs izmantot ar pārliecību un pilnībā tikai tad, kad varēs uzticēties tehnoloģijai, kā arī tehnoloģijas pamatā esošajiem procesiem un tehnoloģijas radītājiem. Tādēļ mūsu pamatmērķis, izstrādājot šīs vadlīnijas, ir uzticams AI.

(139) Uzticamam AI ir trīs iezīmes: 1) tam jābūt likumīgam – jāievēro visi piemērojamie normatīvie akti; 2) tam jābūt ētiskam – jānodrošina ētikas principu un vērtību ievērošana; 3) tam jābūt noturīgam no tehniskā un sociālā viedokļa, lai nodrošinātu, ka AI sistēmas nenodara nevilšu kaitējumu pat tad, ja nodomi bijuši labi. Katra iezīme ir nepieciešama uzticama AI sasniegšanai, bet pati par sevi nav pietiekama. Ideālā gadījumā visas trīs iezīmes

[75] Eiropas Parlamenta rezolūcija 2018/2752(RSP).

[76] Lai gan atsevišķi cilvēki ir pārliecināti, ka mākslīgais vispārējais intelekts, mākslīgā apziņa, mākslīgie morāles aģenti, superintelekts vai pārveidojošs AI varētu būt šādu ilgtermiņa bažu piemēri (kas patlaban neeksistē), daudzi citi uzskata, ka tie ir nereāli.

[77] "Melnais gulbis" ir ļoti rets, bet ievērojams notikums, kas var būt tik rets, ka nekad nav ticis novērots. Tādēļ tā iestāšanās varbūtību parasti var aplēst tikai ar lielu nenoteiktību.

darbojas kā saskaņoti komponenti, kuru darbība pārklājas. Pretrunu rašanās gadījumā mums ir jācenšas tās salāgot.

(140) I nodaļā mēs izklāstījām pamattiesības un attiecīgos ētikas principus, kam ir būtiska nozīme AI kontekstā. II nodaļā mēs norādījām septiņas galvenās prasības, kuras AI sistēmām jāievēro, lai īstenotu uzticamu AI. Mēs ierosinājām tehniskus un netehniskus paņēmienus, kas varētu palīdzēt izpildīt šīs prasības. Visbeidzot, III nodaļā mēs sniedzām uzticama AI novērtēšanas sarakstu, kas var palīdzēt šīs septiņas prasības piemērot praksē. Pēdējā sadaļā esam iekļāvuši piemērus tam, kādas izdevīgas iespējas rada AI sistēmas un kādas ir to kritiskās problēmjomas, un vēlamies stimulēt šo piemēru tālāku apspriešanu.

(141) Eiropai ir unikāla priekšrocība, ko nosaka tās orientācija uz to, ka visu darbību centrā ir iedzīvotāji. Šis princips ir iedzīvināts Eiropas Savienības saknēs un nostiprināts Līgumos, uz kuru pamata tā ir izveidota. Šis dokuments ir daļa no iecerējuma, kas atbalsta uzticamu AI; pēc mūsu domām, tam jābūt pamatam, uz kura Eiropa varētu tiekties ieņemt līdera lomu novatorisku augsto tehnoloģiju AI sistēmu jomā. Šis vērienīgais redzējums palīdzēs kā atsevišķi, tā kolektīvi nodrošināt Eiropas iedzīvotājiem uzplaukumu. Mūsu mērķis ir radīt "uzticama AI Eiropai" kultūru, kurā AI sniegtos ieguvumus ikviens var izmantot tā, lai tiktu nodrošināta cieņa pret mūsu pamatvērtībām: pamattiesībām, demokrātiju un tiesiskumu.

GLOSĀRIJS

(142) Šis glosārijs attiecas uz vadlīnijām un ir paredzēts tam, lai palīdzētu izprast šajā dokumentā izmantotos terminus.

Mākslīgā intelekta jeb AI sistēmas

(143) Mākslīgā intelekta (AI) sistēmas ir programmatūras (iespējams, arī aparatūras) sistēmas, kuras izstrādājis cilvēks[78] un kuras, pastāvot sarežģītam mērķim, darbojas fiziski vai digitāli, uztverot apkārtējo vidi kā ievadītus datus, interpretējot savāktos strukturētos vai nestrukturētos datus, izdarot spriedumus par zināšanām vai apstrādājot no šiem datiem iegūto informāciju, kā arī pieņemot lēmumus par labāko rīcību konkrētā mērķa sasniegšanai. AI sistēmas var izmantot simboliskus noteikumus vai mācīties no cipariska modeļa, kā arī pielāgot savu darbību, analizējot, kā to iepriekšējā rīcība ir ietekmējusi vidi.

(144) AI kā zinātnes disciplīna ietver dažādas pieejas un paņēmienus, piemēram, mašīnu mācīšanos (konkrēti piemēri – mašīnu dziļā mācīšanās un stimulētā mācīšanās), mašīnu spriešanu (ietver plānošanu, programmu veidošanu, zināšanu reprezentāciju un spriešanu, meklēšanu un optimizēšanu) un robotiku (ietver kontroli, uztveri, sensorus un iedarbinātājus, kā arī pārējo paņēmienu integrēšanu kiberfiziskās sistēmās).

(145) Līdztekus tiek publicēts atsevišķs AI ALEG sagatavots dokuments "AI definīcija. Galvenās spējas un zinātnes disciplīnas", kur izklāstīta šajā dokumentā izmantotā AI sistēmu definīcija.

AI jomas praktiķi

(146) AI jomas praktiķi ir visas personas un organizācijas, kas izstrādā (ieskaitot pētniecību, projektēšanu vai datu nodrošināšanu), izvieto (ieskaitot īstenošanu) vai izmanto AI sistēmas, izņemot personas, kas lieto AI sistēmas kā galalietotāji vai patērētāji.

AI sistēmas dzīves cikls

(147) AI sistēmas dzīves cikls ietver tās izstrādes (ieskaitot pētniecību, projektēšanu, datu nodrošināšanu un ierobežotus izmēģinājumus), izvietošanas (ieskaitot īstenošanu) un izmantošanas stadiju.

Pārbaudāmība

(148) Pārbaudāmība ir iespēja veikt AI sistēmas algoritmu, datu un projektēšanas procesu izvērtēšanu. Tā ir viena no septiņām prasībām, kas jāievēro uzticamai AI sistēmai. Teiktais nenozīmē, ka ar AI sistēmu saistītajai informācijai par darbības modeļiem un intelektuālo īpašumu vienmēr jābūt publiski pieejamai. Sistēmas pārbaudāmību var palīdzēt nodrošināt trasējamības un žurnalēšanas mehānismu ieviešana, sākot ar AI sistēmas agrīnās projektēšanas stadiju.

Neobjektivitāte

(149) Neobjektivitāte ir tieksme uz aizspriedumiem par labu vai sliktu kādai personai, objektam vai stāvoklim. Neobjektivitāte AI sistēmās var rasties dažādos veidos. Piemēram, tādās AI sistēmās kā mašīnu mācīšanās ceļā radītās sistēmas, kas ir balstītas uz datiem, neobjektivitāte datu vākšanā un apmācībā var novest pie neobjektivitātes AI sistēmā. Uz loģiku balstītā AI, piemēram, kārtulbalstītās sistēmās, neobjektivitāte var rasties atkarā no tā, kā zināšanu jomas inženieris uztver konkrētā situācijā piemērojamus noteikumus. Neobjektivitāte var rasties arī sakarā ar tiešsaistes mācīšanos un pielāgošanos saskarsmes ceļā. Tāpat tā var rasties personalizācijas rezultātā, kad lietotājiem tiek piedāvāti viņu gaumei pielāgoti ieteikumi vai informācijas plūsma. Tā var nebūt saistīta ar cilvēku neobjektivitāti vai cilvēku vadītu datu vākšanu. Tā var rasties, piemēram, sistēmas ierobežotā izmantošanas konteksta dēļ, un šajā gadījumā to nav iespējams vispārināt uz citu kontekstu. Neobjektivitāte var būt laba vai slikta, tīša vai netīša. Dažos gadījumos neobjektivitāte var novest pie diskriminējošiem un/vai netaisnīgiem rezultātiem, kas šajā dokumentā apzīmēti kā netaisnīga

[78] Cilvēki AI sistēmas projektē tieši, bet projektu optimizēšanā var tikt izmantoti arī AI paņēmieni.

neobjektivitāte.

Ētika

(150) Ētika ir akadēmiska disciplīna – filozofijas apakšnozare. Kopumā tā atbild uz tādiem jautājumiem kā "Kas ir laba rīcība?", "Kāda ir cilvēka dzīvības vērtība?", "Kas ir taisnīgums?" vai "Kas ir laba dzīve?". Akadēmiskajā ētikā ir četri galvenie pētniecības virzieni: i) metaētika, kas galvenokārt skar normatīvu izteikumu nozīmi un atsauci, kā arī jautājumu, kā iespējams noteikt to patiesumvērtības (ja tādas ir); ii) normatīvā ētika – praktiskais veids, kā noteikt morāli pareizo rīcību, izvērtējot pareizas un nepareizas rīcības standartus un piešķirot vērtību konkrētai rīcībai; iii) deskriptīvā ētika, kuras mērķis ir empīriski pētīt cilvēku morālo uzvedību un uzskatus; iv) lietišķā ētika, kura attiecas uz to, ko mums ir pienākums (vai ir atļauts) darīt konkrētā (bieži vien vēsturiski jaunā) situācijā vai konkrētā (bieži vien vēsturiski neeksistējušā) rīcības variantu jomā. Lietišķā ētika attiecas uz reālās dzīves situācijām, kurās lēmumi jāpieņem ierobežotā laikposmā un bieži vien ar ierobežotu pamatojumu. AI ētika pārsvarā tiek uztverta kā lietišķās ētikas apakšnozare un ir orientēta uz AI projektēšanas, izstrādes, izvietošanas un izmantošanas radītajām normatīvajām bažām.

(151) Diskusijās par ētiku bieži lieto jēdzienus "morāls" un "ētisks". Jēdziens "morāls" attiecas uz konkrētiem, faktiskiem uzvedības modeļiem, paražām un ieradumiem, kas noteiktā laikā pastāv konkrētā kultūrā vai grupā vai raksturīgi konkrētam indivīdam. Jēdziens "ētisks" attiecas uz šo konkrēto darbību un uzvedības vērtējumu no sistēmiska, akadēmiska viedokļa.

Ētisks AI

(152) Šajā dokumentā jēdziens "ētisks AI" lietots attiecībā uz tādu AI izstrādi, izvietošanu un izmantošanu, kas nodrošina atbilstību ētikas normām, ieskaitot pamattiesības kā īpašas morālas tiesības, ētikas principus un ar tiem saistītas pamatvērtības. Tas ir otrais no trim pamatelementiem, kas nepieciešami uzticama AI sasniegšanai.

Antropocentrisks AI

(153) Antropocentriskas pieejas AI mērķis ir nodrošināt, lai cilvēciskajām vērtībām būtu noteicošā loma AI sistēmu izstrādē, izvietošanā, izmantošanā un uzraudzībā, garantējot pamattiesību, tostarp Eiropas Savienības līgumos un Eiropas Savienības Pamattiesību hartā noteikto tiesību, ievērošanu, tā kā visas minētās tiesības vieno kopīgs pamats, kurš balstās uz cilvēka cieņas ievērošanu un kurā cilvēks bauda unikālu un neatņemamu morālo statusu. Minētais ietver arī apsvērumus attiecībā uz dabas vidi un citām dzīvajām būtnēm, kas ir daļa no cilvēku ekosistēmas, kā arī ilgtspējīgu pieeju, kas ļautu uzplaukumu piedzīvot arī turpmākajām paaudzēm.

Sarkanās komandas izmantošana

(154) Sarkanās komandas izmantošana ir prakse izmantot "sarkano komandu" jeb neatkarīgu grupu, kas, uzņemoties pretinieka lomu vai uzskatus, panāk, lai organizācija uzlabo tās efektivitāti. Tā īpaši noder, lai palīdzētu atklāt un novērst varbūtējās vājās vietas no drošības viedokļa.

Reproducējamība

(155) Reproducējamība attiecas uz to, vai AI eksperimenta iznākums ir tāds pats, ja to atkārto ar tiem pašiem nosacījumiem.

Noturīgs AI

(156) AI sistēmas noturīgums aptver gan tās tehnisko noturīgumu (atbilstīgi konkrētam kontekstam, piemēram, lietojuma jomai vai dzīves cikla posmam), gan noturīgumu no sociālā viedokļa (nodrošināšanu, ka AI sistēma pienācīgi ņem vērā sistēmas darbības kontekstu un vidi). Tas ir svarīgi, lai nodrošinātu, ka netiek nodarīts netīšs kaitējums pat tad, ja nodomi ir bijuši labi. Noturība ir trešā no trim iezīmēm, kas nepieciešamas uzticama AI sasniegšanai.

Ieinteresētās personas

(157) Ieinteresētās personas ir visi, kuri pēta, izstrādā, izvieto vai izmanto AI, kā arī personas, kuras (tieši vai netieši) ietekmē AI, tostarp (bet ne tikai) uzņēmumi, organizācijas, pētnieki, sabiedrisko pakalpojumu sniedzēji, iestādes, pilsoniskās sabiedrības organizācijas, valdības, regulatori, sociālie partneri, indivīdi, pilsoņi, darba ņēmēji un patērētāji.

Trasējamība

(158) AI sistēmas trasējamība ir spēja izsekot sistēmas datiem, izstrādes un izvietošanas procesiem, parasti – izmantojot dokumentētu un reģistrētu identifikāciju.

Uzticēšanās

(159) Tiek izmantota šāda literatūrā atrodama definīcija: "Ar uzticēšanos saprot: 1) konkrētu ar labvēlību, kompetenci, godprātību un paredzamību saistītu uzskatu kopumu (uz uzticēšanos orientēti uzskati); 2) vienas puses gatavību būt atkarīgai no otras bīstamā situācijā (uzticēšanās nodoms); vai 3) šo elementu kombināciju."[79] Lai gan "uzticēšanās" parasti nav uz automātiem attiecināms jēdziens, šā dokumenta mērķis ir uzsvērt, cik svarīgi ir spēt ne vien uzticēties tam, ka AI sistēmas atbilst tiesību aktiem, ievēro ētikas normas un ir noturīgas, bet arī šādu uzticēšanos attiecināt uz visiem AI sistēmas dzīves ciklā iesaistītajiem cilvēkiem un procesiem.

Uzticams AI

(160) Uzticamam AI ir trīs iezīmes: 1) tam jābūt likumīgam un jāievēro visi piemērojamie normatīvie akti; 2) tam jābūt ētiskam – jāpauž cieņa un jānodrošina ētikas principu un vērtību ievērošana; 3) tam jābūt noturīgam no tehniskā un sociālā viedokļa, jo AI sistēmas var nodarīt netīšu kaitējumu pat tad, ja nodomi ir bijuši labi. Uzticams AI ietver ne tikai pašas AI sistēmas uzticamību, bet arī visu sistēmas dzīves ciklā ietilpstošo procesu un iesaistīto personu uzticamību.

Neaizsargātas personas un grupas

(161) Tiesībās nav vispārpieņemtas vai plaši atzītas neaizsargātu personu definīcijas, jo tās ir ļoti dažādas. Bieži vien tikai no konteksta izriet, kas ir neaizsargāta persona vai grupa. To var ietekmēt pārejošas dzīves norises (piemēram, bērnība, slimība), tirgus faktori (piemēram, informācijas asimetrija vai tirgus vara), ekonomiskie faktori (piemēram, nabadzība), ar identitāti saistīti (piemēram, dzimums, reliģija, kultūra) vai citi faktori. ES Pamattiesību hartas 21. pants aptver nediskriminēšanu uz turpmāk minētajiem pamatiem, kas cita starpā var būt par atskaites punktu: ir aizliegta diskriminācija dzimuma, rases, ādas krāsas, etniskās vai sociālās izcelsmes, ģenētisko īpatnību, valodas, reliģijas vai ticības, politisko vai citādu uzskatu, piederības pie mazākumtautības, mantas, kārtas, invaliditātes, vecuma un dzimumorientācijas dēļ. Citas juridiskās normas attiecas uz konkrētu grupu tiesībām papildus jau minētajām. Jebkurš šāds saraksts nav izsmeļošs un laika gaitā var mainīties. Neaizsargāta grupa ir tādu personu grupa, kam piemīt viena vai vairākas neaizsargātības pazīmes.

[79] *Siau, K., Wang, W.* (2018. g.), "*Building Trust in Artificial Intelligence, Machine Learning, and Robotics*", *CUTTER BUSINESS TECHNOLOGY JOURNAL* (31), 47.–53.

Dokumentu izstrādājuši AI augsta līmeņa ekspertu grupas locekļi,

kuru saraksts dots alfabētiskā secībā

Pekka Ala-Pietilä, AI ALEG priekšsēdētājs
AI Finland, *Huhtamaki*, *Sanoma*

Wilhelm Bauer
Fraunhofer
Urs Bergmann, līdzziņotājs
Zalando

Mária Bieliková
Slovākijas Tehniskā universitāte Bratislavā
Cecilia Bonefeld-Dahl, līdzziņotāja
DigitalEurope
Yann Bonnet
ANSSI
Loubna Bouarfa
OKRA
Stéphan Brunessaux
Airbus
Raja Chatila
IEEE iniciatīva "Intelektisku un autonomu sistēmu ētika" & Sorbonnas Universitāte
Mark Coeckelbergh
Vīnes Universitāte
Virginia Dignum, līdzziņotāja
Ūmeo Universitāte
Luciano Floridi
Oksfordas Universitāte
Jean-Francois Gagné, līdzziņotājs
Element AI
Chiara Giovannini
ANEC
Joanna Goodey
Pamattiesību aģentūra
Sami Haddadin
Minhenes Robotikas un AI skola
Gry Hasselbalch
Thinkdotank DataEthics & Kopenhāgenas Universitāte
Fredrik Heintz
Linšēpingas Universitāte
Fanny Hidvegi
Access Now
Eric Hilgendorf
Vircburgas Universitāte
Klaus Höckner
Hilfsgemeinschaft der Blinden und Sehschwachen
Mari-Noëlle Jégo-Laveissière
Orange
Leo Kärkkäinen
Nokia Bell Labs
Sabine Theresia Köszegi
Vīnes Tehniskā universitāte
Robert Kroplewski,
advokāts un Polijas valdības konsultants
Elisabeth Ling
RELX

Pierre Lucas
Eiropas Mašīnbūves rūpniecības apvienība – Eiropas tehnoloģiju nozare
Ieva Martinkenaite
Telenor
Thomas Metzinger, līdzziņotājs
Maincas Johannesa Gūtenberga universitāte & Eiropas Universitāšu asociācija
Catelijne Muller
ALLAI Netherlands un EESK
Markus Noga
SAP
Barry O'Sullivan, AI ALEG priekšsēdētāja vietnieks
Korkas Universitātes koledža
Ursula Pachl
BEUC
Nicolas Petit, līdzziņotājs
Lježas Universitāte
Christoph Peylo
Bosch

Iris Plöger
BDI
Stefano Quintarelli
Garden Ventures
Andrea Renda
Eiropas Koledžas mācībspēki un *CEPS*
Francesca Rossi
IBM
Cristina San José
Eiropas Banku federācija
George Sharkov
Digital SME Alliance
Philipp Slusallek
Vācijas AI pētniecības centrs (*DFKI*)
Françoise Soulié Fogelman,
AI konsultante
Saskia Steinacker, līdzziņotāja
Bayer
Jaan Tallinn
Ambient Sound Investment
Thierry Tingaud
STMicroelectronics
Jakob Uszkoreit
Google
Aimee Van Wynsberghe, līdzziņotāja
Delftas Tehniskā universitāte
Thiébaut Weber
Eiropas Arodbiedrību konfederācija
Cecile Wendling
AXA
Karen Yeung, līdzziņotāja
Birmingemas Universitāte

Urs Bergmann, Cecilia Bonefeld-Dahl, Virginia Dignum, Jean-François Gagné, Thomas Metzinger, Nicolas Petit, Saskia Steinacker, Aimee Van Wynsberghe un *Karen Yeung* bija šā dokumenta referenti.

Pekka Ala-Pietilä ir AI ALEG priekšsēdētājs. *Barry O'Sullivan* ir AI ALEG priekšsēdētāja vietnieks un koordinē AI ALEG otro iesniedzamo

dokumentu. Dokumenta saturā ieguldījumu sniegusi arī *Nozha Boujemaa*, priekšsēdētāja vietniece līdz 2019. gada 1. februārim, kura koordinēja pirmo iesniedzamo dokumentu.

Redakcionālu atbalstu sniegusi *Nathalie Smuha*.