

Haliucinacijų rizika didelės rizikos dirbtinio intelekto sistemose: Europos Sąjungos teisėkūros bei duomenų mokslo perspektyvos

Emilija Pužauskaitė

Vilniaus Universiteto Teisės fakulteto
4-o kurso studentė
Saulėtekio al. 9, I rūmai, 10222 Vilnius
El. paštas: emilija.puzauskaite@tf.stud.vu.lt

Marija Rapševičiūtė

Vilniaus Universiteto Matematikos ir informatikos fakulteto
2-o kurso studentė
Didlaukio g. 47, 08303 Vilnius
El. paštas: marija.rapseviciute@mif.stud.vu.lt

Mokslinio straipsnio akademinė kuratorė: j. asist. dokt. Deimantė Rimkutė

El. paštas: deimante.rimkute@tf.vu.lt

Mokslinio straipsnio praktinis kuratorius: Dr. Tomas Raila

El. paštas: tomas.raila@mif.vu.lt

Santrauka. Straipsnio tikslas yra pateikti galimus haliucinacijų reiškinių dirbtinio intelekto sistemose mažinimo metodus, siekiant užtikrinti didelės rizikos dirbtinio intelekto sistemų teisėtumą. Siekiant įgyvendinti šį tikslą, darbe bus aptariama ir didelės rizikos dirbtinio intelekto sistemų samprata bei joms keliami reguliaciniai reikalavimai. Ypačingas dėmesys skiriamas tiems didelės rizikos dirbtinio intelekto sistemų reguliaciniams reikalavimams, kurie yra aktualiausi analizuojant haliucinacijų reiškinių pasireiškimą bei valdymą. Straipsnyje keliami hipotezė, kad šiandien duomenų mokslas dar nesudaro prielaidų visiškai eliminuoti haliucinacijų ir jų keliamų rizikų, tačiau leidžia taikyti priemonių rinkinį, kuris sumažina tokių klaidų tikimybę ir poveikį.

Pagrindiniai žodžiai: dirbtinio intelekto haliucinacijos, didelės rizikos dirbtinio intelekto sistemos, Dirbtinio intelekto aktas, didieji kalbos modeliai.

The Risk of Hallucinations in High-Risk Artificial Intelligence Systems: European Union Legislation and Data Science Perspectives

Abstract. *The purpose of the article is to present possible methods of reducing the phenomenon of hallucinations in artificial intelligence systems to ensure the legitimacy of high-risk artificial intelligence systems. To achieve this goal, the paper will also discuss the concept of high-risk artificial intelligence systems and the regulatory requirements associated with them. Special attention is paid to those regulatory requirements for high-risk artificial intelligence systems that are most relevant when analysing the occurrence and management of hallucinations. The article argues that, although data science cannot yet eliminate hallucinations and their risks, it enables the implementation of measures that reduce their likelihood and impact.*

Keywords: *artificial intelligence hallucinations, high-risk artificial intelligence systems, Artificial Intelligence Act, large language models.*

Įvadas

Temos aktualumas ir originalumas. Žmonės vis dažniau kasdieniams sprendimams – nuo kelionių planavimo iki darbo užduočių – pasitelkia generatyvų dirbtinį intelektą (toliau - DI). Pirmas išpūdis būna geras: atsakymai sklandūs, struktūruoti, pateikiami užtikrintu tonu. Tačiau praktika parodo ir kitą pusę – kartais DI „suklysta“ ne kaip paieškos sistema, kuri tiesiog neranda informacijos, o kaip pasakotojas, kuris pateikia įtikinamai skambančius, bet neteisingus faktus, citatas, ar neegzistuojančius teismų praktikos fragmentus. Viena tokių klaidų – 2025 metais įvykusi situacija, kuomet Lietuvos Aukščiausią Teismą pasiekė kasacinis skundas, kuriame buvo sąmoningai remiamasi neegzistuojančia teismų praktika. Tokias klaidas lemia DI sistemose vykstantis reiškiny – haliucinacijos.¹ Haliucinacijų pavyzdys pateikiamas lentelėje Nr. 1.

Šios klaidos kelia grėsmę DI sistemų patikimumui ir teisėtumui, nes naudotojai dažnai nesuabejoja pateiktos informacijos teisingumu. Nors daugelis generatyvinių modelių gali haliucinuoti, ši sąvoka plačiausiai naudojama didžiųjų kalbos modelių (toliau - DKM) kontekste, nes jie generuoja semantiškai interpretuojamą natūralios kalbos tekstą, kurį paprasta suprasti. Teisininkams ši problema svarbi tuo, kad klaidinantis atsakymas nėra tik vartotojo nepatogumas – kai DI naudojamas rengiant

¹ Autorės pažymi, kad šio mokslinio tyrimo kontekste, haliucinacijų reiškinio rizikos valdymas yra analizuojamas abstrakčiai, konkrečiai siejant kylančią riziką su didelės rizikos DI sistemomis, išsamiai neanalizuojant Bendrosios paskirties dirbtinio intelekto (toliau - BPDI) modelių bei jo teisinio reglamentavimo. Tačiau, autorės pažymi, jog BPDI modeliai bus minimi tiek, kiek tai būtina paaiškinti didelės rizikos DI sistemos elgseną ir rizikos kilmę.

procesinius dokumentus, sugalvotas teismų praktikos fragmentas gali klaidinti kitą šalį ar teismą. Analogiškos pasekmės galimos ir kitose srityse – medicinoje, finansuose ar švietime, kur haliucinuojančio DI pateikta netiksli informacija gali lemti ne tik klaidingus sprendimus, bet ir nuostolius ar net pavojų žmogaus sveikatai. Siekiant suvaldyti tokias rizikas (įskaitant haliucinacijas), Europos Sąjunga (toliau - ES) priėmė DI aktą, kuriuo nustatomi techniniai ir kiti reguliaciniai reikalavimai, didžiausią dėmesį skiriant DI sistemoms, keliančioms padidintą riziką žmogaus teisėms. Vis dėlto, haliucinacijų suvaldymas nėra paprastas net ir nustačius DI sistemoms keliamus reikalavimus, todėl praktikoje tai tampa mišria užduotimi, kurios įgyvendinimui gali būti pasitelktos duomenų mokslo žinios. Nors haliucinacijų reiškiny savaimė negali būti laikomas neteisėtu, tačiau didelės rizikos DI kontekste gali daryti esminę įtaką DI akto keliamiems reikalavimams. Siekiant valdyti haliucinacijų reiškinių keliamą riziką didelės rizikos DI sistemose, itin svarbu atsižvelgti į šio reiškinių kilmę – techninius ir statistinius DI sistemų veikimo ypatumus. Todėl, siekiant užtikrinti didelės rizikos DI patikimumą, svarbu analizuoti, kaip duomenų mokslo metodai gali prisidėti prie haliucinacijų reiškinių mažinimo bei rizikos valdymo. Atliktas tyrimas yra svarbus tiek teoriniu, tiek praktiniu požiūriu – šiuo tyrimu siekiama nurodyti galimus haliucinacijų reiškinių valdymo metodus, siekiant užtikrinti, kad naudojamas DI atitiktų jam keliamus reikalavimus. Tikėtina, kad moksliniame straipsnyje analizuojama tema su kiekvienais metais taps vis aktualesnė, turint omenyje tai, jog DI yra naujovė tiek teisės, tiek duomenų mokslo perspektyvomis, ir šios naujovės taikymo sritys dėl jos dinamiškumo galės būti naudojamos šiai dienai dar neaktualiais aspektais.

1 lentelė. Haliucinacijų reiškinių, pasireiškiančio DI sistemose, pavyzdys

Apibūdink 1998 m. lietuvišką filmą „Tylos krantai“ ir jo pagrindinę temą.

Filmas „**Tylos krantai**“ (1998 m.), režisuotas **Valdo Navasaičio**, yra vienas ryškiausių ir vizualiai įsimintiniausių ankstyvojo nepriklausomybės laikotarpio lietuvių kino kūrinių. Tai minimalistinis, mediatyvus darbas, kuris dažnai priskiriamas „lėtojo kino“ (angl. *slow cinema*) tradicijai.

Šis atsakymas yra faktiškai neteisingas, nes 1998 metais nebuvo sukurtas filmas pavadinimu „Tylos krantai“, nors režisierius Valdas Navasaitis egzistuoja.

Šaltinis: sugeneruotas ir aprašytas autorių.

Šio tyrimo tikslas – parodyti, kad haliucinacijų rizikos valdymas didelės rizikos DI sistemose turi remtis ne vien teisinių reikalavimų identifikavimu, bet ir techniniu supratimu apie šio reiškinio kilmę. Siekiant užtikrinti didelės rizikos DI patikimumą, straipsnyje analizuojama, kaip duomenų mokslo metodai gali prisidėti prie haliucinacijų mažinimo ir kaip šios priemonės gali būti integruojamos į rizikos valdymo, testavimo ir žmogaus priežiūros procesus. Straipsnyje tikrinama hipotezė, ar duomenų mokslas šiandien gali pasiūlyti praktiškai taikytinų sprendimų haliucinacijų rizikai mažinti, ir kokios yra tokių sprendimų ribos (t. y. kiek rizika gali būti suvaldoma, bet ne visiškai pašalinama).

Siekiant įgyvendinti numatytą tyrimo tikslą, šiame moksliniame tyrime keliami šie uždaviniai: i) nurodyti haliucinacijų reiškinio sampratą bei iširti atsiradimo priežastis, (ii) aprašyti ES teisės aktuose pateiktą didelės rizikos DI sistemų sampratą bei išanalizuoti didelės rizikos DI sistemoms keliamus reguliacinius bei techninius reikalavimus, kiek tai susiję su haliucinacijų reiškiniu, (iii) aprašyti efektyviausius ir plačiausiai duomenų mokslo siūlomus haliucinacijų reiškinio valdymo metodus didelės rizikos DI sistemose.

Atsižvelgiant į tyrimo tikslą bei iškeltus uždavinius, šio tyrimo objektas yra DI sistemose pasireiškiantis haliucinacijų reiškiny, bei šio reiškinio valdymas didelės rizikos DI sistemose.

Šiame moksliniame straipsnyje pasitelkti šie tyrimo metodai: teisinio teksto interpretavimo metodas taikytas interpretuojant Europos Sąjungos teisės aktų nuostatas, siekiant nuosekliai atskleisti jų turinį, aprašomasis metodas taikytas keliais aspektais: (i) siekiant nurodyti didelės rizikos DI sistemų teisėtumui keliamus reikalavimus; (ii) aprašant haliucinacijų reiškinio sampratą; (iii) analizuojant haliucinacijų atsiradimo priežastis DKM, remiantis duomenų mokslo literatūra; (iv) aprašant ir susisteminant literatūroje pateikiamus haliucinacijų mažinimo metodus bei jų pritaikomumo ribas didelės rizikos DI sistemose, sisteminiu tyrimo metodu, kartu su lyginamuoju tyrimo metodu, atliekama duomenų mokslo autorių literatūroje pateiktų nuomonių analizė, pateikiamos išvados, literatūros analizės metodas taikytas siekiant išanalizuoti Europos Sąjungos teisinio reglamentavimo bei duomenų mokslo perspektyvas, nagrinėjant aktualią mokslinę literatūrą.

Siekiant pagrįsti autorių argumentavimą, šiame moksliniame tyrime naudoti šie pagrindiniai šaltiniai: (i) teisės šaltiniai – DI aktas, Bendrasis duomenų apsaugos reglamentas (toliau - BDAR) ir kiti aktualūs Europos Sąjungos teisės aktai; (ii) doktrina – remtasi užsienio valstybių autorių teisės mokslo veikalais; (iii) kiti šaltiniai – be paminėtų teisės šaltinių ir doktrinos, siekiant įgyvendinti tyrimo tikslą, buvo remtasi aktualia užsienio autorių mokslinė literatūra duomenų mokslo srityje, ypač L. Huang, N. Karanikolas, E. Manga, bei X. Liu, darbais, kuriuose analizuojamos DKM haliucinacijų atsiradimo priežastys, jų aptikimo ir mažinimo metodai bei šių metodų taikymo ribotumai didelės rizikos DI sistemų kontekste. Ypatingas dėmesys darbe

skiriamas L. Huang, W. Yu, W. Ma ir kt. (2025) apžvalginiam straipsniui, publikuotam ACM Transactions on Information Systems, kuris laikomas vienu iš išsamiausių ir reikšmingiausių mokslinių darbų, sistemingai apibendrinančių DKM haliucinacijų principus.

I. DI haliucinacijų reiškinys ir jo atsiradimo priežastys

Haliucinacijos DI kontekste apibrėžiamos kaip šalutinis generatyvinio DI modelio procesas, kurio metu sugeneruojama informacija nepagrįsta šaltiniais, neatitinkanti faktų arba prieštaraujanti pateiktam kontekstui². Reikšminga dalis DKM haliucinacijų kyla iš pačių šių modelių veikimo principų, todėl prieš nagrinėjant haliucinacijų reiškinį būtina trumpai paaiškinti, kaip DKM generuoja tekstą – ši dalis pradeda glaustu DKM veikimo mechanizmų aptarimu. Tuomet, remiantis aptartais DKM veikimo principais, bus analizuojama, kaip dėl šių modelių architektūros atsiranda haliucinacijos. Galiausiai aptariama, kaip haliucinacijų riziką lemia apmokymui naudojamų duomenų kokybė ir kiekis.

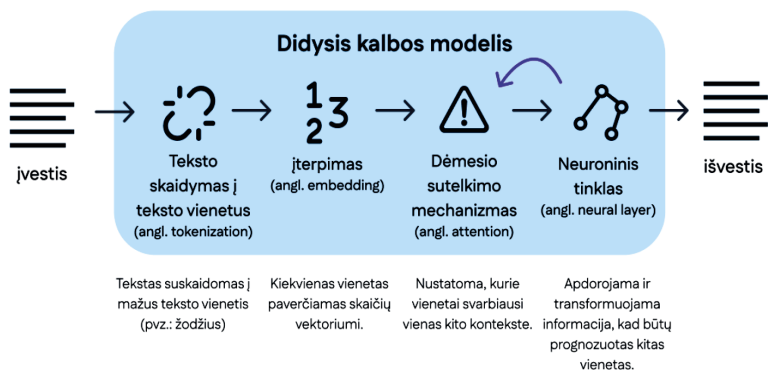
Didžiųjų kalbos modelių veikimo principai. Pagrindinis DKM tikslas yra iš pateiktos natūralios kalbos užklauskos (angl. prompt) pateikti tiksliausią, labiausiai atitinkamą natūralios kalbos atsakymą. Tad, veikimo procesas prasideda nuo vartotojo įvesto teksto suskaidymo į teksto vienetus (angl. tokens) ir apdorojimo per neuroninį tinklą, imituojantį žmogaus mąstymą. Toliau modelis generuoja atsakymą - kiekviename žingsnyje parenka labiausiai tikėtiną kitą teksto vienetą ir jį prideda prie jau savo sugeneruoto teksto, tokiu būdu formuodamas atsakymą³. Esminis DKM veikimo principas – nuspėti labiausiai tikėtiną sekantį teksto vienetą atsižvelgiant į pateiktą užklauską⁴. Žodžių spėjimus modelis daro iš duomenų, kuriais buvo apmokytas. Pavyzdžiui, apmokymo metu modelis apdorojo daug teksto, kuriame buvo minima, kad „pandos gyvena Kinijoje“. Tuomet, generuojant sakinį „Kinijoje galima pamatyti ...“, yra didesnė tikimybė, kad modelis jį užbaigs žodžiu „pandas“, nei žodžiu „triušį“. DKM veikimas yra schemiškai pavaizduotas lentelėje Nr. 2.

² HUANG, L.; YU, W.; MA, W. ir kt. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55, p. 5 [interaktyvus]. <https://dl.acm.org/doi/full/10.1145/3703155>.

³ JI, W., YUAN, W., GETZEN, E. ir kt. (2025). An Overview of Large Language Models for Statistical Analysis. *arXiv [interaktyvus]*. <https://arxiv.org/abs/2502.17814>. p. 2

⁴ CHANG, Y., WANG, X., WANG, J. ir kt. (2024). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1–45 [interaktyvus]. <https://doi.org/10.1145/3641289>. p. 6

2 lentelė. DKM veikimo principo schema



DKM įvestą tekstą suskaido į vienetus, paverčia juos skaičių vektoriais ir naudodamas dėmesio mechanizmą bei neuroninius sluoksnius apdoroja kontekstą. Taip modelis išmoka, kurie teksto elementai svarbiausi ir kaip jie susiję tarpusavyje, kad galėtų prognozuoti išvestį. Violetinė rodyklė žymi, kad dėmesio mechanizmas ir neuroninis tinklas veikia pakartotinai – tie patys blokai kartojami daug kartų.

Šaltinis: sugeneruotas ir aprašytas autorių.

Kadangi atsakymų generavimas priklauso nuo mokymo duomenų, yra sudėtinga nuspėti kokius atsakymus gali pateikti DKM⁵. Dar atsakymų nuspėjamumą apsunkina tai, kad šiuolaikiniai DKM (pvz.: GPT-5 (sukurtas OpenAI)), turi itin didelį nustatymų skaičių, kuris lemia sudėtingą modelių architektūrą⁶. Dėl sunkumo atsekti DKM mąstymo procesą, sakoma, kad šie modeliai pasižymi „juodosios dėžės“ (angl. black-box) bruožais. Dėl sprendimų proceso neskaidrumo sunku tiksliai nustatyti, kokie veiksniai lemia DKM generuojamas haliucinacijas. Literatūroje haliucinacijų pasireiškimo priežastys dažnai skirstomos į dvi plačias kategorijas pabrėžiant jų atsiradimo priežastį: (i) dėl pačio modelio architektūros, ir (ii) duomenų, naudojamų modelio apmokymo proceso metu⁷.

⁵ SU, W. (2025). Do Large Language Models (Really) Need Statistical Foundations? *arXiv* [interaktyvus]. <https://arxiv.org/abs/2505.19145>. p. 5

⁶ SU, W. (2025). Do Large Language Models (Really) Need Statistical Foundations? *arXiv* [interaktyvus]. <https://arxiv.org/abs/2505.19145>. p.7

⁷ HUANG, L., YU, W., MA, W. ir kt. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2), 1–55 [interaktyvus]. <https://dl.acm.org/doi/full/10.1145/3703155>. p. 7

Haliucinacijos, kylančios iš modelio architektūros. Viena iš pagrindinių haliucinacijų atsiradimo priežasčių yra susijusi su DKM architektūra. Tai galėtų paaiškinti anksčiau minėtą 2024 m. advokato kasacinio skundo Lietuvos Aukščiausiajam Teismui atvejį. Viena iš tokio neigiamo rezultato priežasčių galėjo būti nuoseklaus teksto generavimo mechanizmas, esantis DKM. Sugeneravęs pirmąjį melagingą bylos pavadinimą, modelis vėlesnėje generacijoje jį naudojo kaip pagrindą tolimesniems argumentams konstruoti. Situacija, kai modelis, remdamasis savo klaidingu tekstu, generuoja tolesnį tekstą ir daro papildomas klaidas, literatūroje vadinama „ekspozicijos šališkumo“ (angl. *exposure bias*) problema⁸. Minėtame pavyzdyje naudojamas DKM tikriausiai neturėjo tinkamų šaltinių, kuriais galėjo paremti aprašomą skundą, tad nutartis išgalvojo, nes privarėjo pateikti įtikinimą atsakymą. Tai susiję su modelio programavimu visada pateikti atsakymą – net tada, kai jo neturi⁹. DKM negali išreikšti neužtikrintumų – net ir tokiais atvejais, kai DKM neturi tinkamo atsakymo, jie privalo daryti spėjimą, kuris vartotojui pateikiamas kaip užtikrintas faktas¹⁰. Šie mechanizmai, keliantys haliucinacijas DKM, yra tiesiogiai susiję su šių modelių architektūra.

Duomenų kokybės ir kiekio sąsaja su haliucinacijų reiškiniu. Kita reikšminga haliucinacijų atsiradimo priežastis yra duomenų, kuriais modelis apmokomas, kokybė bei kiekis. DKM apmokomi naudojant įvairius duomenų šaltinius – knygas, mokslinius straipsnius, enciklopedijas, interneto svetaines ir socialinius tinklus, kuriuose dažnai pasitaiko neteisinga ar tarpusavyje prieštaringa informacija. Pavyzdžiui, pateikus klausimą „Ar žmonės naudoja tik 10% savo smegenų?“, DKM atsakydavo, kad taip, nors tai nebūdavo tiesa. Toks DKM pateiktas atsakymas gali būti siejamas su tuo, kad šitas teiginys dažnai pasirodydavo populiariosios psichologijos tinklalapiuose socialiniuose tinkluose. Nors duomenų parengimo metu siekiama pašalinti šališką, netiksliai visuomenės tendencijas atspindinčią, faktiškai neteisingą informaciją, filtravimo procesas yra ribotas¹¹. Nors šališka informacija, tokia kaip visuomeniniai stereotipai, savaime nėra laikoma haliucinacija, jos buvimas apmokymo duomenyse didina haliucinacijų riziką¹². 2016 m. „Microsoft Tay“ pokalbių asistento atvejis pa-

⁸ LIN, Z., GUAN, S., ZHANG, W. ir kt. (2024). Towards trustworthy LLMs: a review on debiasing and dehallaucinating in large language models. *Artificial Intelligence Review*, 57(9), 243 [interaktyvus]. <https://link.springer.com/article/10.1007/s10462-024-10896-y>. p. 20

⁹ HUANG, L., YU, W., MA, W. ir kt. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2), 1–55 [interaktyvus]. <https://dl.acm.org/doi/full/10.1145/3703155>. p. 11

¹⁰ HUANG, L., YU, W., MA, W. ir kt. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2), 1–55 [interaktyvus]. <https://dl.acm.org/doi/full/10.1145/3703155>. p. 11

¹¹ ALBALAK, A., ELAZAR, Y., XIE, S. M. ir kt. (2024). A Survey on Data Selection for Language Models. *arXiv* [interaktyvus]. <https://arxiv.org/abs/2402.16827>. p.21

¹² HUANG, L., YU, W., MA, W. ir kt. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2), 1–55 [interaktyvus]. <https://dl.acm.org/doi/full/10.1145/3703155>. p. 8

rodė, kaip nekontroliuojami vieši duomenys gali per kelias valandas suformuoti žalingą modelio elgseną¹³. Pokalbių asistentas „Microsoft Tay“ buvo sukurtas mokytis iš pokalbių „Twitter“ tinkle, tačiau jis perėmė ten vyraujančią rasizmą ir ksenofobiją bei pradėjo juos kartoti savo atsakymuose. Atsižvelgiant į tai, kad DKM išmoka bendras reprezentacijas ir šablonus, šališka informacija lemia polinkį generuoti neteisingą informaciją. Panašų neigiamą poveikį turi ir prieštaringa informacija, būdinga socialiniuose tinkluose, internetiniuose forumuose (pvz., Reddit), kurių duomenys plačiai naudojami apmokant DKM. Vis dėlto ne visi duomenys modelio elgseną formuoja vienodai. Informacija, su kuria DKM retai susiduria, turi mažesnę įtaką modelio mokymosi metu, todėl kyla didesnė išgalvotų atsakymų grėsmė. DKM atveju matyti, kad kuo daugiau modelis buvo mokytas tam tikrais duomenimis, tuo tiksliau jis atsako į bendro pobūdžio klausimus¹⁴. Vadinasi, pateikus detalią užklausą, susijusią su tokiais sritimis kaip medicina ar teisė, bendros paskirties DKM turi didesnę riziką haliucinuoti, nei pateikus užklausas susijusias su bendrais, visuomenei žinomais faktais.

Atsižvelgiant į tai, kad DKM per se yra BPD modeliai, iš kurių yra sudarytos DI sistemos, haliucinacijų reiškinyms yra reikšmingas siekiant išanalizuoti DI akto nustatytus reikalavimus. Autorių aptarta haliucinacijų problema bei atsiradimo priežastys aktualiausia tampa didelės rizikos DI sistemų bei joms keliamų reikalavimų kontekste, atsižvelgiant į tai, kad šiai grupei priskirtų DI sistemų generuojama informacija itin glaudžiai susijusi su asmens sauga, sveikata bei teisėmis.

II. Didelės rizikos DI sistemoms keliami reikalavimai haliucinacijų reiškinių kontekste

Vien tai, kad DI sistemos dažnai veikia kaip „juodoji dėžė“, savaime nereiškia, kad jų keliamų rizikų negalima suvaldyti. ES priėmė DI aktą, kurio paskirtis – nustatyti vienodą teisinę sistemą, skirtą DI sistemoms, skatinant į žmogų orientuoto ir patikimo DI įsisavinimą ir užtikrinti Europos Sąjungos pagrindinių teisių chartijoje (toliau – Chartija) numatytą žmogaus teisių apsaugą¹⁵. Siekiant įgyvendinti šio straipsnio tikslą, būtina įvardyti, kokie reikalavimai taikomi DI sistemoms. Šioje da-

¹³ WOLF, M. J., MILLER, K. & GRODZINSKY, F. S. (2017). Why we should have seen that coming: comments on Microsoft's Tay „experiment,“ and wider implications. *ACM SIGCAS Computers and Society*, 47(3), 54–64 [interaktyvus]. <https://doi.org/10.1145/3135043.3135050>. p. 1

2016 m. „Microsoft Tay“ nebuvo DKM, o paprastesnę architektūrą turintis tekstą generuojančių modelis, tačiau apmokymo duomenų įtaka modelio veikimui yra tokia pati ar panaši į DKM.

¹⁴ HUANG, L., YU, W., MA, W. ir kt. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2), 1–55 [interaktyvus]. <https://dl.acm.org/doi/full/10.1145/3703155>. p. 9

¹⁵ DI akto 1 konstatuojamoji dalis.

lyje bus analizuojama DI sistemų klasifikacija, esminį dėmesį skiriant didelės rizikos klasifikacijai priskiriamoms DI sistemoms bei joms keliamiems reguliaciniams reikalavimams, kurie yra aktualiausi nagrinėjant haliucinacijas.

2.1. DI sistemų klasifikacija pagal DI aktą

Esminis DI akto komponentas yra keturių DI sistemų rizikos¹⁶ klasių apibrėžimas, pateikiamas lentelėje Nr. 3.

3 lentelė. DI sistemų klasifikavimas į rizikos grupes pagal DI aktą



Vaizduojamoje piramidėje aukščiausioje pakopoje nurodoma nepriimtinos rizikos klasifikacija, į kurią priskiriamos visos DI sistemos, kurios pagal savo veikimo principą kelia grėsmę žmonių saugumui, bei jų teisių apsaugai, ir yra uždraustos (pvz., socialiniai reitingai). Viena pakopa žemiau nurodoma didelės rizikos klasifikacija, į kurią priskiriami DI naudojimo atvejai, kurie gali kelti grėsmę sveikatai, saugai, pagrindinėms teisėms (pvz., DI priemonės, naudojamos nuotoliniam biometriniam tapatybės nustatymui). Dar žemiau piramidėje pateikiama skaidrumo rizikos klasifikacija, kurioms DI aktas nustato konkrečias informacijos atskleidimo pareigas, siekiant užtikrinti kad žmonės būtų informuoti dėl DI sistemos veikimo (pvz., internetinių puslapių konsultantai). Žemiausios rizikos klasifikacijos, nurodomos piramidėje, DI aktas nereglementuoja dėl minimalaus bei rizikos nekeliančio DI sistemų pobūdžio.

Šaltinis: sugeneruotas ir aprašytas autorių.

Sistemų rizikos kategorija nustatoma atsižvelgiant į numatytą DI sistemos paskirtį, t.y. sistemų klasifikacija priklauso nuo DI sistemos atliekamos funkcijos ir nuo konkretaus sistemos naudojimo tikslo bei sąlygų. Remiantis DI akto dėstomosios

¹⁶ Rizikos sąvoka turėtų būti suprantama taip, kaip ji numatyta DI akto 1 straipsnio 2 dalyje.

dalies trečiame skyriuje įtvirtintais kriterijais, DI sistemos gali būti laikomos keliančiomis didelę riziką dviem atvejais: (i) kai DI sistema yra kaip saugos komponentas integruota į gaminius, kuriems taikomi galiojantys gaminių teisės aktai¹⁷, arba pati yra toks gaminys¹⁸ (pvz., DI grindžiama medicininė programinė įranga), arba (ii) kai DI sistema yra skirta naudoti į didelės rizikos kategoriją¹⁹ patenkančias naudojimo atvejais²⁰ (pvz., vertinant fizinių asmenų kreditingumą).

Pirmuoju atveju, didelė rizika kyla nepriklausomai nuo konkretaus naudojimo konteksto, nes jei ir DI sistema būtų naudojama saugioje aplinkoje, jos numatomas pavojus yra pakankamai didelis ir reikalaujantis griežtos kontrolės (pvz., medicininis įrenginys su DI palaikoma diagnostika²¹, skirtas analizuoti medicininius vaizdus). Nepaisant to, kad DI sistema naudojama tik su patyrusių gydytojų priežiūra, ji vis tiek priskiriama didelės rizikos klasifikacijai, nes klaida, įvykusi dėl netinkamai tokios DI sistemos sugeneruotos informacijos, gali sukelti tiesioginę grėsmę pacientų sveikatai ar gyvybei.

Antruoju atveju, didelė rizika kyla atsižvelgiant į jų numatytą paskirtį, kuomet jos kelia žalą fizinių asmenų sveikatai ir saugai arba pagrindinėms teisėms, vertinant tiek galimos žalos sunkumą, tiek jos atsiradimo tikimybę, ir DI sistemos naudojamos keliose konkrečiai iš anksto apibrėžtose srityse, nurodytose DI akte²². Šiuo aspektu numatoma išimtis iš bendrosios taisyklės – DI sistema, skirta naudoti į didelės rizikos kategoriją patenkančias naudojimo atvejais, nebus laikoma didelės rizikos, jeigu: (i) ji nekels esminės žalos sveikatai, grėsmės saugumui ar pagrindinių žmogaus teisių užtikrinimui, ir (ii) atvejais, kai DI sistema bus naudojama tik atliekant iš anksto siaurai apibrėžtą pagalbinę funkciją, pavyzdžiui, atliekant parengiamąją funkciją prieš galutinį vertinimą, ar tais atvejais kai DI sistema neatlieka fizinių asmenų profiliavimo paskirties²³. Nors tokia DI sistema būtų naudojama kaip pagalbinė priemonė, siekiant palengvinti žmogaus priimamą galutinį sprendimą, tokia sistema profiliuoja fizinius asmenis pagal jų kompetencijas ir asmenines savybes, todėl visais atvejais bus klasifikuojama kaip DI sistema, kelianti padidintą riziką žmogaus gerovei bei teisėtiems interesams.

¹⁷ DI akto I priedas.

¹⁸ DI akto 6 straipsnio 1 dalies a) punktas.

¹⁹ DI akto III priedas.

²⁰ DI akto 6 straipsnio 1 dalies b) punktas.

²¹ Išsamus medicininis priemonių saugos reglamentavimas nustatomas 2017 m. balandžio 5 d. Europos Parlamento ir Tarybos Reglamente 2017/745 dėl medicinos priemonių, L 117.

²² DI akto 52 konstatuojamoji dalis.

²³ Fizinių asmenų profiliavimo paskirties pavyzdžiu, autorių nuomone, gali būti laikoma DI sistema, kurios paskirtis – įvertinti kandidatų tinkamumą konkrečiai darbovietai, pagal kandidatų pateiktus gyvenimo aprašymus.

2.2. Didelės rizikos DI sistemoms keliami reikalavimai

DI sistemoms, kurios pagal DI aktą priskiriamos didelės rizikos kategorijai, taikomi griežti ex ante reikalavimai, susiję su: rizikos valdymu, DI sistemų naudojamų duomenų kokybe, technine DI sistemų dokumentacija, dokumentų registracija, skaidrumu, saugumu, patikimumu bei tikslumu ir žmogaus atliekamos priežiūros būtinybe. Šie reikalavimai gali būti skirstomi į materialiuosius ir procedūrinius, atsižvelgiant į tai, kokiam aspektui reguliuoti reikalavimai yra numatyti. Procedūriniai (pvz., rizikos valdymo sistemos diegimas, žmogaus atliekama priežiūra) skirti reguliuoti didelės rizikos DI sistemos atitiktį, tuo tarpu materialieji reikalavimai (pvz., duomenų kokybė, skaidrumas naudotojams) skirti nustatyti standartus, kurie užkirstų kelią neigiamam rezultatui. Tačiau ne visi reikalavimai prima facie gali būti tiesiogiai siejami su haliucinacijų reiškiniu – autorių nuomone, atsižvelgiant į haliucinacijų reiškinio specifiką, esminis dėmesys turėtų būti skiriamas: (i) rizikos valdymo sistemos etapams, (ii) DI sistemų naudojamų duomenų kokybei, bei (iii) žmogaus atliekamai priežiūrai.

Rizikos valdymo sistema. DI akto 9 straipsnyje numatoma rizikos valdymo sistema, kuri nustato pozityvią ir tęstinę pareigą didelės rizikos DI sistemų diegėjams sukurti, įgyvendinti, dokumentuoti ir prižiūrėti rizikos valdymo sistemą, skirtą keliama rizikai valdyti. Tai yra tęstinis procesas, kuriame rizikos valdymas privalo būti atliekamas visą DI sistemos veikimo ciklą – kuriant, plėtojant, diegiant bei naudojant didelės rizikos DI sistemas. DI aktas numato keturis rizikos valdymo sistemos etapus – rizikos identifikavimą, rizikos vertinimą, rizikos peržiūrą po pateikimo rinkai bei rizikos valdymo priemonių taikymą. Rizikos identifikavimo bei vertinimo esmė – nustatyti kuriamos DI sistemos numatomas rizikas fizinio asmens sveikatai, saugai bei Chartijos saugomoms teisėms ir įvertinti galimos žalos sunkumą bei tikimybę. Šiuose etapuose išsamiai atlikta analizė sudaro prielaidas taikyti rizikos mažinimo priemones, tokias kaip numatomą žmogaus priežiūrą, diegiant papildomas apsaugos priemones nuo šališkumo ir diskriminacijos. Tačiau, tobulėjant technologijoms svarbu numatyti tai, kad pati DI sistema, po pateikimo rinkai gali sukelti incidentų, susijusių su DI akte numatytais saugomomis vertybėmis, dėl to pradėjus diegti DI sistemas, siekiant suvaldyti riziką svarbu nuolat stebėti sistemos veikimą, o įvykus incidentams – imtis prevencinių veiksmų. Šie keturi rizikos valdymo sistemos etapai privalo būti įgyvendinami visą DI sistemos gyvavimo ciklą, tokiu būdu nustatant DI sistemų diegėjų nuolatinę pareigą valdyti didelės rizikos DI sistemų keliama riziką.

Haliucinacijų reiškinio kontekste, rizikos valdymo sistema: (i) įvardija haliucinacijas kaip rizikos šaltinį (pvz., galimą Chartijoje numatytų teisių pažeidimą), (ii) įvertina DI sistemos haliucinavimo tikimybę (pvz., kai DI sistemos nebuvo mokytos panašiais į pateiktą užklausą pavyzdžiais) bei haliucinavimo poveikį (pvz., jei DI sistema pateiks realybės neatitinkančią atsakymą, kuris potencialiai pažeis asmens teises ar jo teisėtus interesus), (iii) nustato rizikos mažinimo priemones (pvz., apriboja DI sistemą, kuomet trūksta duomenų pateikti atsakymą). Haliucinavimo tikimybės bei

poveikio vertinimas didelės rizikos DI sistemos kontekste yra itin svarbus, atsižvelgiant į tai, kad šios DI sistemos glaudžiai sąveikauja su asmenų sauga, sveikata bei teisėmis, tad tikrovės neatitinkančios informacijos pateikimas, šiuo atveju gali lemti tai, jog bus pažeistas DI akto tikslas bei jame nustatyti reikalavimai.

DI sistemų naudojamų duomenų kokybė. DI akto 9 straipsnyje numatytų privalomų duomenų kokybės ir duomenų valdymo reikalavimų didelės rizikos DI sistemoms tikslas – užtikrinti, kad kuriamos ir vėliau rinkai pateikiamos DI sistemos nebūtų šališkos, nekeltų diskriminacijos rizikos bei veiktų saugiai ir patikimai, atsižvelgiant į jų numatytą paskirtį²⁴. DI modelių mokymo mokymo, validavimo ir bandymo duomenų rinkiniai privalo atitikti šiame straipsnyje nustatytus reikalavimus. Duomenų rinkimas, bei naudojimo paskirtis turi būti aiškūs bei surinkti, duomenų valdymas privalo būti siejamas su didelės rizikos DI sistemos paskirtimi, jų kilmė bei rinkimo tikslas turi būti aiškiai dokumentuoti bei suderinami su didelės rizikos DI sistemos paskirtimi. Tiekėjai yra įpareigojami nurodyti, ką duomenys atspindi, įvertinti ar jų pakanka, ar jie atitinka konkrečią didelės rizikos DI sistemos paskirtį ir ar juose nėra esminių spragų. Duomenys turi tvarkomi siekiant mažinti DI sistemos veikime pasitaikančių klaidų kiekį bei duomenų šališkumą. Tiekėjai, valdydami kylančią riziką privalo įvertinti, ar duomenys, naudojami apmokyti didelės rizikos DI sistemas gali pakenkti fizinių asmenų sveikatai, saugai ar pažeisti Chartijoje numatytas teises sukeliant draudžiamą diskriminaciją²⁵. Šališkumo nustatymo ir ištaisymo tikslais, DI aktas nustato išimtis, susijusias su specialiu kategorijų asmens duomenimis²⁶, suteikiančią galimybę tvarkyti duomenis, susijusius su rasės, etninės kilmės, religijos, lytinės orientacijos aspektais, pagrindžiant šių duomenų tvarkymo būtinybę bei ribojant prieigos kontrolę. DI akto 10 straipsnio 3 dalyje yra numatomi duomenų rinkinių kokybės reikalavimai – tinkamumas, reprezentatyvumas, išsamumas ir teisingumas. Šiame straipsnyje nurodyti reikalavimai paverčia duomenų kokybę bei valdymą teisiniu faktu, kurio nepaisymas gali būti teisės pažeidimas, iš kurio kyla atitinkamos neigiamos pasekmės²⁷. Atitinkamai, jei duomenys, kuriais yra apmokytos DI sistemos, neatitinka duomenų rinkinių kokybės reikalavimų, kyla pagrįsta abejonė teigti, kad sistema, generuodama atsakymus į pateiktas užklausas, negalės atsakyti teisingai ir taps nepatikima jos naudotojui.

Žmogaus atliekama priežiūra. DI esmė – į žmogų orientuotas požiūris²⁸. Būtent dėl to, didelės rizikos DI sistemų kontrolė ir priežiūra yra vienas iš aktualiausių klausimų.

²⁴ CASONATO, C. & OLIVATO, G. (2024). *AI Regulation in Europe: Exploring the Artificial Intelligence Act. Digital Environments and Human Relations: Ethical Perspectives on AI Issues*, vol. 150. Cham: Springer, p. 101. [interaktyvus] Prieiga per internetą: https://doi.org/10.1007/978-3-031-76961-0_5

²⁵ Chartijos 21 straipsnis.

²⁶ BDAR 9 straipsnio 1 punktą.

²⁷ Haliucinacijų reiškiniu kontekste, duomenų kokybė, autorių nuomone, yra esminis aspektas – DI sistemų pateikiami rezultatai tiesiogiai priklauso nuo duomenų, kuriais jos apmokytos.

²⁸ DI akto 1 konstatuojamoji dalis.

simų. Dėl keliamos rizikos, didelės rizikos DI sistemos autonominis veikimas potencialiai sukeltų grėsmę žmogaus gerovei, todėl DI akto 14 straipsnis numato žmogaus atliekamos priežiūros būtinybę²⁹. DI aktas numato, kad didelės rizikos DI sistema nuo pat projektavimo pradžios turi būti pritaikyta veiksmingai žmogaus priežiūrai, siekiant sistemingai valdyti riziką, kylančią žmogaus sveikatai, saugai ar pagrindinėms teisėms, kai sistema naudojama tiek pagal numatytąją paskirtį, tiek numatomo netinkamo naudojimo sąlygomis. DI aktas numato dvi žmogaus priežiūros priemonių rūšis: priemonės, kurios projektavimo metu diegiamos pačioje sistemoje, pvz., galimybė sustabdyti procesą, o kai šios priemonės techniškai įgyvendinti neįmanoma – taikyti procedūrinės priemonės pačiam diegėjui, pvz., darbuotojų apmokymą. Viena svarbiausių uždavinių, keliamų žmonėms, atsakingiems už priežiūros įgyvendinimą – įvertinti, kad DI sistema ne visuomet pateikia tinkamą rezultatą, ir, atsižvelgiant į tai, vengti automatinio šališkumo (angl. automation bias), ypač tokiame kontekste, kuomet didelės rizikos DI sistema teikia rekomendacijas ar padeda priimti sprendimus. Haliucinacijų reiškinių kontekste, siekiant išvengti tokio šalutinio DKM veikimo, žmogaus priežiūros reikalavimas gali tapti išsigelbėjimu, padedančiu užtikrinti didelės rizikos DI sistemų atitiktį numatytiems reikalavimams.

Apibendrinant, haliucinacijų reiškinyje DI akto kontekste yra aktualiausias didelės rizikos klasifikacijai priskiriamoms DI sistemoms, atsižvelgiant į DI sistemoms atliekamą funkciją ir DI sistemos naudojimo tikslo bei sąlygų. Šiame skyriuje aptarti esminiai DI reikalavimai, haliucinacijų reiškinių kontekste yra aktualūs tuo, kad: (i) rizikos valdymo sistemos reikalavimai įtvirtina pareigą kurti DI sistemą taip, jog haliucinacijų reiškinyje DI sistemose būtų aptinkamas kuo įmanoma greičiau, tokiu būdu siekiant valdyti neigiamus šio šalutinio DKM veikimo padarinius, (ii) duomenų kokybės reikalavimas užtikrina, kad duomenys, kuriais mokoma sistema, atitiktų kriterijus, kurie potencialiai eliminuotų galimybę haliucinacijų reiškiniui pasireikšti didelės rizikos DI sistemose, o (iii) žmogaus atliekamos priežiūros reikalavimas būtų įgyvendinamas kaip prevencinė priemonė, santykiyje tarp DI sistemos ir vartotojo, tokiu būdu valdant haliucinuojančios DI sistemos veikimą bei keliamą riziką.

III. Duomenų mokslo vaidmuo valdant haliucinacijų reiškinių problemą didelės rizikos DI sistemose

Atsižvelgiant į aptartas haliucinacijų priežastis ir didelės rizikos DI sistemoms keliamus reikalavimus, šiame skyriuje bus nagrinėjami haliucinacijų mažinimo metodai. Aukštos rizikos klasifikacijai priskirtose DI sistemose, kuriose sprendimai gali turėti saugos, sveikatos ar fundamentalių žmogaus teisių pažeidimų, tokie netikslumai yra nepriimtini, todėl, siekiant įgyvendinti DI akto reikalavimus ir užtikrinti

²⁹ žr. DI akto 14 straipsnį.

sistemų patikimumą, būtina taikyti papildomus techninius ir organizacinius haliucinacijų mažinimo bei rizikos valdymo metodus.

3.1. Haliucinacijų reiškinių valdymo metodai

Autorių nuomone, siekiant valdyti haliucinacijų reiškinių keliamą riziką didelės rizikos DI sistemų veikimui bei atitikčiai, esminis dėmesys turėtų būti skiriamas šiems metodams: (i) paieškos papildymo generacijai, bei (ii) neapibrėžtumo įvertinimui³⁰.

Paieškos papildymo generacija kaip haliucinacijų mažinimo metodas. Vienas iš metodų, naudojamų mažinti haliucinacijas DKM, yra vadinamas paieškos papildymo generacija (angl. Retrieval-Augmented Generation (RAG))³¹. Šis metodas kombinuoja paieškos ir generavimo principus, kai, prieš pradedant generuoti atsakymą į pateiktą užklausą, remiantis išoriniais šaltiniais atrenkama užklausiai aktuali ir svarbi informacija. Galutinis atsakymas formuojamas remiantis tiek naudotojo užklausa, tiek išoriniais dokumentais, taip sumažinant haliucinacijas, kylančias dėl modelio žinių spragų³². Taip pat yra šio metodo versija, kai modelis papildomai ieško informacijos. Tokiu atveju, papildomi duomenys atrenkami ne tik prieš generavimą, bet ir viso atsakymo formavimo metu. Toks metodas ypač naudingas sprendžiant sudėtingas, kelių žingsnių samprotavimo ar ilgų atsakymų generavimo užduotis, nes leidžia nuosekliai tikslinti modelio išvadas ir sumažinti faktinių klaidų kaupimąsi samprotavimo grandinėje. Kai kuriose šio metodo realizacijose taikoma ir vėlesnė informacijos patikra, kurios metu jau sugeneruotas atsakymas lyginamas su išoriniais šaltiniais ir, esant neatitikimams, koreguojamas³³. Taip mažinamos I dalyje minėtos haliucinacijų atsiradimo priežastys, susijusios su nebeaktualia informacija ar nevienodu DKM apmokymo metu naudotos informacijos kiekiu. Tokio mechanizmo integravimas DKM atveria galimybę taikyti šį metodą srityse, kurios reikalauja itin konkrečios ir tikslios informacijos (pvz., advokatūroje, DI pasitelkiant kaip pagalbinę priemonę generuoti informaciją, naudojamą rengiant ieškinius).

Neapibrėžtumo įvertinimas kaip haliucinacijų rizikos mažinimo metodas. Paieškos papildymo generacijos metodas neapsaugo nuo atvejų, kai DKM neturi užtikrinto

³⁰ Autorės pažymi, kad šiame darbe nėra aptariami visi galimi haliucinacijų reiškinių mažinimo metodai, o akcentuojami yra tie, kurie, autorių nuomone, yra veiksmingiausi bandant suvaldyti haliucinacijų reiškinių keliamą riziką.

³¹ ALANSARI, A. & LUQMAN, H. (2025). Large Language Models Hallucination: A Comprehensive Survey. *arXiv* [interaktyvus]. p. 2

³² HUANG, L., YU, W., MA, W. ir kt. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2), 1–55 [interaktyvus]. <https://dl.acm.org/doi/full/10.1145/3703155>. p. 21–22

³³ GAO, L., DAI, Z., PASUPAT, P. ir kt. (2023). RARR: Researching and Revising What Language Models Say, Using Language Models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16477–16508 [interaktyvus]. 10.18653/v1/2023.acl-long.910. p. 2

atsakymo. Neapibrėžtumas (angl. uncertainty) atsakyme atsiranda dėl informacijos trūkumo, duomenyse esančių klaidų ir prieštarių, ar sprendimo proceso sudėtingumo, kurie neleidžia modeliui pateikti tikslių rezultatų³⁴. Tokiems atvejams yra naudojami neapibrėžtumo įvertinimo (angl. uncertainty quantification) metodai, kurie prie DKM atsakymų pateikia užtikrintumo (angl. confidence) balą, nusakantį, kaip stipriai modelis užtikrintas konkrečiu sugeneruotu atsakymu³⁵. Svarbu pabrėžti, kad modelio apibrėžtumas negali būti prilyginamas modelio užtikrintumui. Modelis gali turėti žemą apibrėžtumo lygį, bet aukštą užtikrintumo lygį, pavyzdžiui, kai jis užtikrintai pateikia atsakymą apie retą medicininę būklę, nors jo apmokymo duomenyse tokios informacijos buvo itin mažai arba iš viso nebuvo. Neapibrėžtumo įvertinimo metodai remiasi vertinant kelių to pačio modelio atsakymų į tos pačios užklauskos variaciją³⁶. Jei modelis nebuvo apmokytas duomenimis, turinčiais informaciją, kurios reikalauja užklausa, atsakymų variacija, tikėtina, bus didelė. Dėl šios priežasties neapibrėžtumo įvertinimo metodai tampa ypač reikšmingi aukštos rizikos DI sistemoje, kuriose neteisingas, bet užtikrintai pateiktas atsakymas gali pažeisti DI akto numatytą tikslą – asmens gerovę. Tokie metodai leidžia ne tik identifikuoti situacijas, kuriose DKM atsakymais pasikliauti pavojinga, bet ir iš anksto numatyti papildomus saugiklius, pavyzdžiui, žmogaus įsikišimą ar sprendimo atidėjimą, taip didinant bendrą sistemos patikimumą ir saugumą. Medicinoje, neapibrėžtumo įvertinimo metodai gali pažymėti nepakankamai užtikrintas DKM diagnozes ir reikalauti papildomo specialisto įvertinimo, taip sumažindamos diagnostinių klaidų skaičių iki 40 proc.³⁷.

3.2. DI haliucinacijų mažinimo metodų vertinimas

Nors aptarti metodai sumažina haliucinacijų riziką, jų taikymas praktikoje susiduria su reikšmingais techniniais ir organizaciniais iššūkiais, neleidžiančiais šiais metodais remtis besąlygiškai. Pavyzdžiui, paieškos papildymo generacijos metodo pagrindinis trūkumas yra jo priklausomumas nuo atrinktų išorinių šaltinių kokybės

³⁴ FAKOUR, F., MOSLEH, A. & RAMEZANI, R. (2024). A Structured Review of Literature on Uncertainty in Machine Learning & Deep Learning. *arXiv* [interaktyvus]. <https://arxiv.org/abs/2406.00332>. p. 6-7

³⁵ LIU, X., CHEN, T., DA, L. ir kt. (2025). Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. V. 2, 6107–6117 [interaktyvus]. <https://doi.org/10.1145/3711896.3736569>. p. 1

³⁶ LIU, X., CHEN, T., DA, L. ir kt. (2025). Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. V. 2, 6107–6117 [interaktyvus]. <https://doi.org/10.1145/3711896.3736569>. p. 3

³⁷ LIU, X., CHEN, T., DA, L. ir kt. (2025). Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. V. 2, 6107–6117 [interaktyvus]. <https://doi.org/10.1145/3711896.3736569>. p. 1

ir aktualumo³⁸. Jei paieškos metu pateikiami nereikšmingi ar netikslūs duomenys, jie gali pabloginti galutinio atsakymo tikslumą. Be to, dabartinėse papildymo generacijos metodo realizacijose, paieškos ir generavimo komponentai dažnai neveikia optimaliai, todėl kai DKM remiasi informacija iš šaltinių, jie ne visada yra tinkamai išnaudojami³⁹. Neapibrėžtumo įvertinimo metodo trūkumai yra susiję su didelėmis skaičiavimo sąnaudomis ir menku patikimumo pagerėjimu, ypač taikant neapibrėžtumo metodus DMK⁴⁰. Taip pat, haliucinacijas lemiantis neapibrėžtumas gali kilti ir iš pateiktos užklauskos dviprasmiškumo. Be to, BPGI modelių atveju gali atsirasti nesuderinami neapibrėžtumo įvertinimai, nes tokio pobūdžio modeliai gali užklauskas priimti ne tik natūralia kalba, bet ir nuotraukų ar garso įrašų pavidalu. Akcentuojama, kad dauguma neapibrėžtumo įvertinimo metodų neturi aiškaus skaitinio įvertinimo, tad gali būti sudėtinga interpretuoti gautą neapibrėžtumo įvertinimą⁴¹. Šie trūkumai parodo, kad nei vienas metodas savaime neužtikrina patikimo haliucinacijų suvaldymo, nes jie paveldi tas pačias sistemines problemas kaip ir patys modeliai. Žmogaus atliekama priežiūra, tiek teoriniu, tiek praktiniu aspektu, gali būti išieitimi, siekiant valdyti riziką, kylančią haliucinacijų reiškinių paveiktoje DI sistemoje. Derinant skirtingus haliucinacijų mažinimo metodus, galima iš dalies kompensuoti atskirų metodų trūkumus. Vis dėlto efektyvus haliucinacijų suvaldymas reikalauja ne pavienių sprendimų, o nuosekliai vertinamų metodų kombinacijų, taikomų realiomis sąlygomis. Be to, neturint aiškių vertinimo metrikų ir aukštos kokybės duomenų, net ir pažangiausių metodų taikymas praktikoje išlieka riboto patikimumo.

Atlikta literatūrinė analizė atskleidžia, kad bendros paskirties DKM, nors ir pasižymintys plačiu funkcionalumu, susiduria su struktūriniais apribojimais taikant juos didelės rizikos DI sistemų kontekste, kuriuose būtinas aukštas informacijos tikslumo, atsakymų atsekamumo ir elgsenos nuspėjamumo lygis. Didžioji dalis haliucinacijų reiškinių priežasčių yra susiję su DKM architektūra, kuri yra sudėtinga ir sunkiai interpretuojama. Dėl itin didelio apmokymui naudojamų duomenų kiekio tampa sudėtinga tiksliai nustatyti haliucinacijų kilmę, kokybę bei kontroliuoti filtravimo

³⁸ HUANG, L., YU, W., MA, W. ir kt. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2), 1–55 [interaktyvus]. <https://dl.acm.org/doi/full/10.1145/3703155>. p. 26

³⁹ BARNETT, S., KURNIAWAN, S., THUDUMU, S. ir kt. (2024). Seven Failure Points When Engineering a Retrieval Augmented Generation System. *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI*, 194–199 [interaktyvus]. <https://doi.org/10.1145/3644815.3644945>. p. 4

⁴⁰ LIU, X., CHEN, T., DA, L. ir kt. (2025). Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, V. 2, 6107–6117 [interaktyvus]. <https://doi.org/10.1145/3711896.3736569>. p. 8

⁴¹ LIU, X., CHEN, T., DA, L. ir kt. (2025). Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, V. 2, 6107–6117 [interaktyvus]. <https://doi.org/10.1145/3711896.3736569>. p. 4

procesus; be to, sunku numatyti, kokį šališkumą tokie modeliai gali perimti iš duomenų ir kokiame kontekste DKM turi didesnę riziką haliucinuoti. Papildomą riziką didina ir tai, kad didžiausius ir populiariausius modelius (tokius kaip GPT, Gemini) kuriančios privačios įmonės tarpusavyje konkuruoja, siekdamos kuo greičiau pristatyti pažangesnius sprendimus. Toks spaudimas dažnai lemia vis didesnių duomenų rinkinių naudojimą, o tai gali neigiamai paveikti duomenų atrankos kokybę. Be to, apmokymo proceso metu skiriamas dėmesys modelių suderinamumui su vartotoju, siekiant pagerinti naudojimo patirtį. Nors tai padidina naudojimo patogumą, toks optimizavimas gali kartu didinti ir haliucinacijų riziką, nes modeliai labiau orientuojasi į norimą atsakymo stilių, o ne į faktinį tikslumą⁴².

Srityse, kuriose sprendimai tiesiogiai veikia žmonių sveikatą, saugumą ar gerovę, DKM taikymas turėtų būti vertinamas itin atsargiai. Tokiose srityse net ir nedidelės informacinės paklaidos gali turėti reikšmingų pasekmių, todėl modelių polinkis generuoti haliucinacijas tampa kritiniu veiksniu. Šiame straipsnyje aptarti haliucinacijų mažinimo metodai neeliminuoja haliucinacijų reiškinių bei jo keliamos rizikos, todėl, atitinkamai, autorės pabrėžia, kad DI sistemų pateikiamais atsakymais besąlygiškai remtis nereikėtų.

Tačiau, kaip buvo aptarta analizėje, DKM gali pasitarnauti atsakant į bendrus klausimus, tad gali netiesiogiai padėti spręsti profesinius iššūkius. Priimant svarbiausius sprendimus, būtų galima naudotis mažesniais kalbos modeliais. Tyrimai rodo, kad mažesni, vienoje srityje apmokyti kalbos modeliai gali pateikti tikslesnius ir patikimesnius atsakymus nei didelio masto bendros paskirties modeliai, ypač kai užklausa reikalauja specifinių, kontekstinių žinių^{43 44}. Tokie modeliai nėra priversti apibendrinti informacijos iš labai plataus teminio lauko, todėl jų atsakymai dažniau remiasi tiksliais ir aktualiais duomenimis, priešingai nei DKM, kurie apmokomi naudojant viešai prieinamus duomenimis, kuriuose dažnai trūksta specializuotos ar profesionalios informacijos. Tuo tarpu kuriant mažesnius, siauresnės paskirties modelius, atsiranda galimybė sąmoningai kontroliuoti, kokie duomenų šaltiniai naudojami, kokia informacija patenka į mokymo procesą ir kokios žinios yra formuojamos. Tai leidžia ne tik geriau suprasti modelio ribas, bet ir užtikrinti didesnę skaidrumą bei patikimumą, ypač tais atvejais, kai modelio sprendimai turi realų poveikį žmonėms.

⁴² ALANSARI, A. & LUQMAN, H. (2025). Large Language Models Hallucination: A Comprehensive Survey. *arXiv* [interaktyvus]. <https://arxiv.org/abs/2510.06265>. p. 8

⁴³ LI, H.; AI, Q., CHEN, J. ir kt. (2025). BLADE: Enhancing Black-Box Large Language Models With Small Domain-Specific Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23), 24422–24430 [interaktyvus]. <https://doi.org/10.1609/aaai.v39i23.34620>. p. 1

⁴⁴ LU, Y., YAO, B., ZHANG, S. ir kt. (2023). Human Still Wins over LLM: An Empirical Study of Active Learning on Domain-Specific Annotation tasks. *arXiv* [interaktyvus]. <https://arxiv.org/abs/2311.09825>. p. 4

Apibendrinant, būtų galima teigti, kad haliucinacijų rizika nėra atsitiktinio pobūdžio reiškinys – jo pasireiškimas kyla iš DKM architektūros ir veikimo principų. Modeliai, sukurti veikti itin plačiame žinių ir užduočių spektre, neišvengiamai susiduria su ribotu gyliu specifiniuose kontekstuose. Haliucinacijų mažinimo metodai susiduria su tais pačiais sisteminiais iššūkiais, kaip ir patys DKM. Dėl šios priežasties bendrosios paskirties DKM taikymas didelės rizikos DI sistemose sukelia saugos iššūkius, kurių neįmanoma visiškai eliminuoti pavienėmis techninėmis priemonėmis. Autorės siūlo alternatyvą, siekiant valdyti haliucinacijų reiškinio keliamą riziką ir užtikrinti DI akte numatytus reikalavimus – nuolatinę žmogaus atliekamą didelės rizikos DI sistemos priežiūrą, kuri tuo pačiu pabrėžtų, kad: (i) DI sistemos vis dar turi techninių trūkumų, ir (ii) jog šiai dienai besąlygiškai remtis DI sistemų pateikiama informacija galimybės nėra.

Išvados

DI haliucinacijos yra neatsiejamos nuo DKM architektūros ir šių modelių apmokymui skirtų duomenų pobūdžio. Kadangi DKM pažingsniui dėlioja tekstą, jie gali prarasti generuojamo teksto kontekstą ir taip sugeneruoti neteisingą atsakymą. Be to, DKM negeba išreikšti neužtikrintumo, tad, kai susiduria su užklausa, į kurią atsakymo neturi, didėja tikimybė, kad modelis haliucinuos. Šias problemas dar labiau sustiprina mokymo duomenyse esanti klaidinga informacija, kuria modelis remiasi, pateikdamas atsakymus į užklausas. Tai patvirtina, kad techninis DKM supratimas yra būtinas pamatas nustatant veiksmingas DI sistemų valdymo priemones. Kadangi DKM per se yra DI modeliai, kurie sudaro DI sistemas, ši problema tampa aktuali ir DI akto kontekste, siekiant užtikrinti nustatytus reikalavimus.

DI akte įtvirtinta DI sistemų rizikos klasifikacija pagrįsta DI sistemos galimo poveikio asmens saugai, sveikatai ar Chartijoje numatytoms teisėms. Didelės rizikos DI sistemų kvalifikavimas grindžiamas ne vien jų taikymo sritimi, bet ir numatytąja paskirtimi, kuri lemia padidintą žalos atsiradimo tikimybę. Haliucinacijos šių sistemų veikime įgauna ypatingą reikšmę, kadangi tikrovės neatitinkantys DI generuojami rezultatai gali sukelti reikšmingas neigiamas pasekmes fiziniams asmenims. Haliucinacijų reiškinio kontekste, autorių nuomone, pagrindinis dėmesys turi būti skiriamas: (i) rizikos valdymo sistemos, (ii) duomenų kokybės (iii) žmogaus atliekamos priežiūros reikalavimams.

Haliucinacijų mažinimo metodų analizė rodo, kad paieškos papildymo generacijos metodas efektyviai sprendžia žinių spragų ir informacijos aktualumo problemas, leidžiant modeliui remtis išoriniais šaltiniais tiek prieš generavimą, tiek jo metu. Kartu su neapibrėžtumo įvertinimo metodais, kurie leidžia identifikuoti atvejus, kai DKM neturi užtikrinto atsakymo, šios priemonės iš esmės padidina didelės rizikos DI sistemų patikimumą. Tačiau tyrimas, atliktas šiame straipsnyje, patvirtina, kad duomenų mokslo metodai, nors ir leidžia nuosekliai tikslinti išvadas bei mažinti ha-

liucinacijų kiekį, visiškai haliucinacijų rizikos nepanaikina. Tad siekiant užtikrinti DI sistemų atitiktį DI akto keliams reikalavimams, autorės siūlo akcentuoti nuolatinę žmogaus atliekamą priežiūrą, kaip išeitį valdant haliucinacijų reiškinių keliamą riziką didelės rizikos DI sistemose.

Šaltinių sąrašas

Europos Sąjungos teisės aktai

Europos Parlamento ir Tarybos (ES) 2024 m. birželio 13 d. Reglamentas 2024/1689 kuriuo nustatomos suderintos dirbtinio intelekto taisyklės ir iš dalies keičiami reglamentai (EB) Nr. 300/2008, (ES) Nr. 167/2013, (ES) Nr. 168/2013, (ES) 2018/858, (ES) 2018/1139 ir (ES) 2019/2144 ir direktyvos 2014/90/ES, (ES) 2016/797 ir (ES) 2020/1828 (Dirbtinio intelekto aktas), OJ L, 2024/1689, 12.7.2024, p. 1 – 84, 127 – 129.

Europos Parlamento ir Tarybos (ES) 2016 m. balandžio 27 d. Reglamentas 2016/679 dėl fizinių asmenų apsaugos tvarkant asmens duomenis ir dėl laisvo tokių duomenų judėjimo ir kuriuo panaikinama Direktyva 95/46/EB (Bendrasis duomenų apsaugos reglamentas), OL L 119, 2016 5 4, p. 1–88.

Europos Parlamento ir Tarybos (ES) 2017 m. balandžio 5 d. Reglamentas Nr. 2017/745 dėl medicinos priemonių, kuriuo iš dalies keičiama Direktyva 2001/83/EB, Reglamentas (EB) Nr. 178/2002 ir Reglamentas (EB) Nr. 1223/2009, ir kuriuo panaikinamos Tarybos direktyvos 90/385/EEB ir 93/42/EEB, L 117, 5.5.2017, p. 1–175.

Specialioji literatūra

ALANSARI, A. & LUQMAN, H. (2025). Large Language Models Hallucination: A Comprehensive Survey. arXiv [interaktyvus]. <https://arxiv.org/abs/2510.06265>. [žiūrėta 2025 m. gruodžio 20 d.]

ALBALAK, A., ELAZAR, Y. XIE, S. M. ir kt.(2024). A Survey on Data Selection for Language Models. arXiv [interaktyvus]. <https://arxiv.org/abs/2402.16827>. [žiūrėta 2025 m. gruodžio 19 d.]

BARNETT, S., KURNIAWAN, S., THUDUMU, S. ir kt. (2024). Seven Failure Points When Engineering a Retrieval Augmented Generation System. Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI, 194–199 [interaktyvus]. <https://doi.org/10.1145/3644815.3644945>. [žiūrėta 2025 m. gruodžio 31 d.]

CASONATO, C. & OLIVATO, G. (2024). AI Regulation in Europe: Exploring the Artificial Intelligence Act. Digital Environments and Human Relations: Ethical Perspectives on AI Issues, vol. 150. Cham: Springer, p. 101. [interaktyvus] Prieiga per internetą: https://doi.org/10.1007/978-3-031-76961-0_5 [Žiūrėta 2026 m. sausio 1 d.]

CHANG, Y., WANG, X., WANG, J. ir kt. (2024). A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology, 15(3), 1–45 [interaktyvus]. <https://doi.org/10.1145/3641289>. [žiūrėta 2025 m. gruodžio 15 d.]

- FAKOUR, F., MOSLEH, A. & RAMEZANI, R. (2024). A Structured Review of Literature on Uncertainty in Machine Learning & Deep Learning. arXiv [interaktyvus]. <https://arxiv.org/abs/2406.00332>. [žiūrėta 2025 m. gruodžio 29 d.]
- GAO, L., DAI, Z., PASUPAT, P. ir kt. (2023). RARR: Researching and Revising What Language Models Say, Using Language Models. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 16477–16508 [interaktyvus]. 10.18653/v1/2023.acl-long.910. [žiūrėta 2025 m. gruodžio 28 d.]
- HUANG, L., YU, W., MA, W. ir kt. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems, 43(2), 1–55 [interaktyvus]. <https://dl.acm.org/doi/full/10.1145/3703155>. [žiūrėta 2026 m. sausio 3 d.]
- Ji, W., YUAN, W., GETZEN, E. ir kt. (2025). An Overview of Large Language Models for Statisticians. arXiv [interaktyvus]. <https://arxiv.org/abs/2502.17814>. [žiūrėta 2025 m. gruodžio 15 d.]
- LI, H.; AI, Q., CHEN, J. ir kt. (2025). BLADE: Enhancing Black-Box Large Language Models With Small Domain-Specific Models. Proceedings of the AAAI Conference on Artificial Intelligence, 39(23), 24422–24430 [interaktyvus]. <https://doi.org/10.1609/aaai.v39i23.34620>. [žiūrėta 2025 m. gruodžio 2 d.]
- LIN, Z., GUAN, S., ZHANG, W. ir kt. (2024). Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. Artificial Intelligence Review, 57(9), 243 [interaktyvus]. <https://link.springer.com/article/10.1007/s10462-024-10896-y>. [žiūrėta 2025 m. gruodžio 27 d.]
- LIU, X., CHEN, T., DA, L. ir kt. (2025). Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V. 2, 6107–6117 [interaktyvus]. <https://doi.org/10.1145/3711896.3736569>. [žiūrėta 2025 m. gruodžio 29 d.]
- LIU, Y., YAO, Y., TON, J. F. ir kt. (2023). Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv [interaktyvus]. <https://arxiv.org/abs/2308.05374>. [žiūrėta 2025 m. gruodžio 2 d.]
- LIU, Y., YAO, B., ZHANG, S. ir kt. (2023). Human Still Wins over LLM: An Empirical Study of Active Learning on Domain-Specific Annotation tasks. arXiv [interaktyvus]. <https://arxiv.org/abs/2311.09825>. [žiūrėta 2025 m. gruodžio 31 d.]
- SU, W. (2025). Do Large Language Models (Really) Need Statistical Foundations? arXiv [interaktyvus]. <https://arxiv.org/abs/2505.19145>. [žiūrėta 2025 m. gruodžio 15 d.]
- WOLF, M. J., MILLER, K. & GRODZINSKY, F. S. (2017). Why we should have seen that coming: comments on Microsoft's Tay “experiment,” and wider implications. ACM SIGCAS Computers and Society, 47(3), 54–64 [interaktyvus]. <https://doi.org/10.1145/3135043.3135050>. [žiūrėta 2026 m. vasario 01 d.]