MYKOLO ROMERIO UNIVERSITETO
VIEŠOJO VALDYMO IR VERSLO FAKULTETAS
VERSLO IR EKONOMIKOS INSTITUTAS

IVETA GASPARAVIČIENĖ
KIBERNETINIO SAUGUMO VALDYMAS

DIRBTINIO INTELEKTO RIZIKŲ VALDYMAS

Magistro baigiamasis darbas

Vadovas: **Prof. dr. Mindaugas Kiškis**

Vilnius, 2025

MYKOLAS ROMERIS UNIVERSITY

FACULTY OF PUBLIC ADMINISTRATION AND BUSINESS

INSTITUTE OF BUSINESS AND ECONOMY

IVETA GASPARAVIČIENĖ

CYBER SECURITY MANAGEMENT

RISK MANAGEMENT OF ARTIFICIAL INTELLIGENCE

Master Thesis

Supervisor: **Prof. dr. Mindaugas Kiškis**

Vilnius, 2025

TURINYS

LENTELIŲ SĄRAŠAS

PAVEIKSLŲ SĄRAŠAS

SANTRUMPOS IR SĄVOKOS

1. AI – Dirbtinis intelektas (*angl*. *Artificial Intelligence*)
2. DI – Dirbtinis intelektas (*angl*. *Artificial Intelligence*)
3. ES – Europos Sąjunga (*angl*. *European Union*)
4. EU – Europos Sąjunga (*angl*. *European Union*)
5. JAV – Jungtinės Amerikos Valstijos (*angl*. *United States of America*)
6. ISO – Tarptautinė standartizacijos organizacija (*angl.International Organization for Standardization*)
7. NIST – Nacionalinis standartų ir technologijų institutas (*angl*. *National Institute of Standards and Technology*)
8. RMF – Rizikos valdymo sistema (*angl*. *Risk Management Framework*)
9. GRC – Valdymas, rizika ir atitiktis (*angl*. *Governance, Risk and Compliance*)
10. LLM – Didelis kalbos modelis (*angl*. *Large Language Model*)
11. GPU – Grafinis procesorius (*angl*. *Graphics Processing Unit*)
12. ML – Mašininis mokymasis (*angl*. *Machine Learning*)
13. NLP – Natūralios kalbos apdorojimas (*angl*. *Natural Language Processing*)
14. OECD – Ekonominio bendradarbiavimo ir plėtros organizacija (*angl*. *Organisation for Economic Co-operation and Development*)
15. AI ACT – Europos dirbtinio intelekto reglamentas (*angl*. *Artificial Intelligence Act*)
16. GDPR – Bendrasis duomenų apsaugos reglamentas (*angl*. *General Data Protection Regulation*)
17. BDAR – Bendrasis duomenų apsaugos reglamentas (*lietuviškas GDPR trumpinys*)
18. DSA – Skaitmeninių paslaugų aktas (*angl*. *Digital Services Act*)
19. DMA – Skaitmeninių rinkų aktas (*angl*. *Digital Markets Act*)
20. NIS2 – Tinklo ir informacinių sistemų saugumo direktyva (*angl*. *Directive on Security of Network and Information Systems*)
21. TIS2 – Tinklo ir informacinių sistemų direktyva (*lietuviškas NIS2 trumpinys*)
22. VDAI – Valstybinė duomenų apsaugos inspekcija
23. AI BIAS – Situacija, kai DI sistema priima šališkus, neteisingus ar diskriminacinius sprendimus (*angl*. *Artificial Intelligence Bias*)

ĮVADAS

Temos aktualumas. Spartus dirbtinio intelekto (DI) technologijų vystymasis atveria naujas galimybes ne tik pramonėje ar akademinėje aplinkoje, bet ir kelia vis daugiau klausimų apie jų etišką taikymą, galimus pavojus žmogaus privatumui bei visuomenės saugumui. Šiandien DI sprendimai taikomi labai plačiai – nuo buitinių įrenginių iki pažangių valstybės sektoriaus analitinių sistemų. Nors pagrindinis DI tikslas – padėti spręsti įvairaus pobūdžio užduotis greičiau ir efektyviau, jo gebėjimas apdoroti milžiniškus informacijos kiekius iškelia iššūkių, susijusių su asmens duomenų apsauga, diskriminacijos rizika bei sistemų elgsenos nenuspėjamumu.

Europos institucijos (2024) [1] akcentuoja, kad žmogaus priežiūra turėtų išlikti pagrindine AI sprendimų dalimi. Kitaip tariant, vien technologiškai pažangių sistemų nepakanka – jos turi būti kuriamos taip, kad atitiktų teisės aktus, būtų skaidrios bei atsekamos. Tik suderinus inžinerinius sprendimus su aiškiomis teisinėmis normomis ir etiškai pagrįstomis gairėmis, įmanoma veiksmingai suvaldyti iš naujų technologijų kylančias rizikas.

Augantis DI algoritmų sudėtingumas ir sistemų savarankiškumas skatina tiek mokslininkus, tiek praktikus vis dažniau atkreipti dėmesį į rizikų prevenciją ir valdymą. Galimybė autonomiškai apdoroti ir interpretuoti duomenis, priimant sprendimus su ribota žmogaus intervencija, reikalauja integruoto požiūrio: būtina ne tik taikyti kibernetinio saugumo sprendimus, bet ir užtikrinti atitiktį teisiniams bei etiniams standartams. Šiame darbe išryškinamas sisteminis požiūris į dirbtinio intelekto rizikų vertinimą, pabrėžiant, kad kompleksiškas įvairių sričių – technologinės, socialinės ir teisinės – integravimas leidžia nuosekliau valdyti iššūkius ir apsaugoti visuomenės gerovę

Temos naujumas / teorinis reikšmingumas. Pastaraisiais metais dirbtinio intelekto (DI) keliamų rizikų tema vis dažniau analizuojama ne tik technologiniu, bet ir kur kas platesniu – etiniu, teisiniu bei socialiniu – kampu. Toks poslinkis rodo, kad tradiciniai rizikų valdymo modeliai, orientuoti tik į saugumą ar duomenų apsaugą, ne visada pajėgūs atliepti sudėtingą DI prigimtį. Temos aktualumą ir naujumą lemia poreikis peržengti siaurą disciplininę specializaciją bei taikyti integruotą, tarpdisciplininį požiūrį, aprėpiantį skirtingų sričių sąveiką.

Šiame darbe sąmoningai pasirinkta neapsiriboti vien teoriniais modeliais, bet gilintis į tai, kaip organizacijos realiai atpažįsta, vertina ir sprendžia su DI susijusias rizikas. Tokia praktika grįsta tyrimo kryptis leidžia įvertinti, su kokiais iššūkiais susiduriama taikant egzistuojančias metodikas, ir kokie sprendimai galėtų padidinti jų efektyvumą ar reikalingumą adaptuoti.

[1] „EUROPOS PARLAMENTO IR TARYBOS REGLAMENTAS (ES) 2024/1689 2024 m. birželio 13 d. kuriuo nustatomos suderintos dirbtinio intelekto taisyklės.

Teoriniu požiūriu šis tyrimas reikšmingas tuo, kad siekia sujungti technologinius, organizacinius ir reguliacinius aspektus į nuoseklų analitinį karkasą. Tokiu būdu formuojamas pagrindas atsakingiems sprendimams, kurie padeda užtikrinti DI sistemų patikimumą, etišką naudojimą ir teisinį atitikties užtikrinimą organizacijų veikloje.

Mokslinė problema – Kaip užtikrinti, kad dirbtinio intelekto technologijų taikymas organizacijose vyktų saugiai, valdant su tuo susijusias rizikas?

Tyrimo objektas – Dirbtinio intelekto rizikų valdymo procesai Lietuvos organizacijose.

Tyrimo dalykas – Praktinės priemonės ir sprendimai, padedantys organizacijoms identifikuoti, vertinti bei valdyti DI keliamas rizikas.

Tyrimo tikslas – Išanalizuoti ir įvertinti, kaip Lietuvos organizacijose identifikuojamos ir valdomos su dirbtiniu intelektu susijusios rizikos, siekiant nustatyti taikomų sprendimų efektyvumą bei atotrūkius nuo tarptautinių modelių.

Tyrimo uždaviniai:

- Išanalizuoti mokslinę literatūrą, nagrinėjančią dirbtinio intelekto keliamas rizikas ir jų valdymo metodus;
- Įvertinti tarptautinius DI rizikų valdymo modelius (pvz., ISO, NIST AI RMF, GRC) bei jų pritaikomumą organizacijų kontekste;
- Atlikti ekspertinį tyrimą, siekiant suprasti, kaip Lietuvos organizacijos identifikuoja, vertina ir sprendžia su DI susijusias rizikas;
- Pasiūlyti praktines rekomendacijas ar modelio koncepciją, kuri galėtų padėti organizacijoms veiksmingiau valdyti DI rizikas.

Duomenų rinkimo metodai:

1. Mokslinės literatūros analizė ir sisteminimas;
2. Pusiau struktūruoti ekspertiniai interviu (kokybinio tyrimo strategija).

Duomenų analizės metodai:

1. Kokybinė turinio analizė;
2. Teminė interviu duomenų analizė ir interpretavimas;
3. Lyginamoji analizė su teoriniais modeliais.

1. DIRBTINIO INTELEKTO RAIDA IR SAMPRATA

Šiuolaikinė visuomenė vis labiau priklauso nuo skaitmeninių technologijų, todėl klausimas, kas yra dirbtinis intelektas (DI), įgauna ypatingą reikšmę tiek akademiniu, tiek praktiniu požiūriu. Mokslinėje literatūroje vieningos DI sąvokos iki šiol nėra, nes šios technologijos aprėptis ir taikymo sritys yra labai plačios – nuo algoritmų, kurie analizuoja ir mokosi iš didelių duomenų rinkinių, iki autonomiškai sprendimus priimančių robotų realiose situacijose. Kadangi DI gali atlikti tokias užduotis, kurios anksčiau buvo siejamos tik su žmogaus intelektu – pavyzdžiui, kalbos atpažinimą, vertimą ar kompleksinių problemų sprendimą – aiškaus ir vieningo jo apibrėžimo klausimas iki šiol kelia daug diskusijų . [2] (Russell ir Norvig, 2021).

1.1. Dirbtinio intelekto sąvoka

Pirmieji žingsniai, lėmę dirbtinio intelekto (DI) tyrimų pradžią, buvo žengti XX a. pirmoje pusėje, kai matematika ir mechanizuotas skaičiavimas pradėjo užimti reikšmingą vietą kriptografijos srityje. Vienas iš ryškesnių to meto veikėjų – lenkų matematikas ir kriptoanalitikas Marian Rejewski, kuris 1932 m. sugebėjo iššifruoti vokiečių „Enigmos“ kodą. Be to, jis sukūrė pirmąją automatizuotą dešifravimo sistemą – mechanizmą, padėjusį pamatus vėlesniems Alano Turingo darbams, įskaitant ir garsųjį „Bombe“ įrenginį, sukurtą Bletchley parke. Rejewski savo pasiekimais pasidalijo su sąjungininkais Britanijoje ir Prancūzijoje, taip svariai prisidėdamas prie pažangos sprendžiant problemas mechanizuotai. Šie ankstyvieji laimėjimai vėliau tapo pamatiniais principais, formuojant DI tyrimų metodikas ir technologinius pagrindus.

Ryškesni DI tyrimų užuomazgų kontūrai išryškėjo Antrojo pasaulinio karo metais, kuomet Bletchley parko kriptografijos laboratorijose (Didžioji Britanija) buvo taikomi pažangūs matematikos ir inžinerijos sprendimai, siekiant perprasti milžiniškus kiekius koduotų pranešimų. Šioje aplinkoje išsiskyrė A. Turingas – ne tik kaip kriptoanalitikas, bet ir kaip mąstytojas, kėlęs esminį klausimą: ar įmanoma sukurti mąstančią mašiną?

Po karo jis ėmėsi plėtoti naujas, netradicines idėjas, tęsiant ir plėtojant tokių mokslininkų kaip M. Rejewski pradėtus darbus. 1950 m. jis publikavo žymųjį straipsnį „Computing Machinery and Intelligence“, kuriame pasiūlė Turingo testą[3] . Šio testo esmė – jei žmogus, bendraudamas su mašina (dažnai tekstiniu būdu), negali atskirti jos nuo kito žmogaus, galima teigti, kad mašina „mąsto“. Šis principas suformavo pagrindą DI tyrimams, orientuojant mokslininkus ne tik į techninius sprendimus, bet ir į „intelekto“ esmės tyrimus.

[2] Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.
[3] Turing, A. (1950). *Computing Machinery and Intelligence.* Mind, 59(236), 433–460.

Vis dėlto tuo metu nebuvo vieningo supratimo apie tai, kas apskritai yra „intelektas“. Psichologai, tokie kaip D. Wechsler (1944), jį apibrėžė kaip gebėjimą mokytis, mąstyti racionaliai ir veikti efektyviai aplinkoje[4] , o filosofai intelektą siejo su sąmone ir kūrybingumu. Todėl vystantis „dirbtinio“ intelekto idėjai, į ją įsitraukė įvairių disciplinų – matematikos, inžinerijos, psichologijos, lingvistikos – tyrėjai, kurių požiūriai į intelektą dažnai skyrėsi.

Šis disciplinų susiliejimas sukūrė terpę, kurioje pradėti svarstyti keli esminiai klausimai:

- Ar mašina gali tiksliai atkartoti žmogaus mąstymo procesą?
- Ar DI veikiau yra praktinis metodas, skirtas sistemingai spręsti sudėtingas problemas, reikalaujančias didžiulių duomenų kiekių apdorojimo?
- Kaip į sparčiai tobulėjančias „mąstančias“ sistemas integruoti etinius, teisės ir socialinius aspektus?

Dartmuto konferencija ir sąvokos „artificial intelligence“ gimimas

Po ankstyvųjų A. Turingo teorijų ir kriptologijos pasiekimų Antrojo pasaulinio karo metu, reikšmingas lūžis dirbtinio intelekto (DI) istorijoje įvyko 1956 m. Dartmuto konferencijoje. Šios konferencijos iniciatorius, kompiuterių mokslo pionierius Johnas McCarthy, pakvietė Marviną Minsky, vieną iš MIT Dirbtinio intelekto laboratorijos steigėjų, Claude’ą Shannoną, pripažįstamą „informacijos teorijos tėvu“, Nathanaeilį Rochesterį, IBM inžinierių bei pirmųjų kompiuterių architektą, ir kitus žymius to laikotarpio mokslininkus aptarti, kaip kompiuteriais galima imituoti arba net pranokti žmogaus mąstymo procesus.

Būtent šio susitikimo metu prigijo terminas „***artificial intelligence***[5]“ (liet.k „Dirbtinis Intelektas“), kuris netrukus tapo naujos mokslo disciplinos pavadinimu. DI pradėjo formuotis kaip tarpdisciplininis laukas, apjungiantis informatiką, matematiką, psichologiją ir net lingvistiką, taip siekiant išspręsti sudėtingas problemas ir gilintis į žmogaus intelekto esmę.

Simbolinio DI pakilimai ir „dirbtinio intelekto žiemos“

XX a. septintajame–aštuntajame dešimtmečiuose formavosi dirbtinio intelekto (DI) disciplina, dar vadinama simboliniu (taisyklių pagrįstu) DI, kuri tuo metu išgyveno savo pakilimo etapą. Kuriamos ekspertinės sistemos, tokios kaip DENDRAL[6] ir MYCIN[7], rėmėsi ranka užrašytomis taisyklėmis bei

[4] Wechsler, D. (1944). The Measurement of Adult Intelligence. Williams & Wilkins. – 3psl.

[5] McCarthy, J. (1956). *Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.* Dartmouth College.

[6] DENDRAL: A case study of the first expert system for scientific hypothesis formation R. K. Lindsay , B. G. Buchanan , E. A. Feigenbaum , J. Lederberg Volume 61, Issue 2, 1993, 209-261psl.

[7] MYCIN artificial intelligence program B.J. Copeland „Encyclopedia Britannica“

loginiais algoritmais [8](Feigenbaum, 1977). Tačiau tuometiniai techniniai ribojimai ir sudėtinga duomenų rinkimo aplinka greitai atskleidė šių sistemų ribotumą sprendžiant dinamiškus ir sunkiai nuspėjamus realaus pasaulio procesus. Todėl daugelio lūkesčiai smarkiai prasilenkė su galimybėmis[9] (Minsky, 1985; Russell ir Norvig, 2021). Susidariusi atotrūkio tarp teorijos ir praktikos situacija lėmė kelis „dirbtinio intelekto žiemų“ laikotarpius – laikus, kai finansavimas ir visuomenės susidomėjimas DI technologijomis pastebimai sumažėdavo.

Mašininio mokymosi pradmenys atgimstantys neuroniniai tinklai

Kol simbolinis DI dominavo, septintajame dešimtmetyje kai kurie mokslininkai pradėjo tirti biologiniais principais grįstas sistemas, remdamiesi idėja, kad žmogaus smegenų neuroninę veiklą galima modeliuoti matematiškai [10] [11](McCulloch ir Pitts, 1943; Rosenblatt, 1958). Tačiau tuo metu menka aparatinės įrangos galia, neefektyvūs mokymo algoritmai ir riboti duomenų kiekiai neleido šioms idėjoms pilnai atsiskleisti. Apie devintąjį dešimtmetį neuroniniai tinklai vėl atsidūrė mokslininkų akiratyje. Vienas iš lūžio taškų buvo 1986 metai, kai Rumelhart, Hinton ir Williams pristatė vadinamąjį atgalinio sklidimo (angl. backpropagation) algoritmą[12]. Tai leido gerokai tiksliau reguliuoti vidinius tinklo parametrus ir ženkliai padidino tinklų gebėjimą mokytis iš duomenų. Netrukus vis daugiau dėmesio pradėta skirti ne taisyklėmis grįstam programavimui, o modelių mokymui remiantis pačiais duomenimis – būtent tai ir nulėmė tolesnį šios srities augimą.

Didieji duomenys ir gilusis mokymasis: kelias į dabartį

XXI amžiaus pradžia žymėjo esminį DI evoliucijos etapą. Interneto plėtra, socialinių tinklų populiarumas ir kiti skaitmeniniai kanalai pradėjo generuoti milžiniškus duomenų masyvus (angl. ***big data***). Šis procesas, kartu su sparčiai tobulėjančiomis aparatinės įrangos technologijomis, ypač grafikos procesoriais (GPU), sukūrė sąlygas giliojo mokymosi (angl. deep learning) atgimimui [13] (Goodfellow, Bengio ir Courville, 2016). Šiuolaikinės DI sistemos išsiskiria tuo, kad jos jau nebėra griežtomis taisyklėmis pagrįstos programos – tai savaime besimokantys modeliai, gebantys aptikti sudėtingus dėsningumus iš milžiniškų duomenų šaltinių.

Duomenimis grįsto mokymosi pranašumai itin išryškėjo tokiose srityse kaip vaizdų atpažinimas, kalbos sintezė, autonominis judėjimas ar natūralios kalbos apdorojimas (angl. NLP). Pastarasis sektorius sparčiai progresuoja – ypač pokalbių robotų (angl. chatbot) srityje, kurie šiandien jau tapo įprasta

[8] Feigenbaum, E. A**.** (1977). The Art of Artificial Intelligence: Themes and Case Studies of Knowledge Engineering. *IJCAI*, 77, 1014–1029.

[9] Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson

[10] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133

[11] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408.

[12] Rumelhart, D. E., Hinton, G.E., & Villiams, R.J. (1986). Learning representations by back-propagating errors. Nature, 323, 533-536.

[13] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

priemone ne tik verslo procesams automatizuoti, bet ir kasdieniam vartotojų bendravimui su technologijomis. Skirtingai nei ankstyvosios ekspertinės sistemos, gilusis mokymasis leidžia algoritmams ne tik spręsti konkrečias problemas, bet ir tobulėti bei prisitaikyti prie naujų sąlygų be tiesioginės žmogaus intervencijos.

Šiuolaikinė DI samprata: tarp kognityvinio atkartojimo ir praktinio efektyvumo

Būtent nuo 1956 m. Dartmuto konferencijos, kur ***artificial intelligence*** sąvoka įgavo oficialų statusą, iki šių dienų DI suvokimas nuolat keitėsi.

1 lentelė. „Dirbtinio intelekto apibrėžimai“

Autoriai	**Metai**	**D.I. apibrėžimas / požiūris**	**Raktažodžiai**
A. Turing	1950	Mašinos gebėjimas taip imituoti žmogų, kad bendraujant neįmanoma pasakyti, ar kalbamasi su žmogumi, ar su kompiuteriu.	Turingo testas, žmogaus imitacija
McCarthy (1956)	1956	Mokslas ir inžinerija, skirti protingoms mašinoms kurti, orientuojantis į loginių, racionalių funkcijų automatizavimą.	„Protingos mašinos“, inžinerija
Minsky (1985)	1985	Dirbtinis intelektas – psichologijos ir kompiuterių mokslo sandūra, kurioje ieškoma, kaip algoritmiškai suvokti ir atkartoti žmogaus protą.	Sąmonės modeliavimas, „proto visuomenė“
Goodfellow, Bengio ir Courville (2016)	2021	Metodų visuma, leidžianti kompiuterinėms sistemoms mokytis spręsti uždavinius tiesiogiai iš duomenų, neturint iš anksto užprogramuotų taisyklių.	Giliojo mokymosi algoritmai, didieji duomenys
Russell ir Norvig (2021)	2021	„Sistemų, galinčių suvokti aplinką ir veikti taip, kad padidintų sėkmingo rezultato tikimybę, tyrimas ir kūrimas.“	Racionalus veikimas, aplinkos suvokimas

Šaltinis: parengta autorės, remiantis Turing (1950), McCarthy (1956), Minsky (1985), Goodfellow, Bengio ir Courville (2016), Russell ir Norvig (2021)

Lentelėje pateiktos skirtingų autorių dirbtinio intelekto (DI) apibrėžimai atskleidžia šios srities sampratos įvairovę – nuo žmogaus proto imitacijos iki pragmatiško, duomenimis grįsto metodo. Nepaisant šios įvairovės, DI vystymosi kontekste išskiriamos trys pagrindinės kryptys, kurios atsispindi tiek teoriniuose tyrimuose, tiek praktiniame taikyme.

Tačiau galima išskirti tris pagrindines DI raidos kryptis:

- **Simbolinis (taisyklėmis pagrįstas) DI** Tikėta, kad logika ir formalių taisyklių rinkiniai gali atkartoti žmogaus mąstymą. Viliantis, jog tinkamai „užprogramavus protą“, kompiuteriai galėtų atlikti net sudėtingas užduotis, pavyzdžiui, medicinines diagnozes.
- **Biologinė inspiracija** Neuroniniai tinklai ir mokymosi algoritmai, kurių tikslas – modeliuoti smegenų veiklą. Manyta, kad „dirbtinės smegenys“ gebės mokytis iš patirties ir prisitaikyti prie dinamiškos aplinkos.
- **Duomenimis grįstas mokymasis (**angl. **„mashine learning“)** Pastaraisiais dešimtmečiais dėmesys sutelktas į didžiulių duomenų analizę, kur algoritmai autonomiškai atranda dėsningumus. Ši kryptis svarbi srityse nuo rekomendacinių sistemų iki autonominių transporto priemonių.

Dėl šių skirtingų krypčių DI sąvoka išlieka nevienareikšmė. Vieni mokslininkai siekia kuo tiksliau atkurti žmogaus mąstymo procesus, o kiti labiau koncentruojasi į efektyvius, praktiškus sprendimus, nesvarbu, ar jie atitinka žmogiškojo intelekto logiką. Vis dažniau pritariama požiūriui, kad dirbtinis intelektas – tai ne vien technologijų rinkinys, bet ir tarpdisciplininė mokslo bei inžinerijos sritis, leidžianti kompiuteriams savarankiškai atlikti užduotis, kurias iki šiol laikėme būdingomis žmogaus intelektui (Russell ir Norvig, 2021). Šis apibrėžimas neapsiriboja vien techniniu aspektu – jis pabrėžia ir DI vaidmenį visuomenėje, kartu iškeldamas klausimus apie etiką, teisinį reguliavimą bei atsakomybę, kurie būtini norint užtikrinti šių technologijų saugų ir patikimą taikymą.

1.2. Dirbtinio intelekto tipai ir klasifikacijos

Kaip jau minėta anksčiau, dirbtinio intelekto (DI) sąvoka yra plati ir nevienareikšmė – ji aprėpia ne tik technologinius sprendimus, bet ir skirtingus teorinius požiūrius. DI raida parodė, kad šios srities formavimuisi įtakos turėjo įvairios disciplinos – nuo matematikos iki psichologijos – kurios nulėmė dabartines tyrimų kryptis ir taikymo sritis. Tačiau vien tik bendras apibrėžimas nesuteikia visapusiško supratimo apie šią sritį. Todėl svarbu atskirai analizuoti DI tipus ir jų klasifikavimo principus – tai leidžia geriau suprasti, kokias pažintines funkcijas atlieka skirtingos sistemos, kaip jos veikia ir kokiose situacijose jas tikslinga taikyti.

Skirtingos DI klasifikacijų kryptys

Nors keturių rūšių modelis (angl. Reactive AI, Limited Memory AI, Theory– of– Mind AI, Self–Aware AI) dažnai išskiriamas kaip vienas žinomiausių DI klasifikacijos pavyzdžių, mokslinėje

literatūroje aptinkama ir kitų skirstymo būdų. Klasifikacijos gali skirtis priklausomai nuo to, ar dėmesys sutelktas į:

- **Funkcionalumą.** Pavyzdžiui, pokalbių robotai (chatbots), robotinės platformos, rekomendacinės sistemos ar autonominio vairavimo algoritmai.
- **Mokymosi metodą.** Akcentuojama, ar DI veikia pagal simbolinį mokymąsi (taisyklių rinkiniai), mašininį mokymąsi, gilųjį mokymąsi (angl. deep learning), evoliucinius algoritmus ir kt.
- **Autonomiškumo lygmenį.** Aiškinama, kiek sistema geba priimti sprendimus savarankiškai ir kokio žmogaus įsikišimo jai reikia.

Nepaisant to, kasdienėje praktikoje šioms sistemoms dažniausiai taikomas būtent „keturių DI rūšių“ matmuo (angl. *four types of AI*), nes jis aiškiai parodo, kokią kognityvinę veiklą bei prisitaikymą sugeba pasiūlyti tam tikras algoritmas ar robotas [14] (IBM data and AI Team 2024).

1. Reaktyviosios mašinos (angl. *Reactive AI*) – tai tokio tipo dirbtinio intelekto sistemos, kurios sukurtos vykdyti aiškiai apibrėžtas, konkrečias užduotis. Jos neturi atminties, todėl veikia tik tuo momentu turima informacija – tai reiškia, kad jos negali pasimokyti iš ankstesnės patirties ar kaupti žinių apie aplinką. Nors šis apribojimas riboja jų gebėjimą adaptuotis, tokios sistemos pasižymi aukštu efektyvumu siaurose srityse. Pavyzdžiui, žaidimuose kaip šachmatai ar Go jos dažnai pranoksta net ir itin patyrusius žaidėjus (Russell ir Norvig, 2021). 1997 m. ji nugalėjo pasaulio čempioną G. Kasparovą, vertindama tik esamą žaidimo figūrų būseną ir galimus ėjimus, nes neturėjo platesnės strateginės atminties. Panašiu principu veikia Google AlphaGo, kuri daugiausia remiasi momentinių žaidimų figūrų išdėstymu. Nors tokios sistemos jau turi tam tikrų mokymosi savybių ir geba prisitaikyti prie priešininko veiksmų, joms vis dar trūksta ilgalaikės atminties bei galimybės kaupti patirtį.

2. Ribotosios atminties (angl. Limited Memory AI) sistemos jau gali trumpam išsaugoti informaciją apie neseniai įvykusius įvykius, o tai leidžia joms tiksliau reaguoti į besikeičiančią situaciją. Skirtingai nei reaktyviosios mašinos, šios sistemos sugeba mokytis iš trumpalaikės patirties, tačiau jos vis dar neturi gebėjimo apdoroti ilgalaikių duomenų ar suvokti platesnio konteksto (Russell ir Norvig, 2021). Vienas geriausių pavyzdžių – autonominiai automobiliai. Jie fiksuoja ir analizuoja neseniai matytus ženklus, kitų automobilių judėjimą ar kelio sąlygas, kad galėtų realiuoju laiku priimti sprendimus – kada stabdyti, keisti eismo juostą ar koreguoti kryptį. Tokios sistemos naudojamos projektuose kaip „Tesla Autopilot“ ar „Waymo15“, kur DI jutikliai ir kameros renka bei apdoroja informaciją, ribotai modeliuojant dinamišką aplinką16 (Goodfellow, Bengio ir Courville, 2016).

[14] IBM Data and AI Team 2024 „The four types of AI based on functionalities“ prieiga per internetą https://www.ibm.com/think/topics/artificial-intelligence-types (žiūrėta 2025-01-02)

[15] Waymo 2024 „Self-driving cars – Autonomuos Vehicles“ https://waymo.com (žiūrėta 2025-01-02)

[16] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

Nepaisant privalumų, ribotosios atminties DI nesukuria „istorinės“ atminties: surinkti duomenys naudojami tik tol, kol jie svarbūs artimiausiam sprendimui. Jei aplinka iš esmės pasikeičia (pvz., pasikeičia kelių infrastruktūra ar atsiranda netikėti trikdžiai), sistema negali pasikliauti sukaupta ilgalaike patirtimi, todėl ją reikia adaptuoti naujai situacijai. Nepaisant to, ribotosios atminties DI pasižymi didesniu lankstumu nei reaktyviosios sistemos, nes geba bent iš dalies įvertinti aplinkos kontekstą ir prisitaikyti prie netikėtų situacijų.

3. **„Proto teorijos“ dirbtinis intelektas (angl. Theory– of– Mind AI) –** šios DI sistemos remiasi psichologine idėja, teigiančia, kad agentas (žmogus ar mašina) geba suprasti, jog kiti veikėjai turi savus tikslus, nuostatas ir emocijas, darančius įtaką jų elgsenai. Nors žmonėms šis gebėjimas vystosi nuo vaikystės, dirbtinis intelektas dar tik pradeda kurti sistemas, galinčias modeliuoti sudėtingą socialinę sąveiką 17(Hintze, 2016).

Pagrindinis šios DI sistemos tikslas – plėtoti mašinas ar agentus, gebančius prognozuoti ir interpretuoti kitų veiksmus, pavyzdžiui, suprasti, kodėl žmogus renkasi tam tikrą sprendimą ar kaip jis elgsis tam tikrose situacijose [18](Bostrom, 2014). Tam kuriamos emocijų atpažinimo funkcijos, kūno kalbos bei kalbos intonacijos analizės technologijos ir kiti socialiniai signalai. Pavyzdžiui, eksperimentiniai socialiniai robotai bando įvertinti žmogaus veido išraiškas ar balso toną ir atitinkamai pakoreguoti savo reakcijas.

Nors šioje srityje jau nemažai nuveikta, dirbtinis intelektas, paremtas vadinamąja „proto teorija“, vis dar išlieka daugiausia mokslinių tyrimų bei laboratorinių eksperimentų objektu. Tokios sistemos susiduria su reikšmingais iššūkiais – nuo milžiniškos duomenų įvairovės iki žmogaus elgsenos nepastovumo ir sunkiai prognozuojamų reakcijų. Be to, jų kūrimui reikia itin daug skaičiavimo galios, o kartu atsiveria ir neišvengiami etiniai bei teisiniai klausimai, kuriems kol kas trūksta aiškių atsakymų. Kiek toli turėtume leisti mašinai „skaityti mintis“ ar bandyti numatyti žmogaus ketinimus? Tokie klausimai tampa vis aktualesni, kai DI pradedamas taikyti socialiniuose kontekstuose, kurie gali paveikti žmonių privatumą ar sukelti diskriminacines situacijas [19] (Poole, Mackworth ir Goebel, 1998).

4. **Savimonę turintis DI (angl. Self– Aware AI)** – tai kol kas tik teorinis DI raidos etapas, kuriame sistema gebėtų ne tik atpažinti kitus agentus, bet ir sąmoningai suvoktų save kaip atskirą egzistuojantį subjektą. Tokia sistema galėtų priimti sprendimus remdamasi savarankiškais „tikslais“ ar „interesais“, kas kelia ypač sudėtingų etinių ir teisinių klausimų[20] (Searle, 1980). Šiuo metu tokio lygio DI egzistuoja tik mokslinės fantastikos kūriniuose ar filosofiniuose svarstymuose – praktikoje jis dar nepasiektas. Savimonę turinčio DI vystymas kelia rimtų klausimų, susijusių su technologiniais

[17] A. Hintze 2016 „Understanding the Four Types of Artificial Intelligence“ https://www.govtech.com/computing/understanding-the-four-types-of-artificial-intelligence.html (žiūrėta 2025-01-02)

[18] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

[19] Poole, D., Mackworth, A., & Goebel, R. (1998). *Computational Intelligence: A Logical Approach*. Oxford University Press.

[20] Searle, J.R. (1980). Mindi, Brains and Programs. Behavioral and Brain Sciences, 3(3), 417-424

sprendimais, etinėmis dilemomis ir teisinėmis atsakomybės ribomis. Tarp svarbiausių iššūkių iškyla sistemos autonomija, aiškaus atsakomybės paskirstymo poreikis bei galimi interesų konfliktai tarp žmogaus ir mašinos. Jei tokio tipo dirbtinis intelektas kada nors būtų sukurtas, jo poveikis žmogaus privatumui bei visuomeniniams santykiams galėtų būti milžiniškas. Todėl jau dabar pradedama kalbėti apie tai, kokių naujų teisinių priemonių ir reguliavimo mechanizmų gali prireikti (Bostrom, 2014).

Nors savimonę turinčio DI dar nėra, pati jo idėja jau kurį laiką skatina aktyvias diskusijas – apie tai, kokį vaidmenį tokios technologijos galėtų užimti visuomenėje ir kaip jas reikėtų valdyti. Tokie svarstymai padeda formuoti atsakingesnį požiūrį į inovacijas, kad jos neprasilenktų su etikos normomis ir teisiniais principais.

Apibendrinant keturias pagrindines DI kategorijas – nuo paprasčiausių reaktyviųjų sistemų iki teorinio savimonės lygio – aiškėja, kad skirtingi tipai pasižymi nevienodu pažinimo gebėjimų ir savarankiškumo lygiu. Šie skirtumai svarbūs ne tik technologiniu, bet ir etiniu bei teisiniu požiūriu. Todėl, kalbant apie DI veikimą, rizikas ir reguliavimą, tampa aišku, kad vieno požiūrio nepakanka – kiekvienai sistemai būtina rasti tinkamiausią, atsakingą taikymo būdą, atsižvelgiant į bendrą visuomenės interesą.

1.3. Dirbtinio intelekto veikimo komponentai ir mokymosi procesai

Dirbtinio intelekto (DI) veikimas remiasi glaudžia tarpusavyje susijusių komponentų sąveika, kuri leidžia sistemoms analizuoti aplinką, priimti sprendimus ir vykdyti sudėtingas užduotis[21]. Vienas pagrindinių šių sistemų elementų – DI agentai, kurie, kaip nurodo Wooldridge (2009), yra sistemos branduoliai, atsakingi už informacijos rinkimą, analizę ir veiksmų atlikimą, siekiant optimizuoti veiklą pagal aplinkos sąlygas[22].

Ne mažiau svarbų vaidmenį atlieka neuroniniai tinklai, kurie, kaip pažymi Goodfellow, Bengio ir Courville (2016), suteikia DI galimybę apdoroti didžiulius duomenų kiekius, atpažinti sudėtingus dėsningumus ir mokytis iš patirties. Šios technologijos plačiai naudojamos vaizdų atpažinime, kalbos sintezėje ir rekomendacinėse sistemose[23].

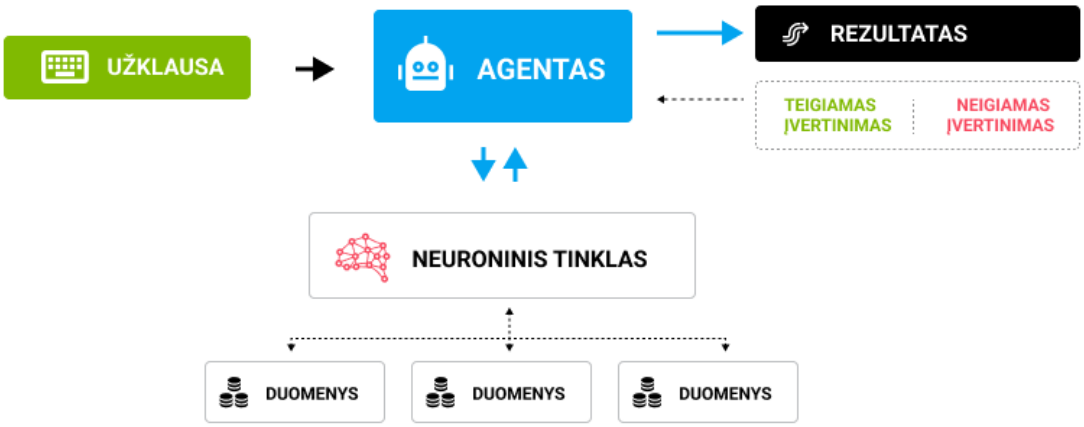
Šio skyriaus tikslas – aptarti DI veikimo principus, nuo agentų funkcijų iki mokymosi procesų, pateikiant praktinio taikymo pavyzdžių.

[21] Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson

[22] Wooldridge, M. (2009). *An Introduction to MultiAgent Systems* (2nd ed.). John Wiley & Sons.

[23] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. Prieiga per: http://www.deeplearningbook.org (žiūrėta 2025-01-09)

1 pav. Dirbtinio intelekto schema



Šaltinis: Sudaryta autorės remiantis S.Russell ir P.Norvig knyga "Artificial Intelligence: A Modern Approach" (2021 m.).

1.3.1. Dirbtinio intelekto agentai ir jų funkcijos

DI agentas yra centrinis sistemos elementas, kuris apdoroja informaciją ir priima sprendimus pagal iš anksto nustatytus algoritmus arba naudojant mokymosi modelius. Agentas gali būti paprastas, taisyklėmis pagrįstas mechanizmas arba sudėtinga savarankiškai besimokanti sistema, gebanti adaptuotis prie kintančių sąlygų.

Pagrindinės agentų funkcijos:

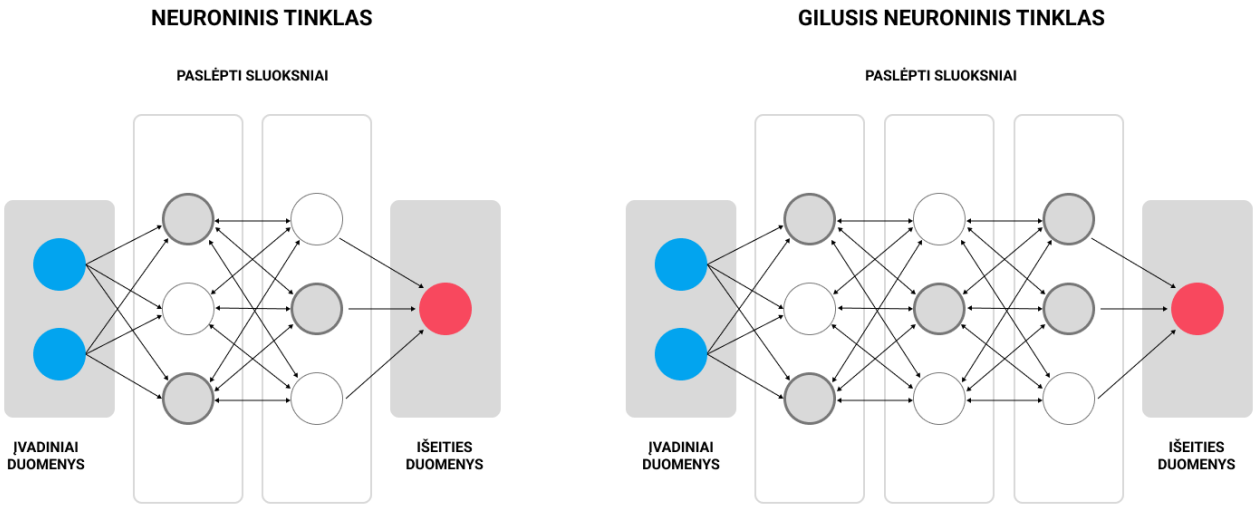
1. **Suvokimas** *(angl. Perception)* – dirbtinio intelekto agentai aplinkos informaciją renka per įvairius jutiklius: vaizdo kameras, mikrofonus, infraraudonuosius spindulius ar kitus sensorinius įrenginius. Pavyzdžiui, autonominiai automobiliai pasitelkia kameras, radarus ir lidarus tam, kad galėtų stebėti kelią, eismo dalyvius ir ženklus.
2. **Apdorojimas ir mokymasis** *(angl. Processing and Learning)* kai duomenys jau surinkti, kitas žingsnis – jų apdorojimas naudojant įvairius algoritmus. Būtent čia ypač svarbią vietą užima mašininis mokymasis – jo dėka sistemos ne tik analizuoja informaciją, bet ir su kiekviena patirtimi tampa išmanesnės, gebėdamos nuolat tobulinti savo veikimą pagal ankstesnius rezultatus.
3. **Sprendimų priėmimas** *(angl. Decision Making)* – išanalizavę informaciją, agentai modeliuoja galimus veiksmų scenarijus ir, įvertinę situacijos kontekstą bei tikėtinus rezultatus, pasirenka tinkamiausią veikimo kryptį.
4. **Veikimas** *(angl. Actuation)* – po sprendimo priėmimo agentas vykdo konkrečius veiksmus : pavyzdžiui, autonominis automobilis valdo vairą, stabdžius ar akseleratorių, o pramoninis robotas vykdo nurodytą judesį gamybos linijoje.

Ši struktūra padeda užtikrinti, kad DI agentai galėtų efektyviai sąveikauti su savo aplinka ir spręsti užduotis, remiantis savo suvokimu ir mokymosi procesais.

1.3.2. Neuroninių tinklų struktūra ir mokymasis

Neuroniniai tinklai – tai viena iš kertinių dirbtinio intelekto dalių, pagrįsta idėja, kad kompiuterinės sistemos gali veikti panašiai kaip žmogaus smegenys, atkuriant dirbtinių neuronų veikimą ir jų tarpusavio ryšius. Šis principas leidžia sistemoms mokytis iš duomenų ir spręsti išties sudėtingas užduotis – nuo vaizdų atpažinimo iki kalbos sintezės ar natūralios kalbos analizės. Pagrindiniai neuroninių tinklų komponentai – duomenų sluoksniai, neurono reikšmingumo svoriai ir aktyvacijos funkcijos – užtikrina, kad sistema galėtų analizuoti įvairius duomenų tipus ir aptikti sudėtingus dėsningumus[24].

2 Pav. Neuroniniai tinklai



Šaltinis: Sudaryta autorės remiantis Y. LeCun, Y. Bengio ir G. Hinton (2015) bei I. Goodfellow, Y. Bengio ir A. Courville (2016) darbais.

Dirbtinio neurono sudėtis ir funkcijos

Kiekvienas dirbtinis neuronas yra informacijos apdorojimo vienetas, sudarytas iš šių pagrindinių komponentų:

- **Įvestis** (*angl. Input*) – signalai, gaunami iš aplinkos arba ankstesnių tinklo sluoksnių.
- **Svoriai** (*angl. Weights*) – kiekviena įvestis yra įvertinama atitinkamu reikšmingumo matu, vadinamu „svoriu“, kuris nusako jos svarbą priimant sprendimus. Mokymosi proceso metu šie svoriai dinamiškai koreguojami.

[24] T. Hastie, R. Tibshirani, J Friedman „The Elements of Statistical Learning Data Mining, Inference, and Prediction Second Edition“ (2017)

- **Sumavimo funkcija** (*angl. Summation Function*) – susumuoja visų įvesties signalų ir jų svorių sandaugas, pateikdama sumažintą reikšmę, kuri toliau perduodama aktyvacijos funkcijai.
- **Aktyvacijos funkcija** (*angl. Activation Function*) – nustato, ar gautas sumažintas įverčio rezultatas yra pakankamas, kad neuronas perduotų signalą kitam sluoksniui. Šis procesas leidžia tinklui spręsti sudėtingas ir netiesines problemas.

Šių komponentų veikimo sinergija leidžia neuronui efektyviai reaguoti į įvesties duomenis ir perduoti apdorotą informaciją tolesniam apdorojimui [25](Bishop, 2006).

Neuroninio tinklo struktūra

Neuroninis tinklas paprastai susideda iš trijų pagrindinių sluoksnių:

1. **Įvesties sluoksnis** (*angl. Input Layer*) – pirmasis tinklo sluoksnis, į kurį patenka pradiniai duomenys; jis perduoda informaciją tolesniems, vidiniams tinklo sluoksniams..
2. **Paslėpti sluoksniai** (*angl. Hidden Layers*) – vienas ar keli sluoksniai tarp įvesties ir išvesties dalių, kuriuose vyksta pagrindinis informacijos apdorojimas: čia tinklas mokosi atpažinti duomenų struktūras, dėsningumus ir ryšius.
3. **Išvesties sluoksnis** (*angl.* Output Layer) – paskutinis sluoksnis, pateikiantis galutinį tinklo rezultatą – atsaką, kuris atspindi mokymosi proceso išvadas.

Mokymosi procesas neuroniniuose tinkluose

Neuroniniuose tinkluose mokymasis vyksta tobulinant svorius ir parametrus, naudojant klaidos skleidimo atgal (backpropagation) metodą[26]. Šis metodas analizuoja klaidą, gaunamą lyginant tinklo prognozes su tiksliais atsakymais, ir koreguoja svorius taip, kad klaida būtų minimizuota[27] (LeCun, Bengio ir Hinton, 2015). Klaidos atgalinio sklidimo algoritmas yra neuroninių tinklų mokymosi dalis, leidžianti tinklui nuolat tobulėti, gerinti savo tikslumą ir gebėjimą spręsti sudėtingas problemas.

1.3.3. Duomenų apdorojimo ir saugojimo mechanizmai

Dirbtinio intelekto (DI) veiksmingumas glaudžiai susijęs su tuo, kaip efektyviai valdomi duomenys. Jei sistema neturi galimybės greitai ir patikimai pasiekti reikalingos informacijos, tai gali trukdyti tiek neuroninių tinklų mokymosi procesui, tiek ir trukdyti agentams priimti kokybiškus sprendimus. Todėl itin svarbu sukurti pažangią duomenų saugojimo infrastruktūrą, kuri leistų užtikrinti greitą ir stabilų duomenų pasiekiamumą.

Duomenų saugyklos (*angl.. Data Warehouses*) – dažniausiai naudojamos tais atvejais, kai reikia kaupti ir analizuoti didelius kiekius istorinių duomenų. Jos padeda užtikrinti ilgalaikį duomenų

[25] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[26] Danielis Nelsonas (2024 Unite.AI) „Kas yra Backpropagation?“ https://www.unite.ai/lt/kas-yra-backpropagation/ (žiūrėta 2025-01-09)

[27] Y. LeCun, Y.Bengio & Hinton, G**.** (2015). Deep learning. Nature, 521(7553), 436-444. DOI: 10.1038/nature14539

vientisumą ir yra ypač naudingos sprendžiant analitines užduotis. Kaip pastebėjo vienas šios srities pradininkų William H. Inmon (2002), tinkamai prižiūrimos duomenų saugyklos leidžia ne tik analizuoti praeities įvykius, bet ir išlaikyti aukštą duomenų kokybės lygį ilgalaikėje perspektyvoje[28]. Tačiau norint, kad šios saugyklos būtų naudingos, privalu nuolat jas atnaujinti, kad jos išliktų aktualios ir atitiktų besikeičiančius poreikius.

Duomenų ežerai (*angl. Data Lakes*). Kai duomenų struktūrizavimas iš anksto nėra būtinas, naudojami duomenų ežerai – lankstūs saugojimo sprendimai, leidžiantys talpinti informaciją jos pradiniu formatu. Tokiu būdu galima saugoti ne tik struktūrizuotus, bet ir nestruktūrizuotus duomenis, kurie gali būti analizuojami ar naudojami modelių mokymui vėliau. Bart Baesens (B.Baesens, 2014) savo tyrimuose akcentuoja[29], kad duomenų ežerai palengvina naujų duomenų šaltinių įtraukimą ir suteikia erdvės eksperimentavimui, taip sudarydami palankias sąlygas inovacijoms ir pažangių analitinių modelių vystymui.

Duomenų srautai (*angl. Data Streams*). Kai kurie projektai reikalauja realaus laiko duomenų apdorojimo, todėl pasyvių saugyklų ar duomenų ežerų nepakanka. Tokiais atvejais pasitelkiami vadinamieji duomenų srautai – jie leidžia DI sistemoms reaguoti į įvykius beveik akimirksniu. Tai ypač svarbu situacijose, kur greitis yra lemiamas, pavyzdžiui, finansinių operacijų stebėsenoje ar valdant autonominius automobilius. Tyrimai rodo, kad dirbant su srautine informacija DI geba priimti sprendimus itin greitai, nes sistema nuolat apdoroja naujus duomenis realiuoju laiku.

Apibendrinant galima pasakyti, jog DI veiksmingumas labai priklauso nuo to, kaip tiksliai ir lanksčiai parinkta duomenų infrastruktūra konkrečiai užduočiai. Duomenų saugyklos padeda analizuoti ilgalaikes tendencijas ir išlaikyti kokybę, duomenų ežerai skatina inovacijas bei greitą prisitaikymą prie naujų šaltinių, o duomenų srautai leidžia reaguoti čia ir dabar. Šių sprendimų derinimas leidžia DI sistemoms išlikti patikimoms net ir itin dinamiškoje aplinkoje.

1.3.4. Dirbtinio intelekto mokymo procesai ir duomenų panaudojimas

Dirbtinio intelekto (DI) sistemų tikslumas ir veikimo kokybė labai priklauso nuo to, kaip jos buvo mokytos. Kadangi šie modeliai mokosi iš jau turimų duomenų, ypač svarbu užtikrinti, kad tie duomenys būtų ne tik gausūs, bet ir kokybiški bei patikimi – kitaip sistema gali mokytis neteisingų dėsningumų ar pateikti klaidingus rezultatus. T. Mitchell (1997) teigia, jog gerai atrinkti mokymo duomenys yra vienas iš pagrindinių veiksnių, lemiančių sėkmingą modelio veikimą. Šiame skyriuje aptariamas visas mokymo procesas – nuo duomenų šaltinių atrankos ir paruošimo iki galutinio modelio testavimo bei validacijos.

[28] W.H. Inmon (2002) „Building the data warehouse“ 35psl

[29] Baesens, B. (2014). Analytics in a Big Data World: The Essential Guide to Data Science and its Applications. Wiley.

Duomenų rinkimas ir paruošimas. Kuriant dirbtinio intelekto (DI) sistemą, viskas prasideda nuo duomenų – pirmasis žingsnis yra jų tinkamas surinkimas. Tačiau vien kiekybės čia neužtenka. Labai svarbu, kad tie duomenys atspindėtų realybę, kurioje modelis vėliau bus taikomas – kitaip sistema gali tiesiog „neišlaikyti“ praktikos testo. Jei duomenų šaltiniai parinkti netinkamai arba informacija nekokybiška, tai gali lemti, jog sistema veiks netiksliai, ypač realiose situacijose. Surinkus duomenis, atliekamas jų valymas – pašalinamos trūkstamos, pasikartojančios ar klaidingos reikšmės. K. Murphy (2012) pabrėžia, kad menkiausi neatitikimai duomenyse gali reikšmingai pabloginti modelio veikimą. Be to, tam tikrais atvejais (pvz., kuriant vaizdų atpažinimo sistemas) duomenys turi būti anotuojami, t. y. pažymimi reikšmingi objektai arba atributai[30].

Pagrindiniai mokymo metodai

Baigus duomenų paruošimą, pasirenkamas tinkamiausias mokymo metodas. Metodai skiriasi pagal tai, kokie duomenys naudojami ir kokį grįžtamąjį ryšį modelis gauna.

- **Prižiūrimasis mokymas** (*angl. Supervised Learning*) metodo pagrindas – anotuoti duomenys, turintys aiškias etiketes arba atsakymus. Modelis mokosi lygindamas savo prognozes su faktinėmis reikšmėmis, koreguodamas parametrus klaidų minimizavimui. Šis metodas dažnai taikomas klasifikavimo ir regresijos užduotyse, tokiose kaip teksto priskyrimas kategorijoms ar kainų prognozavimas.
- **Neprižiūrimasis mokymas** (*angl. Unsupervised Learning*) – naudojamas, kai duomenys nėra anotuoti. Modelis pats ieško dėsningumų ir grupių duomenyse, kas leidžia atlikti klasterizavimą arba atrasti naujus šablonus. Toks metodas itin naudingas analizuojant didelius duomenų rinkinius, pavyzdžiui, klientų segmentavime.
- **Sustiprinamasis mokymas** (*angl. Reinforcement Learning*) – metodas, kuris remiasi nuolatine modelio sąveika su aplinka ir „atlygio“ principu, kai už teisingus veiksmus modelis gauna teigiamą grįžtamąjį ryšį, o už netinkamus – neigiamą. Šis modelis dažnai naudojamas autonominiuose robotuose, žaidimų algoritmuose ir autonominio vairavimo sistemose.

Modelio mokymas ir vertinimas

Pasirinkus mokymo metodą, vykdomas pats modelio mokymas. Algoritmas iteratyviai apdoroja duomenis, koreguodamas savo parametrus taip, kad būtų kuo tiksliau sprendžiamos užduotys. Norint išvengti per didelio prisitaikymo prie mokymo duomenų (angl. overfitting), įprastai naudojami du atskiri duomenų rinkiniai:

[30] K. P. Murphy (2012) „Machine learning a probabilistic perspective“

- **Validacijos rinkinys** – skirtas modelio hiperparametrų derinimui ir per didelio prisirišimo prie mokymo duomenų prevencijai.
- **Testavimo rinkinys** – galutinis rinkinys, leidžiantis įvertinti modelio tikslumą su naujais, iki tol nenaudotais duomenimis.

Iteracinis tobulinimas

Dirbtinio intelekto (DI) sistemų kūrimas yra dinamiškas, ciklinškas procesas. Net ir įdiegus modelį praktikoje, jis nėra galutinis – sistemingai papildomi nauji duomenys, koreguojami hiperparametrai, diegiami patobulinimai, leidžiantys palaikyti veikimo tikslumą kintančiomis sąlygomis. Toks nuolatinis prisitaikymas užtikrina, kad modelis išliktų efektyvus ir ilgainiui tobulėtų.

Dirbtinio intelekto (DI) mokymo kokybė glaudžiai siejasi su tuo, kaip paruošiami ir tvarkomi duomenys. Net pažangiausias modelis nebus veiksmingas, jei duomenys bus nepakankami, netikslūs ar tiesiog prastai struktūruoti. Prižiūrimo, neprižiūrimo ar sustiprinamojo mokymo metodai leidžia DI spręsti labai skirtingas užduotis, tačiau jų efektyvumui būtinas ne tik geras startas, bet ir nuolatinis tobulinimas. Sistema turi mokytis ne vien iš praeities, bet ir iš kintančios aplinkos.

Šiandien DI vis dažniau pristatomas kaip viena svarbiausių skaitmeninės pažangos jėgų – ir ne be pagrindo. Jo įtaka plečiasi: nuo automatizuotų logistikos sprendimų iki personalizuotų sveikatos technologijų. Bet kartu su galimybėmis auga ir klausimų skaičius. Neužtenka vien kalbėti apie duomenų saugumą – svarbu suprasti, kaip sprendimai priimami, kas už juos atsako ir kaip jie keičia mūsų kasdienybę. Galiausiai, technologija tik tuomet tampa pažangi, kai ją lydi atsakingas ir vertybinis požiūris.

Nūdieną vis garsiau keliami pagrindiniai klausimai, be kurių atsakingas dirbtinio intelekto taikymas tampa neįmanomas. Pavyzdžiui :

- Kaip užtikrinti, kad DI sprendimai nebūtų šališki ir išliktų objektyvūs?
- Kas prisiima atsakomybę, kai sistema suklysta ir sukelia realią žalą?
- Kaip apsaugoti žmogaus privatumą, kai vis daugiau sprendimų perleidžiama algoritmams?

Tai ne abstraktūs teoriniai svarstymai. Tokie klausimai turi labai aiškią praktinę reikšmę – nuo jų priklausys, kaip bus reguliuojamos DI technologijos, kokiomis sąlygomis jos bus diegiamos ir kokias ribas joms nustatysime.

Todėl siekiant kurti tvarius ir saugius DI sprendimus, būtina neapsiriboti vien techniniu išmanymu. Lygiai taip pat svarbu atpažinti susijusias rizikas – tiek etines, tiek teisines. Tik taip galime būti tikri, kad technologijų raida iš tiesų bus naudinga visuomenei, o ne taps nekontroliuojamu eksperimentu.

2. RIZIKOS VALDYMO TEORINIAI PAGRINDAI

Rizikos valdymas – tai vienas iš esminių kiekvienos organizacijos veiklos ramsčių. Jo paskirtis – ne tik apsaugoti nuo galimų neigiamų padarinių ir užtikrinti veiklos tęstinumą, bet ir padėti išnaudoti netikėtai atsirandančias galimybes. Ši sritis išaugo kaip atsakas į nuolat kintančią verslo, finansų ir technologijų aplinką, kurioje vis svarbiau tampa gebėjimas greitai atpažinti rizikas ir į jas tinkamai sureaguoti. Rizikos valdymo pagrindas – metodai ir principai, kurie leidžia įžvelgti galimas grėsmes, įvertinti jų poveikį ir laiku imtis reikiamų veiksmų. Kai šie principai taikomi sąmoningai ir nuosekliai, organizacija tampa daug atsparesnė netikėtumams ir gali išlaikyti stabilumą net esant spaudimui.

Šio skyriaus tikslas – aiškiai ir nuosekliai supažindinti su pagrindinėmis rizikos valdymo sąvokomis bei svarbiausiais jos raidos etapais. Čia taip pat bus aptarti esminiai rizikos valdymo komponentai ir tai, kaip jie taikomi realiose situacijose. Gilindamiesi į istorinius šios srities pagrindus, peržvelgsime teorinius modelius, kurie padėjo suformuoti šiandien taikomas strategijas. Galiausiai išskirsime pagrindines rizikų rūšis ir pažvelgsime į analizės metodus, kurie padeda organizacijoms efektyviai valdyti rizikas tiek kasdienėje veikloje, tiek vykdant projektus.

2.1 Rizikos sampratos raida ir rūšys

Rizikos valdymas – tai sritis, lydinti žmoniją nuo seniausių laikų. Net senovės civilizacijos, stokodamos šiuolaikinių technologijų, ieškojo būdų, kaip sušvelninti neapibrėžtumo padarinius – pasitelkdavo ankstyvas draudimo formas, dalydavosi rizika prekyboje ar statybų procese. Esminis lūžis įvyko XVII amžiuje, kai Blaise'as Pascalis ir Pierre'as de Fermatas pradėjo kurti tikimybių teoriją. Būtent šie darbai tapo kertiniu tašku, nuo kurio rizikos vertinimas įgijo matematinį pagrindą – tai leido vystytis šiuolaikiniams rizikos valdymo modeliams, kuriuos šiandien taikome versle, finansuose ir technologijose. Vėliau, XIX–XX amžiuose, rizikos valdymas išaugo į savarankišką discipliną, kurioje vis daugiau dėmesio skiriama draudimui, finansiniam saugumui ir veiklos tęstinumui.

Pasak F. Knight (1921), rizika reiškia situacijas, kuriose neapibrėžtumas gali būti kiekybiškai įvertintas, o neapibrėžtumas – tai aplinkybės, kuriose tikimybės nėra aiškios[31]. Toks požiūris iš pagrindų atskyrė riziką nuo neapibrėžtumo ir tapo pagrindu tolesnėms rizikos valdymo teorijoms. J. von Neumann ir O. Morgenstern (1944) savo darbuose pabrėžė rizikos reikšmę sprendimų priėmimui, pritaikydami žaidimų teoriją bei matematinę logiką[32]. Kartu su finansų rinkų plėtra iškilo poreikis standartizuoti rizikos valdymo procesus, todėl buvo sukurti formalūs modeliai, tokie kaip H.Markowitz (1952)

[31] F.Knight (1921) „Risk, Uncertainity and Profit"
[32] J. Neuman, O.Morgensten (1944) „Theory of games and economic behavoir"

portfelio teorija, kuri leido kiekybiškai vertinti investicijų riziką ir grąžą[33]. Vėliau, M. Douglas ir A.Wildavsky (1983) pabrėžė socialinius ir kultūrinius rizikos suvokimo aspektus bei jų įtaką valdymo strategijoms[34].

Per pastaruosius dešimtmečius rizikos valdymas įgavo didelę reikšmę technologijų srityje, ypač informacinių sistemų saugumo ir dirbtinio intelekto rizikos srityse. U. Beck (1992) savo knygoje „Rizikos visuomenė“ iškėlė idėją, kad modernioji visuomenė tampa vis labiau priklausoma nuo rizikos valdymo, nes technologinė pažanga ir globalizacija sukuria naujas grėsmes, kurių kontrolė tampa būtina socialinės gerovės užtikrinimui[35].

Lietuvių autorės V. Stasytytė ir L.Aleksienė (2015) teigia, jog rizikos samprata yra dinamiška ir priklauso nuo aplinkos konteksto. Šiuolaikinėje vadybos literatūroje rizika apima įvairių tipų grėsmes ir galimybes, susijusias su organizacijos veikla[36].

A. Balkevičius (2017) biudžeto rizikos valdymo vadovėlyje pabrėžė, kad rizika gali turėti tiek neigiamą, tiek teigiamą poveikį organizacijos veiklai. Jo teigimu, „rizika – tai įvykių ir veiksmų neapibrėžtumo rezultatas, galintis turėti tiek teigiamą, tiek neigiamą poveikį organizacijos tikslų įgyvendinimui“[37]. (A. Balkevičius; 2017)

2 lentelė. Rizikos ir jos valdymo apibrėžimai

Autoriai	**Metai**	**Rizikos valdymo apibrėžimas / požiūris**	**Raktažodžiai**
F.Knight	1921	„Rizika yra situacija, kurioje sprendimus priimantis asmuo turi tris pranašumus: jis žino problemos struktūrą; jis supranta visą galimų pasekmių spektrą; jis gali objektyviai įvertinti kiekvienos pasekmės tikimybę.“	Rizikos ir neapibrėžtumo atskyrimas, kiekybinis įvertinimas
J. Neumann ir O. Morgenstern	1944	„Rizika yra susijusi su sprendimų priėmimu esant neapibrėžtumui, kai tikimybės yra žinomos arba gali būti apskaičiuotos.“	Neapibrėžtumo ir tikimybių apskaičiavimas
H.Markowitz	1952	„Rizika yra investicijų grąžos svyravimų amplitudė, kurią galima kiekybiškai įvertinti ir valdyti diversifikuojant portfelį.“	Rizikos diversifikavimas

[33] H.Markowithz (1952) „Portfolio Selection“ The Journal of Finance: Volume 7, Issue 1

[34] M.Douglas, A. Wildavsky (1983) Risk and Culture An Essay on the Selection of Technological and Environmental Dangers

[35] U.Beck (1992) „Risk Society towards a new modern“

[36] V. Stasytytė, L.Aleksiene (2015) „Operational Risk Assessment and Management in Small and Medium-sized Enterprises“ DOI: 10.3846/btp.2015.568

[37] A.Balkevičius 2018 „Biudžeto rizikos Valdymas“

M.Douglas ir A.Wildavsky	1983	„Rizika yra kultūriškai ir socialiai konstruojama sąvoka, kurios suvokimas ir valdymas priklauso nuo bendruomenės vertybių ir normų.“	Socialinis ir kultūrinis kontekstas yra rizikos dalis
U.Beck	1992	„Rizika yra sisteminė moderniosios visuomenės savybė, kylanti iš technologinės pažangos ir globalizacijos, sukurianti naujas grėsmes, kurių kontrolė būtina socialinės gerovės užtikrinimui.“	Sisteminė ir technologinė rizika
A.Stasytyte ir L.Aleksienė	2015	„Rizikos samprata yra dinamiška ir priklauso nuo aplinkos konteksto; šiuolaikinėje vadybos literatūroje rizika apima įvairių tipų grėsmes ir galimybes, susijusias su organizacijos veikla.“	Dinamiška organizacinė rizika
A.Balkevičius	2017	„Rizika – tai įvykių ir veiksmų neapibrėžtumo rezultatas, galintis turėti tiek teigiamą, tiek neigiamą poveikį organizacijos tikslų įgyvendinimui.“	Neapibrėžtumo rezultatas, teigiamas ir neigiamas poveikis

Šaltinis: Parengta autorės, remiantis Knight (1921), Neumann ir Morgenstern (1944), Markowitz (1952), Douglas ir Wildavsky (1983), Beck (1992), Stasytytė ir Aleksienė (2015), Balkevičius (2017).

Rizikos rūšys ir jų klasifikavimo kriterijai

Rizikos valdymo teorijoje paprastai išskiriamos pagrindinės rizikos rūšys remiantis tarptautiniu standartu ISO 31000 (2018), kuris pateikia nuoseklią rizikos klasifikavimo sistemą. Šis standartas rizikas klasifikuoja pagal jų galimą poveikį organizacijos tikslų pasiekimui bei veiklos procesams[38]:

- **Finansinė rizika** – tai galimybė patirti finansinius nuostolius. Pasak Bromiley ir kt. (2014), ši rizika laikoma viena reikšmingiausių, nes ji tiesiogiai veikia organizacijos finansinį stabilumą ir gyvybingumą[39].
- **Operacinė rizika** – apima vidinius procesus, sistemas, žmogiškuosius išteklius bei išorinius veiksnius, galinčius sutrikdyti įprastą organizacijos veiklą. COSO (2004) pažymi, kad operacinė rizika yra sudėtinga, nes kyla iš įvairių organizacijos veiklos aspektų[40].

[38] A.Sidorenko „Risk – Academy‘s Guide on ISO 31000 https://www.researchgate.net/publication/369916561 (žiūrėta 2025-01-10)

[39] P. Bromiley , M. McShane , A. Nair , E. Rustambekov 2014 „Enterprise Risk Management: Review, Critique, and Research Directions“

[40] COSO ERM modelis 2004 „ Enterprise Risk Management – Integrated Framework“

- **Strateginė rizika** – kyla dėl netinkamų ar neefektyvių strateginių sprendimų, kurie gali lemti rinkos dalies sumažėjimą ar konkurencinio pranašumo praradimą. Tokios rizikos valdymas yra itin svarbus siekiant išlaikyti organizacijos ilgalaikį konkurencingumą.
- **Teisinė rizika** – atsiranda dėl teisinių aktų ar reguliavimo pokyčių, galinčių turėti įtakos organizacijos veiklai. Nesugebėjimas laikytis teisinių reikalavimų gali sukelti didelio mąsto pasekmių, tokių kaip sankcijos ar teisminiai procesai.
- **Technologinė rizika** – susijusi su neapibrėžtumu, kurį sukelia technologinės inovacijos, IT sistemų gedimai ar nepakankamas technologinių naujovių įsisavinimas. Dėl spartaus technologijų vystymosi šios rizikos vertinimas tampa vis svarbesnė, nes organizacijos dar labiau priklauso nuo skaitmeninių sprendimų ir IT infrastruktūros.

2.2 Rizikos valdymo procesas ir metodai

Rizikos valdymo procesas – tai struktūruota ir nuosekli veikla, kurios pagrindinis tikslas yra nustatyti, įvertinti ir kontroliuoti rizikas, galinčias paveikti organizacijos tikslų įgyvendinimą. Remiantis COSO (2004), veiksmingas rizikos valdymas turi apimti visas organizacijos veiklos sritis ir būti integruotas į strateginį planavimą, siekiant užtikrinti visapusišką rizikų kontrolę.

Pagrindiniai rizikos valdymo proceso etapai:

1. **Rizikos identifikavimas.** Tai pradinė ir itin reikšminga rizikos valdymo proceso dalis, kurioje siekiama nustatyti visus galimus įvykius, galinčius daryti įtaką organizacijos tikslams. H. Smith ir J. D. McKeen (2009) pabrėžia, kad sistemingas ir tęstinis rizikos identifikavimo procesas yra pagrindinis IT rizikos valdymo sėkmės faktorius[41]. Rizikos identifikavimo metodai apima istorinių duomenų analizę, ekspertų konsultacijas bei analitinių įrankių taikymą.
2. **Rizikos analizė ir vertinimas**. Nustačius rizikas, jos vertinamos pagal jų poveikio dydį ir tikimybę. Vertinimo procese gali būti naudojami tiek kokybiniai, tiek kiekybiniai metodai:
    - **Kokybiniai metodai** – tai ekspertų vertinimai ir scenarijų analizė, padedantys nustatyti galimas pasekmes.
    - **Kiekybiniai metodai** – įtraukia matematinį modeliavimą, Monte Karlo simuliaciją ir jautrumo analizę, leidžiančius detaliau prognozuoti galimų rizikų pasekmes (A. Balkevičius, 2017).

[41] Smith, H., & McKeen, J. D. (2009). Developments in Practice XXXIII: A Holistic Approach to Managing IT-based Risk. Communications of the Association for Information Systems, 25, pp-pp. https://doi.org/10.17705/1CAIS.02541

3. **Rizikos mažinimo strategijos ir kontrolė.** Šiame etape parenkamos efektyviausios priemonės rizikai valdyti. Galimi rizikos kontrolės būdai:
    - **Rizikos vengimas** – veiklos, galinčios sukelti riziką, atsisakymas.
    - **Rizikos perleidimas** – rizikos perdavimas kitoms šalims, pavyzdžiui, sudarant draudimo sutartis.
    - **Rizikos sumažinimas** – papildomų priemonių diegimas siekiant sumažinti rizikos poveikį arba tikimybę.
    - **Rizikos priėmimas** – sąmoningas sprendimas toleruoti tam tikrą riziką, jei jos poveikis yra priimtinas arba mažos tikimybės.
4. **Rizikos stebėsena ir vertinimas.** Kadangi rizikos valdymas vyksta nuolat, svarbu ne tik stebėti esamas rizikas ir jų pokyčius, bet ir periodiškai peržiūrėti taikomas strategijas, kad jos atitiktų realią situaciją. COSO (2004) akcentuoja, kad nuolatinė stebėsena leidžia laiku pastebėti naujas grėsmes ir užtikrinti, kad rizikos valdymo priemonės išliktų veiksmingos, prisitaikant prie kintančių sąlygų.

2.3 Rizikos identifikavimo ir vertinimo priemonės

Peteliškės metodas (*angl. Bow– Tie*) – vizualus rizikos analizės ir valdymo įrankis, kurio pagrindinis tikslas – nustatyti rizikos įvykio priežastis bei galimas pasekmes ir apibrėžti kontrolės priemones, siekiant sumažinti rizikos tikimybę bei jos poveikį. Šio metodo pavadinimas kilo iš jo vizualinės struktūros, kuri primena „peteliškės“ kaklaraištį, o jo populiarumas ypač išaugo pramonės sektoriuose, kur svarbus veiksmingas didelės rizikos situacijų valdymas.

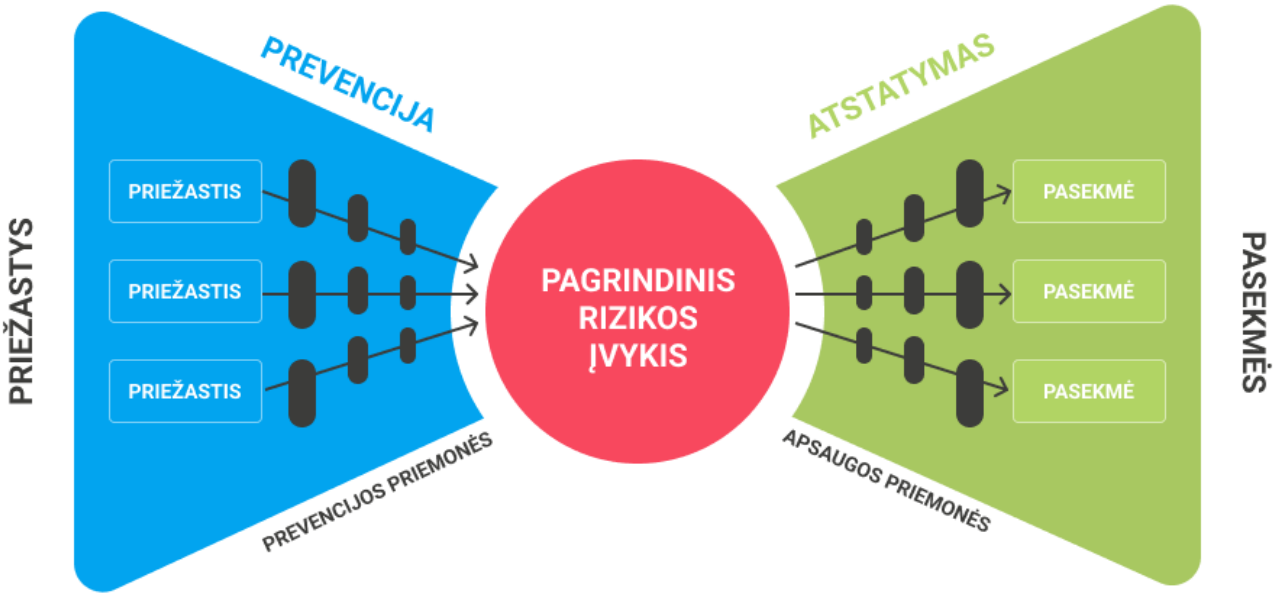
Šis metodas išpopuliarėjo 1990– aisiais metais, kai Olandijos naftos ir dujų bendrovė „Shell“ pradėjo jį taikyti siekdama efektyviai valdyti didelės rizikos situacijas energetikos srityje[42]. Metodo tobulinimas siejamas ir su XX a. antroje pusėje vykdytais moksliniais tyrimais rizikos valdymo srityje, kurių pagrindinė kryptis buvo sudėtingų rizikos scenarijų paprastas ir aiškus pateikimas bei kontrolės priemonių apibrėžimas.

„Peteliškės“ metodas plačiai naudojamas tokiose srityse kaip naftos gavyba, chemijos pramonė ir aviacija, kur yra didelė pavojingų įvykių rizika. Pasak A. de Ruijter ir F. Guldenmund (2016), šis metodas yra naudingas ne tik dėl savo paprastos struktūros, bet ir dėl galimybės viename paveiksle pavaizduoti visą rizikos valdymo grandinę – nuo priežasčių iki kontrolės priemonių ir pasekmių[43].

[42] CCPS (Center for Chemical Process Safety) 2018 Bow Ties in Risk Management ISBN: 9781119490388
[43] A. de Ruijter, F. Guldenmund (2016) „The bowtie method: A review“ 211-218psl

Būtent gebėjimas aiškiai vizualizuoti sudėtingus rizikos scenarijus padarė šį metodą populiarų sveikatos apsaugos, energetikos ir IT infrastruktūros srityse. „Peteliškės“ metodo privalumas yra jo aiškumas ir paprastumas, leidžiantis efektyviai komunikuoti rizikas su suinteresuotosiomis šalimis.

3 pav. „Peteliškės“ metodas (angl. Bow– Tie method)



Šaltinis: Sudaryta autorės pagal „CCPS (Center for Chemical Process Safety) 2018 Bow Ties in Risk Management“

Struktūra ir elementai

„Peteliškės“ metodas išskiria dvi esmines kontrolės priemonių kategorijas:

- **Prevencinės kontrolės priemonės**: taikomos siekiant sumažinti rizikos įvykio tikimybę.
- **Reagavimo kontrolės priemonės:** skirtos sumažinti galimų pasekmių poveikį, jei rizikos įvykis vis dėlto įvyktų.

Metodą sudaro šie pagrindiniai komponentai:

- **Centrinis mazgas** – tai pagrindinis rizikos įvykis, kurio priežastys bei pasekmės nagrinėjamos.
- **Priežastys –** tai visi potencialūs veiksniai, galintys paskatinti rizikos įvykį. Jie pateikiami kairėje „Peteliškės“ diagramos dalyje.
- **Pasekmės** – visos galimos rizikos įvykio pasekmės, vaizduojamos diagramos dešinėje pusėje.
- **Kontrolės priemonės –** priemonės, skirtos sumažinti rizikos įvykio tikimybę (prevencinės priemonės) arba sušvelninti jo pasekmes (reagavimo priemonės).

Bow– Tie „Peteliškės“ metodas itin naudingas situacijose, kuriose būtina aiškiai suprasti, kokie veiksniai gali sukelti riziką ir kokios priemonės galėtų padėti ją suvaldyti. Dėl gebėjimo vizualiai

pavaizduoti sudėtingas rizikos situacijas, šis metodas plačiai naudojamas sveikatos apsaugos, darbo saugos ir aplinkos apsaugos srityse.

Norint geriau suprasti „Peteliškės“ metodo pritaikymą, galima įsivaizduoti situaciją, kai organizacija nagrinėja konfidencialių duomenų nutekėjimo riziką.

Pagrindinis rizikos įvykis (centrinis mazgas) – *„duomenų nutekėjimas“* – vaizduojamas diagramos centre. Kairėje pusėje pateikiamos galimos šio įvykio priežastys, pavyzdžiui, nesaugus darbuotojų elgesys, IT infrastruktūros pažeidžiamumai ar programinės įrangos klaidos. Dešinėje diagramos pusėje pateikiamos galimos pasekmės, tokios kaip finansiniai nuostoliai, žala įmonės reputacijai ar teisiniai ginčai.

Rizikos valdymui schemoje naudojamos kontrolės priemonės yra suskirstytos į dvi dalis: kairėje pusėje, tarp centrinio mazgo ir priežasčių, pateikiamos prevencinės priemonės, tokios kaip darbuotojų mokymai ir IT saugumo auditai, o dešinėje pusėje, tarp centrinio mazgo ir pasekmių, pateikiamos pasekmių valdymo priemonės, tokios kaip duomenų šifravimas ar greito reagavimo planai.

Šis metodas padeda organizacijoms aiškiau suprasti, kurios taikomos priemonės veikia efektyviausiai, o kur dar reikėtų papildomų investicijų rizikos poveikiui sumažinti.

Tačiau pagrindinis trūkumas – šis metodas netinka sudėtingoms rizikos situacijoms, kur būtina taikyti išsamius kiekybinius vertinimo metodus

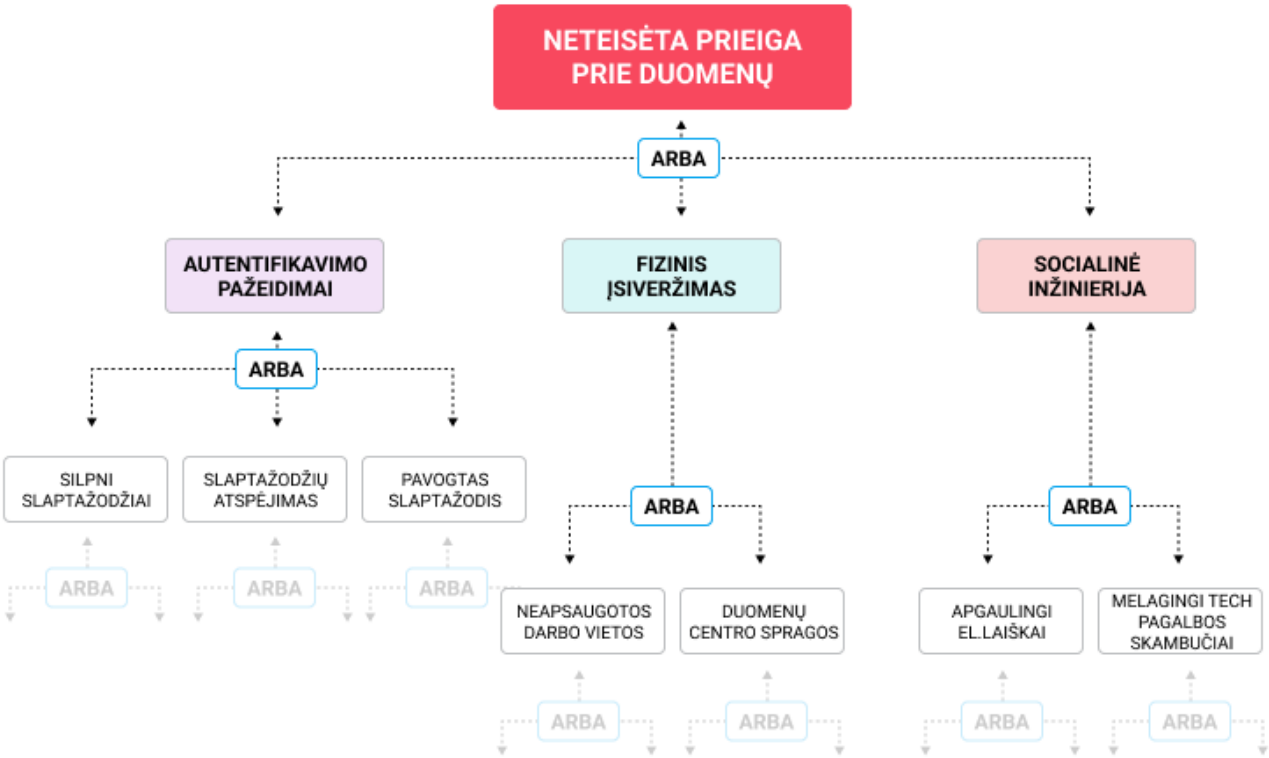
„Atakos medžio“ modelis (*angl. Attack Tree Model*) – tai vizualus rizikos analizės metodas, leidžiantis aiškiai pavaizduoti galimus grėsmių scenarijus ir padedantis parinkti tinkamiausias jų valdymo priemones. Šį metodą 1999 m. pristatė kibernetinio saugumo specialistas Bruce Schneier savo knygoje „Secrets and Lies: Digital Security in a Networked World“. Nuo to laiko modelis tapo plačiai taikomas kibernetinio saugumo, IT bei kritinės infrastruktūros apsaugos srityse.

Pagrindinė „atakos medžio“ modelio idėja – hierarchiškai pavaizduoti įvairius galimus grėsmių kelius, vedančius į konkretaus atakos tikslo pasiekimą. Šis metodas padeda suprasti, kokie veiksmai gali prisidėti prie grėsmės įgyvendinimo, ir leidžia organizacijoms efektyviau planuoti prevencines priemones.

Atakos medžio modelį sudaro šie pagrindiniai komponentai:

- **Pagrindinis mazgas** (atakos tikslas) – centrinis elementas, kuris apibrėžia, kokį konkretų grėsmės rezultatą norima išanalizuoti.
- **Šakos** – galimi veiksmai ar keliai, vedantys link pagrindinio mazgo. Kiekviena šaka atspindi konkretų scenarijų ar veiksmą, kuris prisideda prie grėsmės realizavimo.
- **Tarpiniai mazgai** – papildomi veiksmai arba sąlygos, kurie turi būti įgyvendinti, kad būtų galima pasiekti pagrindinį mazgą.
- **Galutiniai mazgai** – veiksmų ar scenarijų pabaigos taškai, kurie užbaigia tam tikrą atakos kelią.

4 pav. „Atakos medžio“ modelis



Šaltinis: Sudaryta autorės pagal Lietuvos Respublikos Vidaus reikalų ministerija (2005) „Rizikos analizės vadovas“ ISBN 5– 415– 01827– 1 120 psl

Norint geriau suprasti atakos medžio modelį, galima apsvarstyti situaciją, kai organizacija analizuoja galimus neteisėtos prieigos prie svarbių duomenų gavimo scenarijus.

Pagrindinis mazgas: „Neteisėta prieiga prie duomenų“.

Prieigos gavimo būdai:

1. Autentifikavimo pažeidimai:
   - Įsilaužimas per silpną slaptažodį.
   - Slaptažodžių atspėjimas (*angl. „bruteforce“ atakos*).
   - Pavogtos prisijungimo informacijos naudojimas
2. Fizinio įsiveržimo pažeidimai:
   - Neapsaugotos ar lengvai pasiekiamos darbo vietos;
   - Trūkumai duomenų centrų fizinėje apsaugoje.
3. Socialinė inžinierija:
   - apgaulingi el. laiškai (angl. „*phishing*“), skirti išvilioti jautrią informaciją;
   - Skambučiai, kuriuose apsimetama pagalbos specialistais.

Kiekvienai grėsmei galima priskirti tikimybę ir galimą žalą. Tai padeda aiškiau suprasti bendrą rizikos mastą ir nuspręsti, kur pirmiausia reikėtų taikyti apsaugos priemones. Šis pavyzdys padeda

iliustruoti, kaip „Atakos medžio“ modelis gali būti naudojamas identifikuoti įvairius rizikos scenarijus, įvertinti jų tikimybę ir numatyti kontrolės priemones.

Šio modelio privalumai:

- Leidžia išsamiai išanalizuoti įvairius grėsmių scenarijus ir atskleisti silpnąsias sistemos vietas.
- Padeda nustatyti veiksmingiausias kontrolės priemones, kurios sumažintų grėsmių tikimybę arba jų poveikį.
- Skatina sisteminį požiūrį į rizikos analizę, apimant visus galimus atakos kelius.

Trūkumai:

- Analizė gali tapti itin sudėtinga, kai nagrinėjama daug skirtingų grėsmių scenarijų.
- Reikalauja daug laiko ir resursų, ypač jei būtina tiksliai įvertinti kiekvieno scenarijaus tikimybę bei poveikį.
- Nėra tinkamas greitam rizikos vertinimui ar situacijoms, kuriose reikia priimti skubius sprendimus.

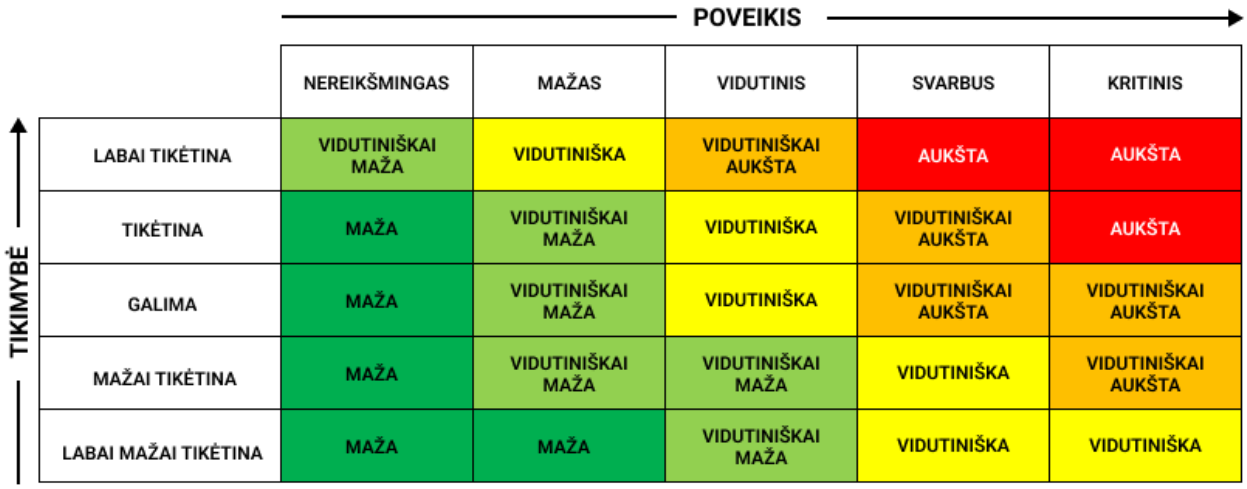
„Atakos medžio“ metodas – vienas iš praktinių įrankių, kurį organizacijos vis dažniau naudoja saugodamosi nuo sudėtingų kibernetinių grėsmių (Rizikos analizė, 2018). Jis padeda aiškiai pavaizduoti galimus atakos scenarijus, įvertinti, kiek jie tikėtini ir kokią žalą galėtų sukelti – tai ypač naudinga, kai reikia tiksliai suplanuoti apsaugos priemones. Dėl savo lankstumo ir aiškios struktūros šis metodas plačiai taikomas tiek kibernetinio saugumo, tiek kritinės infrastruktūros srityse, ir šiuo metu laikomas vienu iš standartinių analizės būdų.

Rizikos vertinimo matrica – vienas iš pirmųjų struktūruotų rizikos analizės įrankių, išpopuliarėjęs XX a. pabaigoje. Ji buvo sukurta kaip greitas ir praktiškas būdas nustatyti, kurios rizikos reikalauja didžiausio dėmesio skirtingose veiklos srityse. Iš pradžių ši matrica buvo taikoma inžinerijos bei projektų valdymo srityse, o vėliau išplito ir į IT, sveikatos apsaugos bei finansų sektorius. Pasak D. Hillson ir R. Murray– Webster (2007), rizikos matrica suteikia aiškų vizualų būdą vertinti rizikas, jas suskirstant pagal dvi dimensijas – tikimybę ir poveikį[44]. Tai leidžia organizacijoms efektyviai paskirstyti išteklius, koncentruojantis į didžiausią grėsmę keliančius veiksnius.

Dažniausiai matrica naudojama atliekant kokybinį rizikos vertinimą. Tačiau, kaip pažymi S. Kaplan ir B. J. Garrick (1981), metodas gali būti pritaikytas ir kiekybiniams vertinimams, siekiant gauti tikslesnius rezultatus sudėtingose situacijose[45]. Rizikos matrica padeda vizualiai pavaizduoti rizikų reikšmingumą, įvertinant jų tikimybę bei poveikio mastą. Tai suteikia galimybę greitai nustatyti, kurios rizikos yra svarbiausios ir reikalauja skubių veiksmų. Šio metodo paprastumas ir veiksmingumas nulėmė jo plačią taikymo sritį tiek mažose, tiek didelėse organizacijose.

[44] D. Hillson, R Murray-Webster (2007) „Understanding and Managing Risk Attitude“
[45] S. Kaplan, B. John Garrik 1981 “On the quantitative definition of risk” Risk Analysis journal 11-27psl

5 pav. Rizikos vertinimo matrica



Šaltinis: Sudaryta autorės pagal Lietuvos Respublikos Vidaus reikalų ministerija (2005) „Rizikos analizės vadovas“ ISBN 5– 415– 01827– 1 120 psl

Rizikos vertinimo matrica paprastai pateikiama kaip dvimatė lentelė, kurioje viena ašis nurodo rizikos įvykio tikimybę, o kita – galimo poveikio dydį. Kiekviename lentelės langelyje pateikiami rizikos reikšmės lygiai, kuriuos galima vizualiai atskirti naudojant skirtingas spalvas:

- **Žalioji zona**: maža ar silpną įtaka turinti rizika, kuriai nereikia skubių veiksmų.
- **Geltonoji zona**: vidutinė rizika, kuriai gali būti taikoma stebėsena arba dalinės kontrolės priemonės.
- **Raudonojoje zonoje**: aukšta ir kritiška rizika, kuriai būtina skirti prioritetinį dėmesį ir nedelsiant įgyvendinti veiksmingas kontrolės priemones.

Rizikos reikšmė apskaičiuojama sudauginus tikimybės ir poveikio koeficientus. Pavyzdžiui, jei rizikos tikimybė įvertinama 4 balais (dažna), o poveikis – 5 balais (kritinis), bendra rizikos reikšmė būtų 20 balų, kas paprastai patektų į raudoną zoną, reikalaujančią skubių valdymo priemonių.

Įsivaizduokime organizaciją, kuri vertina IT sistemos gedimo riziką.

Pirma, nustatomos galimos rizikos įvykio priežastys:

- Įrangos gedimai.
- Programinės įrangos klaidos.
- Netinkamas sistemos administravimas.
- Netyčinis įrangos perkrovimas

Toliau atliekamas šių rizikos veiksnių vertinimas pagal tikimybę ir poveikį:

3 lentelė. Rizikos vertinimo tikimybių matrica

Rizikos veiksnys	Tikimybė	Poveikis	Rizikos reikšmė	Rizikos lygis
Įrangos gedimai	3	4	12	**Vidutinė**
Programinės įrangos klaidos	**4**	**5**	**20**	**Aukšta**
Netinkamas sistemos administravimas	2	3	6	**Vid. Maža**
Netyčinis įrangos perkrovimas	2	1	3	**Maža**

Šaltinis: Parengta autorės, remiantis D. Hilson ir R. Murray– Webster (2007) rizikos matricos modeliu

Pagal šią analizę matyti, kad didžiausią riziką sudaro programinės įrangos klaidos, todėl būtent jų prevencija turėtų būti prioritetinė. Prevencinės priemonės galėtų apimti reguliarius programinės įrangos atnaujinimus, kruopštų testavimą prieš įdiegiant naujas funkcijas bei nuolatinį atsarginių kopijų kūrimą, siekiant sumažinti galimų klaidų poveikį.

Privalumai – metodas išsiskiria tuo, kad yra paprastas naudoti ir lengvai suprantamas tiek techniniam, tiek netechniniam personalui. Jis leidžia greitai atpažinti svarbiausias rizikas ir aiškiai nustatyti prioritetines valdymo sritis. Metodas taikomas įvairiose srityse – nuo IT ir sveikatos apsaugos iki projektų valdymo bei kitų veiklos krypčių.

Trūkumai – rizikos vertinimas gali būti pernelyg subjektyvus, ypač tuomet, kai organizacija neturi pakankamai patikimų duomenų. Šis metodas netinka sudėtingoms situacijoms, kuriose reikalingi išsamūs kiekybiniai vertinimai. Rizikos reikšmė dažnai supaprastinama, nes matrica neapima galimų rizikos įvykių tarpusavio sąveikų.

Nors rizikos vertinimo matrica turi tam tikrų ribotumų, ji vis dar išlieka vienu dažniausiai taikomų praktinių įrankių įvairiose organizacijose. Tai greitas būdas nustatyti svarbiausias rizikas ir priimti sprendimus, kaip kuo veiksmingiau sumažinti jų poveikį.

Apskritai rizikų valdymas grindžiamas gana aiškiais ir universalais principais – jie padeda kryptingai identifikuoti grėsmes ir taikyti tinkamas kontrolės priemones. Šiame skyriuje aptarėme, kaip įvairūs metodai – nepaisant skirtingų požiūrių ar paskirties – gali būti pritaikomi pačiuose įvairiausiuose kontekstuose. Kitas žingsnis – pažvelgti, kaip visa tai veikia DI srityje, kur rizikos dažnai nėra taip lengvai apibrėžiamos ir reikalauja kitokio, labiau kontekstinio požiūrio.

3. DIRBTINIO INTELEKTO GRĖSMĖS, REGULIAVIMAS IR RIZIKŲ VALDYMO STANDARTAI

Sparčiai plintant dirbtinio intelekto (DI) technologijoms, jų galimybės ir pritaikymo sritys iš esmės keičia verslo procesus, viešąsias paslaugas bei kasdienį gyvenimą. Tačiau ši technologinė pažanga neišvengiamai atneša ir naujų rizikų. DI sistemų plėtra atveria daugybę inovatyvių sprendimų, bet tuo pačiu iškelia sudėtingus klausimus, susijusius su saugumu, etika ir atsakomybe. Visa tai verčia tiek verslo subjektus, tiek valstybes bei tarptautines organizacijas ieškoti efektyvių reguliavimo ir valdymo būdų.

Kaip pabrėžia M. Wooldridge (2009), norint, kad dirbtinio intelekto (DI) plėtra būtų atsakinga, reikia plataus, visapusio požiūrio – vien techninių žinių neužtenka, būtinas ir etikos bei teisės principų integravimas. A. Paulauskaitė–Tarasevičienė ir K. Šutienė (2020) taip pat išskiria, kad DI sėkmė priklauso ne vien nuo technologinių sprendimų – labai svarbu ir tai, kiek žmonės pasitiki šių sistemų etiškumu bei patikimumu. Tuo tarpu E. Brynjolfsson ir A. McAfee (2017) atkreipia dėmesį, jog praktinis DI taikymas kelia vis daugiau naujų klausimų – nuo duomenų apsaugos iki dezinformacijos grėsmių ar pačių sistemų nenuspėjamo elgesio.

Kartu su šiomis galimybėmis neišvengiamai atsiranda ir rizikos. Kai DI pasitelkiamas neleistinais tikslais, pavyzdžiui, melagingos informacijos sklaidai ar pasinaudojant technologiniais pažeidžiamumais padariniai gali būti rimti ir toli siekiantys. Tai rodo, kad vien technologinių inovacijų nepakanka, būtina ir atsakinga priežiūra bei tinkami kontrolės mechanizmai.

Šiame skyriuje bus gilinamasi į tai, kaip valdyti DI keliamas grėsmes pasitelkiant teisės aktus, tarptautinius standartus ir praktines strategijas. Dėmesys bus skiriamas svarbiausioms rizikų kategorijoms – nuo kibernetinio saugumo iki netinkamo DI pritaikymo. Pirmiausia bus aptartos pagrindinės grėsmės – nuo kibernetinių atakų iki DI naudojimo ne pagal paskirtį. Toliau analizuojama, kaip į šias grėsmes reaguoja teisės sistemos, kokie sprendimai priimami tiek nacionaliniu, tiek tarptautiniu lygmeniu. Taip pat išryškinamas standartų vaidmuo – jie padeda organizacijoms struktūruotai spręsti DI keliamus iššūkius.

Šio skyriaus tikslas – išsamiai išanalizuoti DI rizikų valdymo priemones ir parodyti, kaip grėsmių valdymas, teisinis reguliavimas ir standartizacija tarpusavyje sąveikauja siekiant sukurti etišką, patikimą ir visuomenei naudingą DI ekosistemą. DI plėtra ir rizikų valdymas yra priklausomi vienas nuo kito procesai, kurių bendras tikslas – užtikrinti saugią ir atsakingą technologijų raidą.

3.1. Išorinės ir vidinės dirbtinio intelekto grėsmės

Dirbtinio intelekto (DI) plėtra atveria daug naujų galimybių, tačiau kartu neišvengiamai iškelia ir įvairių grėsmių, kurios dažniausiai skirstomos į vidines ir išorines. **Vidinės grėsmės** grėsmės dažnai kyla iš pačių technologijų – kai algoritmai reaguoja į klaidinančius ar neišsamius duomenis, sprendimai gali būti ne tik netikslūs, bet ir šališki (Binns, 2018). Tai ne vien skaičiavimo klaida – tokie sprendimai daro poveikį realiems žmonėms ir situacijoms.

Tuo tarpu išorinės rizikos – tai jau tyčinis DI išnaudojimas, pavyzdžiui, dezinformacijos skleidimui ar kibernetinėms atakoms. Tokiais atvejais pasekmės neapsiriboja technologiniu lygiu – jos apima ir visuomenės pasitikėjimą, demokratinius procesus ar net nacionalinį saugumą (Brundage et al., 2018).

Todėl kalbėdami apie DI saugumą, turime mąstyti plačiau – neužtenka vien techninių sprendimų. Reikalingas iš esmės jungtinis požiūris, kuriame technologiniai sprendimai būtų papildyti aiškiomis etikos nuostatomis ir tvirtu teisiniu pagrindu (Floridi et al., 2018).

Tik visapusiška grėsmių analizė sudaro sąlygas kurti atsakingą, etiškai pagrįstą ir visuomenės interesus atitinkančią DI ekosistemą.

3.1.1. Vidinės dirbtinio intelekto grėsmės

Vidinės grėsmės skirstomos į dvi pagrindines kategorijas: ***DI išorinio poveikio rizikos***, kurios kyla dėl piktavališkų veiksmų, ir ***DI vidines pažeidžiamybes***, susijusias su pačių DI sistemų struktūra bei veikimo principais ir jų klaidomis. Abi šios kategorijos reikalauja specifinių rizikos valdymo metodų, siekiant užtikrinti DI sistemų saugumą ir patikimumą[46] (Miller et al., 2020). Šiame skyriuje nagrinėjamos pagrindinės grėsmės, susijusios su šiomis kategorijomis.

DI išorinio poveikio rizikos

Adversarinės atakos *(angl. Adversarial Attacks)* yra viena iš dažniausiai pasitaikančių atakų prieš DI modelius, tai tyčinės manipuliacijos įvesties duomenimis, kurios priverčia DI modelį priimti klaidingus sprendimus. Užpuolikai įterpia mažus, žmogaus akiai beveik nepastebimus trikdžius, tačiau jų pakanka, kad modelis būtų suklaidintas.

Šios atakos skirstomos į du tipus:

1. **Taikytos atakos** *(angl. Targeted Attacks)* – kai užpuolikas siekia, kad modelis pateiktų konkretų neteisingą atsakymą.
2. **Nenutaikytos atakos** *(angl. Untargeted Attacks)* – kai modelio išvestis tampa klaidinga, tačiau nėra nukreipta į konkrečią reikšmę.

[46] T.Miller, P.Howe & L.Sonenberg, L. (2020) Explainable AI: Understanding, trust, and control. Artificial Intelligence, 290, 103385.

Pavyzdžiui, autonominių transporto priemonių sistemose net ir nežymus vaizdo įvesties pokytis gali lemti neteisingą kelio ženklų atpažinimą. Tai gali sukelti pavojingų situacijų, jei, pavyzdžiui, STOP ženklas būtų klaidingai interpretuotas kaip greičio apribojimas[47] (Kurakin et al., 2017).

Modelio inversijos atakos *(angl. „Model Inversion Attacks")* yra skirtos iš modelio išgauti informaciją apie mokymo duomenis. Tokios atakos leidžia užpuolikams rekonstruoti konfidencialius duomenis, kuriais modelis buvo mokomas[48]. Pavyzdžiui, biometrinių duomenų apdorojimo sistemose užpuolikai gali atkurti vartotojų veido ar pirštų atspaudų duomenis, keldami rimtą grėsmę privatumui.

Užnuodijimo atakos *(angl. „Poisoning Attacks")* yra viena pavojingiausių grėsmių DI sistemoms, nes jos trikdo modelio sprendimų priėmimo procesą [49](Biggio & Roli, 2018). Šių atakų metu kenksmingi duomenys tyčia įterpiami į mokymosi aibę, kad modelis pradėtų sistemingai klysti.

Duomenų išgavimo atakos *(angl. „Data Extraction Attacks")* leidžia užpuolikams gauti jautrią informaciją apie mokymo duomenis. Net jei tiesioginė prieiga prie duomenų nėra suteikta, atakos metu modelio išvestys gali būti analizuojamos naudojant specifinius užklausų rinkinius[50]. Tokios atakos ypač pavojingos medicinos ar finansų sektoriuose, kur jautri informacija, pavyzdžiui, pacientų sveikatos duomenys, gali būti neteisėtai pasisavinta.

Narystės išvedimo atakos *(angl. „Membership Inference Attacks")* leidžia užpuolikui nustatyti, ar konkretus duomuo buvo naudotas DI modelio mokyme [51]. Tai ypač pavojinga, kai kalbama apie jautrią informaciją – tarkime, ar žmogus dalyvavo tam tikrame medicininiame tyrime. Tokio tipo pažeidimai itin aktualūs sveikatos srityje, kur net ir menkiausias privatumo pažeidimas gali sukelti rimtų pasekmių tiek asmeniui, tiek duomenis valdančiai organizacijai.

DI vidaus pažeidžiamumai

Dirbtinio intelekto (DI) sistemų patikimumą lemia ne tik išorinės grėsmės – dažnai nemažiau svarbūs ir vidiniai pažeidžiamumai. Jie gali atsirasti dėl ribotos modelių architektūros, per siauro ar nepakankamai kokybiško duomenų rinkinio ar net technologinių spragų. Tokie veiksniai neretai nulemia netikslumus, skatina diskriminaciją ar sukuria sprendimus, kuriuos sudėtinga paaiškinti. Visa tai kelia abejonių dėl DI patikimumo, ypač kai jis taikomas jautriose srityse, tokiose kaip medicina, teisinės paslaugos ar finansai.

Viena iš dažniausių ir sunkiausiai išsprendžiamų problemų – **algoritmų šališkumas** (angl. „algorithmic bias"). Jis dažnai kyla tada, kai mokymo duomenys neatspindi realaus pasaulio įvairovės arba kai modelyje užkoduojamos sisteminės nelygybės. Jei DI mokomas duomenimis, kuriuose jau glūdi diskriminacinės nuostatos, jis ne tik jas perima, bet ir gali dar labiau įtvirtinti. Geras to pavyzdys – veidų

[47] A. Kurakin (2017). Adversarial examples in the physical world.
[48] Fredrikson, M., et al. (2015). Model inversion attacks that exploit confidence information and basic countermeasures.
[49] B.Biggio & F.Roli. (2018). Wild patterns: Ten years after the rise of adversarial machine learning.
[50] N.Carlini, et al. (2021). Extracting Training Data from Large Language Models.
[51] R. Shokri, et al. (2107) Membership inference attacts agains machine learning models

atpažinimo sistemos, kurios, kaip rodo tyrimai, dažnai prasčiau atpažįsta tamsesnės odos žmones. Taip nutinka todėl, kad jos buvo mokomos duomenimis, kuriuose dominuoja šviesiaodžių atvaizdai (Buolamwini & Gebru, 2018). Panašiai finansų sektoriuje DI modeliai gali neteisingai įvertinti tam tikrų demografinių grupių kreditingumą, jei istoriniuose duomenyse egzistavo šališkumo atvejų.

Modelių nepaaiškinamumas *(angl. „Black Box Problem")* – tai dar vienas iš rimčiausių iššūkių, su kuriais susiduriama dirbtinio intelekto (DI) srityje, yra vadinamoji „juodosios dėžės" problema. Tai reiškinys, kai giluminio mokymosi modeliai – ypač neuroniniai tinklai – pateikia rezultatus, bet paaiškinti, kaip jie prie jų priėjo, nesugeba. Modeliai veikia remdamiesi tūkstančiais vidinių parametrų ir algoritminių sąsajų, kurie žmogui nesuprantami. Šis neaiškumas tampa ypač pavojingas ten, kur nuo sprendimo priklauso žmogaus gerovė ar net gyvybė – tarkim, medicinoje ar teisėje. Įsivaizduokite situaciją, kai modelis nustato didelę vėžio tikimybę – gydytojas privalo suprasti, kuo remiantis buvo pateikta tokia išvada[52]. Be aiškaus paaiškinimo, sunku pagrįsti tolesnius sprendimus, jau nekalbant apie atsakomybę ar paciento pasitikėjimą (Carvalho et al., 2019). Jei šis procesas nėra skaidrus, pasitikėjimas tokiais modeliais mažėja, o sprendimų priėmimo kokybė tampa abejotina.

Adaptyvumo problemos ir per didelis pritaikymas *(angl.. Overfitting)* DI modeliai gali tapti pernelyg specializuoti konkretiems duomenų rinkiniams ir prarasti gebėjimą prisitaikyti prie naujų situacijų. Ši problema vadinama per dideliu pritaikymu[53] (Goodfellow et al., 2016). Kai modelis pernelyg tiksliai išmoksta mokymo duomenis, jis gali nesugebėti apdoroti naujų, anksčiau nematytų atvejų buvo mokomi tik esant geroms oro sąlygoms, gali pasirodyti nepatikimai esant rūko ar lietaus sąlygoms. Siekiant išvengti tokių situacijų, taikomi įvairūs modelio tobulinimo metodai, tokie kaip duomenų augmentacija (*data augmentation*), kryžminis patikrinimas (*cross-validation*) ar kitos reguliavimo technikos.[54].

Vidinės dirbtinio intelekto grėsmės rodo, kad DI sistemų patikimumą lemia ne vien išorinės apsaugos priemonės, bet ir pačios kūrimo metodikos kokybė. Todėl norint sumažinti šiuos pažeidžiamumus, svarbu taikyti veiksmingas reguliavimo strategijas, užtikrinti duomenų įvairovę bei kurti paaiškinamus modelius, kurie leistų aiškiau suprasti, kaip priimami sprendimai, ir padėtų geriau juos kontroliuoti.

[52] M.T. Ribeiro, S. Singh, C. Guestrin (2016) – Machine Learning „Why should i trust you?": Explaining the predictions of any classifier" https://doi.org/10.48550/arXiv.1602.04938

[53] I. Goodfellow, Y. Bengio & A. Courville (2016). Deep Learning. MIT Press.

[54] N. Srivastava, G.Hinton, A.Krizhevsky, I.Sutskever & R.Salakhutdinov (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929-1958.

3.1.2. Išorinės dirbtinio intelekto grėsmės

Dirbtinio intelekto technologijos vystosi itin greitai, atverdamos naujas galimybes inovacijoms, tačiau kartu tampančios ir pavojingu įrankiu tiems, kurie siekia manipuliuoti informacija, skleisti dezinformaciją ar pasinaudoti DI automatizavimo privalumais neskaidriems tikslams. Skirtingai nuo vidinių DI grėsmių, kurios kyla dėl technologinių trūkumų ar sąmoningo sistemų pažeidžiamumo išnaudojimo, išorinės grėsmės yra susijusios su tikslingu DI pritaikymu manipuliacijoms ir apgaulei.

Dirbtinio intelekto kuriama dezinformacija tampa vis sudėtingesnė ir, paradoksaliai, vis labiau įtikinanti. Tyrimai rodo, kad tokios žinutės ne tik skamba įtikinamai, bet dažnai beveik nesiskiria nuo tikros informacijos - ypač kai DI modeliai taikosi į auditoriją, atsižvelgdami į žmonių įpročius, nuotaikas ar net emocinius modelius (Zellers et al., 2019).[55]

Dar didesnį nerimą kelia tai, kad DI pagrįstos deepfake technologijos leidžia kurti labai tikroviškus, bet iš esmės suklastotus vaizdo ar garso įrašus. Tokie įrašai gali būti naudojami ne tik propagandai skleisti, bet ir nusikalstamiems tikslams[56] (Chesney & Citron, 2019). Ir tai dar ne viskas - DI neapsiriboja vien tik melagingo turinio kūrimu. Jis taip pat veikia kaip sklaidos variklis, padėdamas parinkti tikslinę auditoriją ir išplatinti turinį ten, kur jo poveikis bus stipriausias. Toks manipuliavimas viešąja nuomone gali ne tik skatinti susipriešinimą, bet ir daryti realią žalą demokratijai[57] (Ferrara, 2020; Brundage et al., 2018).

Tokio tipo grėsmės iškelia rimtų iššūkių ne vien informaciniam saugumui – jos smogia ir tiems, kurie kuria ar naudoja DI technologijas. Net jei kūrėjai ar tiekėjai neturi jokių piktavališkų ketinimų, šios sistemos vis tiek gali būti pasitelktos netinkamiems tikslams – pavyzdžiui, manipuliacijai. Tokie atvejai gali rimtai pakenkti įmonių reputacijai, pakirsti pasitikėjimą pačia technologija ir dar labiau paskatinti reguliavimo griežtinimą. Todėl rizikos, susijusios su DI generuojamu turiniu, vis labiau aktualėja tiek privačiame, tiek viešajame sektoriuje. Abi šios sritys stengiasi ne tik apsaugoti informacinę erdvę nuo piktnaudžiavimo, bet ir užkirsti kelią neigiamoms pasekmėms, kurios gali kilti dėl neatsakingo ar sąmoningai žalingo technologijų panaudojimo.

Socialinių tinklų botai ir algoritminė manipuliacija. Socialiniai tinklai, tokie kaip Facebook, Twitter ar TikTok, tapo vienais pagrindinių kanalų, per kuriuos plinta DI sugeneruotas turinys. Automatizuoti botai, valdomi dirbtinio intelekto sistemų, geba imituoti žmonių elgesį, skleisti tam tikras politines ar komercines žinutes bei stiprinti jų poveikį didelėms auditorijoms[58] (Ferrara, 2020). Tyrimai rodo, kad net apie 15 % socialinių tinklų paskyrų gali būti automatizuoti botai, kurių tikslas –

[55] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending Against Neural Fake News. Advances in Neural Information Processing Systems (NeurIPS).

[56] Chesney, R., & Citron, D. K. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. Foreign Affairs, 98(1), 147-155.

[57] Ferrara, E. (2020). Disinformation and social bot operations in the run up to the 2020 US election. Harvard Kennedy School Misinformation Review.

[58] Ferrara, E. (2020). Social Bots and Social Media Manipulation in 2020: The Year in Review. arXiv preprint arXiv:2102.08436.

programiškai skleisti tam tikrus naratyvus ir daryti įtaką informacijos sklaidai[59] (Shao et al., 2018). JAV rinkimų kontekste buvo užfiksuota, kad netikrų paskyrų tinklai aktyviai dalijosi melagingomis naujienomis, formuodami iškreiptą viešąją nuomonę ir manipuliuodami informacija politiniais tikslais. Tokia koordinuota dezinformacijos sklaida kelia rimtą grėsmę demokratiniams procesams, nes rinkėjų sprendimai gali būti grindžiami klaidinančia arba net visiškai suklastota informacija.

Mikrotaikymas ir personalizuota dezinformacija. DI suteikia galimybę ne tik masiškai skleisti informaciją, bet ir pritaikyti ją kiekvienam vartotojui individualiai. Mikrotaikymas (*angl. Microtargeting and Personalized Disinformation*) – tai strategija, kai DI analizuoja vartotojų elgesį, pomėgius ir demografinius duomenis, kad pateiktų personalizuotą turinį, kuris juos labiausiai veikia [60](Bessi & Ferrara, 2016).

Ši strategija dažnai naudojama rinkodaroje, tačiau tampa pavojinga, kai pasitelkiama manipuliuoti politiniais sprendimais ar visuomenės nuomone. 2018 m. iškilęs Cambridge Analytica skandalas [61] tapo ryškiu pavyzdžiu, kaip dirbtinis intelektas gali būti išnaudotas politiniais tikslais – milžiniški kiekiai vartotojų duomenų buvo renkami be aiškaus jų sutikimo, o vėliau naudoti rinkėjų elgsenai formuoti (Cadwalladr & Graham–Harrison, 2018). Šis įvykis akivaizdžiai parodė, kad DI technologijos gali kelti grėsmę demokratinėms vertybėms ir rinkimų skaidrumui.

Kibernetinės atakos, pasitelkiant DI. Jei anksčiau tokioms atakoms reikėdavo nemažų resursų, šiandien DI leidžia automatizuoti visą procesą: nuo sistemų silpnųjų vietų paieškos iki itin tikslių taikinių nustatymo. Tokios atakos tampa ne tik greitesnės ir labiau pritaikytos prie situacijos, bet ir žymiai sunkiau aptinkamos.

- **Išmanusis sukčiavimas (angl. AI– powered phishing):** anksčiau atpažinti apgaulingus el. laiškus buvo palyginti lengva, juos išduodavo prasta gramatika, keistos formuluotės ar akivaizdžiai įtartini siuntėjo adresai. Tačiau šiandien situacija gerokai pasikeitė. Naudodami natūralios kalbos apdorojimo (NLP) algoritmus, DI modeliai gali kurti itin įtikinamus laiškus - tiek kalbos tonas, tiek stilius dažnai atitinka gavėjo bendravimo manierą. Dar daugiau - šie modeliai geba reaguoti į atsakymus, o tai dar labiau sustiprina apgaulės efektą ir apsunkina jos atpažinimą [62] (Brown et al., 2020).
- **Balsu pagrįstos apgavystės** *(angl. voice phishing, vishing):* naujausi DI modeliai geba per kelias sekundes išanalizuoti žmogaus balsą ir sukurti itin realistišką jo kopiją. Nusikaltėliai gali

[59] Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. Nature Communications, 9, 4787.

[60] Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. First Monday, 21(11).

[61] Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian.

[62]Brown University Office of Information Technology. (2023). Beware AI-Enhanced Phishing Attempts. Skaityta internetu 2025-03-02 https://it.brown.edu/phish-bowl-alerts/beware-ai-enhanced-phishing-attempts it.brown.edu

pasinaudoti viešai prieinamais įrašais, suklastoti balsą ir įtikinti žmones atlikti veiksmus. Vienas iš žinomiausių atvejų įvyko 2020 m., kai Jungtinės Karalystės energetikos įmonės finansų vadovas gavo skambutį, kuris, kaip vėliau paaiškėjo, buvo DI sugeneruotas jo viršininko balsas. Sukčius įtikino pervesti 200 000 eurų į nusikaltėlių sąskaitą [63](Brundage et al., 2018).

- **Slaptažodžių nulaužimas** (*angl. Password Cracking*)**:** jei anksčiau slaptažodžių atakos buvo vykdomos naudojant „brute– force“ metodus (kai bandomi visi įvairūs slaptažodžių deriniai), dabar DI leidžia prognozuoti slaptažodžius, analizuojant vartotojų įpročius ir viešai prieinamą informaciją [64](Hitaj et al., 2019). Tokie metodai labai sutrumpina įsilaužimo laiką ir leidžia vykdyti itin tikslias, individualizuotas atakas.
- **Kenkėjiškos programinės įrangos** *(angl. AI– generated Malware)* **evoliucija:** anksčiau antivirusinės programos galėjo atpažinti virusus pagal jų unikalius kodus, tačiau DI leidžia sukurti kenkėjiškas programas, kurios nuolat keičia savo struktūrą, taip apeidamos tradicines saugumo priemones[65] (Anderson et al., 2018). Tai reiškia, kad įprasti saugumo sprendimai tampa mažiau veiksmingi prieš DI generuojamas grėsmes.
- **Tikslinės atakos prieš organizacijas ir valstybines institucijas** *(angl. Targeted Attacks on Organizations and Government Institutions)*: nusikaltėliai gali naudoti DI modelius, kad analizuotų įmonių IT infrastruktūrą, ieškotų pažeidžiamumų ir automatizuotai juos išnaudotų. 2020 m. buvo užfiksuota DI pagrįsta kibernetinė ataka, kurios metu algoritmai sugebėjo apeiti daugiapakopę autentifikaciją ir įsilaužti į finansinių institucijų sistemas [66](Miller et al., 2020).

Išorinės grėsmės, susijusios su dirbtiniu intelektu, aiškiai parodo, kad ši technologija gali būti panaudota ne tik inovacijoms, bet ir manipuliacijoms ar net kibernetinėms atakoms. Vien techninių sprendimų čia nepakanka, būtinas ir tvirtas teisinis reguliavimas, ir nuoseklus visuomenės švietimas. Tik suderinus šiuos aspektus būtų galima pabandyti, užtikrinti, kad DI vystytųsi ne beatodairiškai, o atsakingai ir skaidriai, atsižvelgiant į visuomenės interesus.

[63] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv preprint arXiv:1802.07228.

[64] Hitaj, B., Ateniese, G., & Pérez-Cruz, F. (2019). PassGAN: A Deep Learning Approach for Password Guessing. International Conference on Applied Cryptography and Network Security, 217-237.

[65] Anderson, H. S., Woodbridge, J., & Filar, B. (2018). DeepDGA: Adversarially-Tuned Domain Generation and Detection. Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security, 13-21.

[66] Miller, J., Smith, R., & Robinson, T. (2020). AI-Driven Cyber Attacks on Financial Institutions: Assessing the Threat Landscape. Journal of Financial Crime, 27(1), 123-137.

3.2. Teisiniai mechanizmai DI rizikų mažinimui

Dirbtinio intelekto plėtra kelia ne tik technologinius, bet ir teisinius iššūkius, susijusius su privatumu, atsakomybe ir skaidrumu. Siekiant užtikrinti saugų ir etišką DI taikymą, pastaraisiais metais Europos Sąjungoje buvo sukurta išsami reguliavimo sistema. Vienas svarbiausių šios sistemos elementų – Dirbtinio intelekto aktas[67] (AI Act, 2024), kuris nustato DI rizikos lygius bei atitikties reikalavimus. Greta jo veikia Bendrasis duomenų apsaugos reglamentas [68](BDAR, 2016), užtikrinantis asmens duomenų apsaugą DI sistemose, bei Skaitmeninių paslaugų aktas[69] (SPA, 2022), skirtas kovai su DI generuojama dezinformacija.

DI atsakomybei reglamentuoti ES priėmė Dirbtinio intelekto atsakomybės direktyvą [70](2022), kuri apibrėžia teisinę atsakomybę už DI sukeltą žalą.

Lietuvos teisė taip pat prisitaiko prie ES reguliavimo – nacionalinė Lietuvos dirbtinio intelekto strategija[71] (2018) numato principus, kuriais vadovaujamasi plėtojant DI reguliavimą ir valdymą šalyje. Šiame skyriuje bus nagrinėjami pagrindiniai ES ir Lietuvos teisės aktai, reglamentuojantys DI rizikų mažinimą, jų taikymo sritis bei poveikį organizacijoms.

3.2.1. Europos Sąjungos teisės aktai DI rizikų reguliavimui

Dirbtinio intelekto aktas[72] *(angl.. AI Act)* – pirmasis pasaulyje išsamus teisės aktas, skirtas DI sistemų reguliavimui. Europos Komisija jį pasiūlė 2021 m., o galutinai priėmė 2024 m., siekdama užtikrinti, kad DI technologijos būtų saugios, skaidrios ir etiškos bei atitiktų pagrindines žmogaus teises ir ES vertybes. Šis reglamentas nustato teisinius reikalavimus DI sistemoms, atsižvelgiant į jų rizikos lygį.

Rizikos klasifikavimas pagal dirbtinio intelekto aktą, viena pagrindinių naujovių – DI sistemų skirstymas į keturias rizikos kategorijas, kurios lemia, kokie reikalavimai joms taikomi:

- **Nepriimtinos rizikos DI** – visiškai draudžiamos DI sistemos, kurios kelia grėsmę žmogaus teisėms (pvz., socialinis reitingavimas, manipuliacinės elgsenos kontrolė, tam tikros biometrinės stebėsenos formos).
- **Didelės rizikos DI** – DI sistemos, naudojamos kritinėse srityse, tokiose kaip sveikatos priežiūra, teisingumo sistema, transportas ar įdarbinimas. Joms taikomi griežti saugumo, skaidrumo ir atitikties reikalavimai.

[67] Europos Parlamento ir Tarybos Reglamentas (ES) 2024/1689
[68] Europos Parlamento ir Tarybos Reglamentas (ES) 2016/679
[69] Europos Parlamento ir Tarybos Reglamentas (ES) 2022/2065
[70] Europos Parlamento ir Tarybos Reglamentas dėl nesutartinės civilinės atsakomybės taisyklių pritaikymo dirbtiniam intelektui (Atsakomybės už dirbtinį intelektą direktyva) (ES) 2022/0303
[71] Lietuvos Dirbtinio Intelekto Strategija | Ateities Vizija (2018)
[72] Europos Parlamento ir Tarybos Reglamentas (ES) 2024/1689

- **Ribotos rizikos DI** – sistemos, kurios tiesiogiai nekelia grėsmės, bet gali paveikti naudotojų sprendimus (pvz., DI generuojamas turinys). Joms taikomi skaidrumo reikalavimai, įskaitant pareigą informuoti naudotojus, kad jie sąveikauja su DI.
- **Minimalios rizikos DI** – visos kitos DI sistemos (pvz., DI vertėjai, paieškos algoritmai), kurioms netaikomi specifiniai reguliavimo reikalavimai.

Be rizikos klasifikavimo, dirbtinio intelekto aktas įpareigoja organizacijas laikytis skaidrumo ir atitikties reikalavimo. Didelės rizikos DI sistemų kūrėjai turės atlikti griežtą vertinimą prieš pateikdami produktą rinkai. Taip pat turi būti privaloma užtikrinti aiškų sprendimų paaiškinamumą bei išvengti algoritmų šališkumo.

Kadangi aktas yra tiesiogiai taikomas visose ES valstybėse narėse, visos organizacijos, kuriančios ir naudojančios DI technologijas, Europos Sąjungos valstybėse turės prisitaikyti prie naujų reikalavimų. Šis reglamentas padės užtikrinti vieningą DI reguliavimo sistemą visoje ES rinkoje, išvengiant fragmentacijos tarp skirtingų šalių reguliavimo modelių.

3.2.2. EU Bendrasis duomenų apsaugos reglamentas ir jo reikšmė DI

ES Bendrasis duomenų apsaugos reglamentas (BDAR) *(angl. General Data Protection Regulation, GDPR)* – tai pagrindinis Europos Sąjungos teisės aktas, skirtas asmens duomenų apsaugai užtikrinti. Jis įsigaliojo 2018 m. gegužės 25 d. ir taikomas visoms organizacijoms, kurios renka, naudoja ar kitaip tvarko ES piliečių duomenis, nepriklausomai nuo to, kur jos įsikūrusios [73].

Kadangi dirbtinio intelekto (DI) sistemos remiasi didžiuliais duomenų kiekiais, dažnai apimančiais ir asmens duomenis, BDAR vaidina pagrindinį vaidmenį reguliuojant, kaip šios technologijos kuriamos ir naudojamos. Reglamentas užtikrina, kad privatumo principai būtų įtraukti į DI sprendimus nuo pat jų kūrimo pradžios, o naudotojų teisės – tinkamai apsaugotos.

BDAR principai ir jų svarba DI sistemoms.

BDAR nustato kelis privatumo ir duomenų apsaugos principus, kurie yra ypač svarbūs DI kūrėjams ir naudotojams:

- **Skaidrumas ir sąžiningumas** – naudotojai turi būti aiškiai informuoti, kaip ir kodėl jų duomenys tvarkomi.
- **Duomenų kiekio mažinimas** – organizacijos neturėtų rinkti daugiau asmens duomenų, nei būtina konkrečiam tikslui pasiekti.
- **Tikslų apribojimas** – surinkti duomenys gali būti naudojami tik aiškiai apibrėžtais ir teisėtais tikslais.

[73] Europos Parlamento ir Tarybos Reglamentas (ES) 2016/679

- **Atsakomybė ir atskaitomybė –** organizacijos turi ne tik laikytis šių reikalavimų, bet ir gebėti įrodyti, kad jų veikla atitinka GDPR standartus.

Šie principai ypač svarbūs, nes DI modeliai dažnai veikia tarsi „juodosios dėžės“ – jų sprendimų priėmimo logika galutiniams naudotojams gali būti sunkiai suprantama arba visiškai neaiški. Todėl BDAR siekia užtikrinti, kad DI sprendimai būtų ne tik tikslūs ir efektyvūs, bet ir etiški bei skaidrūs.

Konkretūs BDAR reikalavimai DI kūrėjams ir naudotojams

Be bendrųjų principų, BDAR numato konkrečias taisykles, kurių privalo laikytis DI kūrėjai:

- **Teisė būti informuotam** – naudotojai turi žinoti, kokiais jų duomenimis remiasi DI sistema.
- **Teisė į duomenų prieigą ir ištrynimą („teisė būti pamirštam“)** – asmenys gali reikalauti ištrinti savo duomenis, o tai gali kelti iššūkių DI modeliams, kurie mokosi iš didžiulių duomenų rinkinių.
- **Draudimas priimti svarbius sprendimus vien automatizuotomis priemonėmis** – tai ypač svarbu tokiose srityse kaip įdarbinimas ar kreditingumo vertinimas, kur DI sprendimai gali tiesiogiai paveikti žmogaus gyvenimą.
- **Duomenų apsaugos poveikio vertinimas (DPIA)** – būtinas tais atvejais, kai dirbtinio intelekto (DI) sistema gali reikšmingai paveikti naudotojų teises ar laisves. Organizacija tokiu atveju turi įvertinti galimus privatumo pažeidimus ir numatyti priemones jiems išvengti. Nors BDAR ir AI Act yra atskiri teisės aktai, jų taikymas dažnai persidengia. BDAR saugo asmens duomenis, o AI Act nustato reikalavimus DI sistemų skaidrumui ir saugumui. Jei DI sistema laikoma didelės rizikos, jai taikomi ir griežti BDAR reikalavimai – tokiose situacijose DPIA tampa ne tik formalumas, bet ir būtina sąlyga užtikrinti atitiktį bei vartotojų pasitikėjimą.

Pavyzdžiui:

- Jei pagal AI Act DI sistema priskiriama didelės rizikos kategorijai, tikėtina, kad jai bus taikomi ir griežtesni BDAR reikalavimai, ypač tie, kurie susiję su asmens duomenų tvarkymo teisėtumu, skaidrumu bei naudotojų teisių apsauga.
- AI Act numato, kad DI sprendimai privalo būti skaidrūs ir paaiškinami, o BDAR suteikia vartotojams teisę suprasti, kaip jų duomenys yra naudojami.

Abi reguliavimo priemonės kartu padeda sukurti etišką, saugią ir skaidrią DI ekosistemą, kurioje suderinami inovacijų plėtra ir žmogaus teisių apsauga.

3.2.3. Skaitmeninių paslaugų aktas (DSA, 2022)

Skaitmeninių paslaugų aktas [74](*angl. Digital Services Act, DSA*) yra vienas svarbiausių Europos Sąjungos dokumentų, skirtų reguliuoti skaitmeninę erdvę ir joje kylančias rizikas. Šis reglamentas priimtas 2022 m. spalį, o visiškai pradėtas taikyti nuo 2024 m. vasario 17 d.

Jo pagrindinis tikslas – užtikrinti saugesnę, skaidresnę interneto aplinką. Aktas numato aiškias taisykles skaitmeninėms platformoms: kaip jos turi elgtis su neteisėtu turiniu, kaip riboti dezinformacijos plitimą ir pasirūpinti, kad dirbtinio intelekto algoritmai nedarytų žalos visuomenei.

DSA taikymo sritis ir pagrindiniai tikslai

Šis aktas taikomas visoms skaitmeninėms paslaugoms, veikiančioms ES teritorijoje, tačiau didžiausią poveikį daro didžiosioms interneto platformoms, tokioms kaip „Google“, „Meta“, „Amazon“ ar „TikTok“. Pagrindiniai DSA tikslai:

1. Kova su neteisėtu turiniu ir dezinformacija – skaitmeninės platformos pagal DSA privalo greitai reaguoti į žalingo turinio plitimą. Tai taikoma ne tik akivaizdžiai neteisėtai informacijai, bet ir DI sukurtam turiniui – tokiam kaip deepfake vaizdo įrašai, klaidinančios naujienos ar kita manipuliatyviai pateiktai medžiagai.
2. DI algoritmų skaidrumas – skaitmeninės platformos privalo atskleisti, kaip veikia jų DI valdomi rekomendacijų algoritmai ir pagal kokius kriterijus personalizuojamas turinys.
3. Vartotojų apsauga – DI pagrįstų sistemų naudojimas reklamoje ir turinio platinime turi būti aiškiai nurodytas, kad žmonės žinotų, kada jie sąveikauja su automatizuotomis sistemomis.
4. Atsakomybė už algoritminį turinio moderavimą – didžiosios platformos turi užtikrinti, kad DI naudojami turinio moderavimo įrankiai veiktų skaidriai ir nepažeistų žmogaus teisių.

DSA ir DI generuojamo turinio reguliavimas

Vienas svarbiausių DSA aspektų yra taisyklės, tiesiogiai susijusios su DI pagrįstu turinio kūrimu ir sklaida. Kadangi DI gali būti naudojamas manipuliaciniui turiniui, dezinformacijai ar klaidinančiai reklamai generuoti, DSA numato keletą reguliavimo priemonių:

- ***Deepfake* ir DI generuojamos dezinformacijos kontrolė** – platformos privalo turėti mechanizmus, leidžiančius atpažinti ir pažymėti automatizuotai sukurtą turinį.

[74] Europos Parlamento ir Tarybos Reglamentas (ES) 2022/2065

- **Reklamos skaidrumas** – visos DI generuojamos personalizuotos reklamos turi būti aiškiai identifikuojamos, o naudotojai turi gauti informaciją apie tai, kodėl jiems rodomas būtent toks turinys.
- **Priežiūros institucijų įgaliojimai** – ES valstybės narės gali reikalauti iš platformų aiškumo apie DI modelius, naudojamus turinio filtravimui ir rekomendacijoms.

Skaitmeninių paslaugų aktas (DSA) glaudžiai susijęs su kitais ES teisės aktais, ypač su Dirbtinio intelekto aktu (AI Act) ir Bendrąja duomenų apsaugos reglamento (BDAR) nuostatomis. Skirtingai nuo AI Act, kuris tiesiogiai reguliuoja dirbtinio intelekto technologijas nustatydamas jų rizikos lygius ir atitikties reikalavimus, Skaitmeninių paslaugų aktas (DSA) daugiausia orientuotas į skaitmeninių paslaugų reguliavimą plačiame kontekste. AI Act užtikrina, kad didelės rizikos DI sistemos veiktų saugiai, skaidriai ir etiškai, o DSA nustato sąlygas, kaip tokios technologijos gali būti taikomos internetinėje erdvėje – nuo turinio moderavimo iki reklamos ir informacijos platinimo.

Tuo tarpu BDAR apibrėžia duomenų apsaugos reikalavimus visoms DI sistemoms, kurios tvarko asmens duomenis. Pavyzdžiui:

- DSA įpareigoja skaitmenines platformas atskleisti, kaip veikia jų DI pagrįsti rekomendacijų algoritmai.
- BDAR užtikrina, kad šie algoritmai nepažeistų naudotojų privatumo teisių.

Kartu šie teisės aktai sudaro ES reguliavimo sistemą, skirtą mažinti DI naudojimo rizikas skaitmeninėje erdvėje ir apsaugoti naudotojus nuo manipuliacinio ar žalingo turinio.

3.2.4. Dirbtinio intelekto atsakomybės direktyva (2022)

Pagal šiuo metu Europos Sąjungoje galiojančią civilinės atsakomybės sistemą, žalą patyręs asmuo privalo įrodyti priežastinį ryšį tarp tam tikro veiksmo ir jo pasekmių. Tačiau DI atveju šis mechanizmas tampa problemiškas – dirbtinio intelekto modeliai veikia kompleksiškai, jų sprendimų priėmimo procesai dažnai nėra visiškai skaidrūs, o atsakomybės nustatymas gali būti sudėtingas.

Siekdama išspręsti šią problemą, Europos Komisija 2022 m. pristatė Dirbtinio intelekto atsakomybės direktyvos projektą[75], kurio tikslas – aiškiau apibrėžti atsakomybės mechanizmus ir palengvinti žalą patyrusių asmenų galimybes įrodyti patirtą nuostolį bei gauti kompensaciją.

Svarbiausios šios direktyvos nuostatos:

- **Įrodinėjimo naštos perkėlimas**. Įprastai nukentėjęs asmuo turi įrodyti priežastinį ryšį tarp DI sistemos veiksmų ir žalos. Tačiau direktyvoje numatytais atvejais ši našta gali būti

[75] Europos Parlamento ir Tarybos Reglamentas dėl nesutartinės civilinės atsakomybės taisyklių pritaikymo dirbtiniam intelektui (Atsakomybės už dirbtinį intelektą direktyva) (ES) 2022/0303

perkeliama DI kūrėjams ar naudotojams. Tai labai svarbu sudėtingoms DI sistemoms, kurios veikia kaip *„juodosios dėžės“* (angl. black box) – kai jų veikimo principai nėra aiškiai suprantami, nukentėjusiesiems gali būti sudėtinga įrodyti ryšį tarp DI sprendimo ir padarytos žalos.

- **Atsakomybės lygmenys pagal DI riziką**. Direktyva glaudžiai susijusi su Dirbtinio intelekto aktu (AI Act), kuris nustato skirtingų DI sistemų rizikos lygius. Kuo didesnę riziką DI sistema kelia, tuo griežtesni atsakomybės standartai jai taikomi.
- **Didesnė apsauga vartotojams**. Jeigu dirbtinio intelekto (DI) sistema pažeidžia Bendrojo duomenų apsaugos reglamento (BDAR) nuostatas ir netinkamai tvarko asmens duomenis, direktyva suteikia papildomus mechanizmus, leidžiančius vartotojams greičiau ir veiksmingiau reikalauti kompensacijos. Tai sustiprina skaidrumo ir žmogaus teisių apsaugos principus DI taikyme.

Ši direktyva papildo kitus svarbius ES teisės aktus – AI Act, reglamentuojantį DI modelių kūrimą ir naudojimą, bei BDAR, saugantį asmens duomenis. Kartu jie sudaro bendrą teisinį pagrindą, užtikrinantį atsakingą dirbtinio intelekto taikymą Europos Sąjungoje.

3.2.5. Skaitmeninių rinkų aktas ir jo sąsajos su dirbtiniu intelektu

vDidžiosios skaitmeninės platformos plačiai diegia dirbtinio intelekto sprendimus – nuo turinio personalizavimo iki pajamų augimo skatinimo. Tačiau tokie algoritmai ne visuomet atneša tik naudą. Jie gali skatinti rinkos koncentraciją, diskriminaciją ar iškraipyti konkurenciją, nes didžiausiems žaidėjams suteikia dar daugiau galios. Būtent tokioms situacijoms spręsti buvo priimtas Skaitmeninių rinkų aktas (Digital Markets Act, DMA), kuris įsigaliojo 2022 m. lapkritį, o realiai taikomas nuo 2023 m. kovo. DMA dėmesys sutelktas į vadinamuosius „vartų sargus“ (angl. gatekeepers)Didžiosios skaitmeninės platformos vis plačiau diegia dirbtinio intelekto technologijas – taip jos siekia pritaikyti turinį kiekvienam vartotojui, pagerinti naudojimosi patirtį ir auginti pajamas. Tačiau toks algoritmų taikymas ne visada atneša tik teigiamus padarinius. Kai kuriais atvejais DI gali skatinti rinkos koncentraciją, palaikyti diskriminacines praktikas ar riboti konkurenciją, ypač kai naudą iš to gauna jau dominuojantys rinkos žaidėjai.tai didžiosios platformos, turinčios dominuojančią padėtį rinkoje ir kontroliuojančios prieigą prie skaitmeninių paslaugų. Tarp jų – Google, Meta, Amazon, Apple, Microsoft ir kiti, valdantys milžiniškus vartotojų tinklus.

Šias problemas spręsti imtasi pasitelkus Skaitmeninių rinkų aktą (Digital Markets Act, DMA), kuris buvo priimtas 2022 m. lapkritį, o taikomas pradėtas nuo 2023 m. kovo mėnesio[76].

[76] Europos Parlamento ir Tarybos Reglamentas(ES) 2022/1925 (2022) dėl atvirų konkurencijai ir sąžiningų skaitmeninio sektoriaus rinkų.

DMA reglamentuoja vadinamąsias „vartų sargų“ (angl. gatekeepers) platformas – tai stambios technologijų įmonės, turinčios reikšmingą įtaką skaitmeninėje erdvėje ir kontroliuojančios prieigą prie svarbiausių internetinių paslaugų.

Tarp tokių bendrovių patenka Google, Meta, Amazon, Apple ir Microsoft – visos jos valdo platformas su milžinišku vartotojų skaičiumi ir reikšminga įtaka skaitmeninėse rinkose.

Pagal DMA reikalavimus, stambūs paslaugų tiekėjai turi laikytis kelių principų:

- DI valdomi algoritmai turi būti skaidrūs ir negali pažeisti sąžiningos konkurencijos. Pavyzdžiui, „Google“ negali manipuliuoti savo paieškos algoritmais tam, kad jos produktai būtų iškelti aukščiau nei konkurentų.
- Nesąžiningas duomenų kaupimas draudžiamas, kad būtų užkirstas kelias rinkos pranašumo siekimui nesąžiningomis priemonėmis. Jei įmonė naudoja DI vartotojų elgsenos analizei keliose platformose ir tai išnaudoja konkurencijos ribojimui, tai gali būti laikoma pažeidimu.
- Vartotojai turi turėti didesnę kontrolę savo sąveikai su DI sistemomis. Platformos turi sudaryti galimybę atsisakyti tam tikrų DI pagrįstų personalizavimo funkcijų arba bent jau aiškiai informuoti vartotojus apie jų veikimo principus.

Nors Skaitmeninių rinkų aktas (DMA) tiesiogiai nereguliuoja dirbtinio intelekto, jo įtaka DI naudojimui skaitmeninėje ekonomikoje yra gana ryški. Šis aktas glaudžiai susijęs su kitais svarbiais ES dokumentais, tokiais kaip AI Act, kuris apibrėžia DI veiklos standartus, ir BDAR, užtikrinančiu asmens duomenų apsaugą dirbtinio intelekto sistemose.

DMA pagrindinis tikslas – užkirsti kelią nesąžiningam DI technologijų naudojimui, ypač kai jos išnaudojamos riboti konkurenciją, stiprinti dominuojančių bendrovių pozicijas ar manipuliuoti vartotojų elgsena naudojantis jų duomenimis. Toks reguliacinis požiūris aiškiai parodo, kad Europos Sąjungai rūpi ne tik technologijų kūrimo procesas, bet ir jų realus poveikis rinkoje.

3.2.6. Lietuvos asmens duomenų teisinės apsaugos įstatymas

Lietuvoje asmens duomenų apsaugą reguliuoja Lietuvos asmens duomenų teisinės apsaugos įstatymas[77], kuris įgyvendina BDAR nacionaliniu lygmeniu ir nustato papildomas taisykles, susijusias su duomenų tvarkymu bei priežiūra. Pirmą kartą šis įstatymas buvo priimtas 1996 m., tačiau vėliau ne kartą atnaujintas, siekiant jį suderinti su ES teisiniais reikalavimais, ypač po GDPR (lietuvių.k. BDAR) įsigaliojimo 2018 m. Šio įstatymo svarba DI kontekste kyla iš to, kad dirbtinio intelekto modeliai dažnai grindžiami didelės apimties duomenų analize, kuri gali apimti jautrią asmeninę informaciją. Lietuvos

[77] Lietuvos Respublikos Asmens Duomenų Teisinės Apsaugos Įstatymas Nr. I-1374

teisės aktas papildo BDAR, suteikdamas aiškesnę vietinę reguliacinę bazę tiek viešajam sektoriui, tiek verslui, siekiančiam taikyti DI sprendimus.

Pagrindinės įstatymo nuostatos, aktualios DI reguliavimui:

- **Duomenų apsaugos priežiūra**. Už įstatymo laikymąsi atsakinga Valstybinė duomenų apsaugos inspekcija (VDAI), turinti teisę atlikti patikrinimus, tirti pažeidimus ir taikyti administracines priemones. Tai labai svarbu DI modeliams, apdorojantiems asmens duomenis, nes VDAI gali įpareigoti įmones keisti DI veikimo principus, jei nustatomi privatumo pažeidimai.
- **Viešojo sektoriaus prievolės**. Lietuvos teisės aktas nustato griežtesnius reikalavimus valstybinėms institucijoms, kurios diegia DI sprendimus. Pavyzdžiui, automatizuoto sprendimų priėmimo ar stebėjimo sistemos viešajame sektoriuje turi būti ne tik teisėtos, bet ir proporcingos bei pagrįstos aiškiais kriterijais.
- **Asmens teisės DI kontekste**. Lietuvos įstatymas išlaiko BDAR numatytas asmens teises, tokias kaip teisė būti informuotam, teisė nesutikti su automatizuotu sprendimų priėmimu ir teisė būti pamirštam. Papildomai akcentuojama teisė kreiptis į nacionalines priežiūros institucijas, jei asmens duomenys naudojami neteisėtai.
- **Sankcijos ir atsakomybė**. Nors BDAR numato griežtas finansines baudas, Lietuvos teisės aktas detalizuoja administracinės atsakomybės procedūras, kurios taikomos tiek fiziniams, tiek juridiniams asmenims už netinkamą asmens duomenų tvarkymą DI sistemose.

Lietuvos asmens duomenų teisinės apsaugos įstatymas papildo BDAR, suteikdamas aiškesnius nacionalinius duomenų apsaugos standartus ir priežiūros mechanizmus. DI sistemų kūrėjai ir naudotojai privalo atsižvelgti į šio įstatymo reikalavimus, ypač jei jų technologijos apdoroja jautrius asmens duomenis.

3.2.7. Kibernetinio saugumo įstatymas ir jo sąsaja su DI rizikų valdymu

Kadangi DI sistemos veikia sudėtingose IT infrastruktūrose, jos yra jautrios įvairiems kibernetiniams pavojams, įskaitant duomenų vagystes, modelių manipuliavimą (angl. adversarial attacks), užnuodijimo atakas (angl. data poisoning) ir kitus trikdžius, galinčius sutrikdyti jų veikimą. Lietuvoje DI sistemų saugumą iš dalies reglamentuoja Kibernetinio saugumo įstatymas, kuris nustato reikalavimus IT sistemų apsaugai, atsakomybę už incidentų valdymą bei prevencines priemones nuo kibernetinių grėsmių.

Kibernetinio saugumo įstatymas[78], pirmą kartą priimtas 2014 m., nuolat atnaujinamas siekiant jį suderinti su Europos Sąjungos direktyvomis, įskaitant TIS2 [79](*angl.* NIS2) direktyvą (Tinklų ir informacinių sistemų saugumo direktyvą). Tai pagrindinis teisės aktas, reglamentuojantis nacionalinę kibernetinio saugumo politiką ir jos įgyvendinimą.

Jis apima šiuos aspektus:

- **Kibernetinio saugumo principai** – įtvirtinamos priemonės, skirtos užtikrinti informacijos konfidencialumą, vientisumą ir prieinamumą.
- **Institucijų atsakomybė** – pagrindinės institucijos, atsakingos už nacionalinį saugumą šioje srityje, yra Krašto apsaugos ministerija ir Nacionalinis kibernetinio saugumo centras, kurios prižiūri strateginį saugumo lygį valstybėje.
- **Viešojo ir privataus sektoriaus pareigos** – tam tikros valstybinės institucijos bei privatūs subjektai, priskiriami prie kritinės IT infrastruktūros, privalo įgyvendinti aukšto lygio saugumo priemones.

Kalbant apie DI technologijų saugumą, šis įstatymas yra reikšmingas keliais aspektais:

1. **DI sistemų atsparumas kibernetinėms atakoms.** Organizacijos, diegiančios dirbtinio intelekto sprendimus jautriose srityse – tokiose kaip finansai, energetika ar sveikatos priežiūra – turi pareigą užtikrinti itin aukštą saugumo lygį. Tokios sistemos neretai tampa kibernetinių atakų taikiniu, todėl joms taikomi sugriežtinti reikalavimai. Todėl saugumo priemonės apima ne tik pačių sistemų testavimą, bet ir išankstinį rizikų įvertinimą bei pasirengimą greitai ir efektyviai reaguoti į galimus saugumo pažeidimus.
2. **Asmens duomenų apsauga ir DI sąveika.** Kadangi DI modeliai dažnai apdoroja jautrią informaciją, reikalinga užtikrinti duomenų saugumą ir vientisumą. Įstatyme numatytos organizacinės ir techninės priemonės padeda sumažinti riziką, susijusią su duomenų nutekėjimu ar neteisėtu jų naudojimu mokymo procesuose.
3. **Grėsmių prevencija autonominėms DI sistemoms.** Augant autonominių DI sprendimų naudojimui, ypač transporto, saugumo ir gynybos srityse, atsiranda naujų grėsmių, susijusių su sistemų manipuliacija. Kibernetinio saugumo įstatymas numato apsaugos standartus, kurie padeda apsaugoti tokias sistemas nuo įsilaužimų ar piktavališko poveikio.

Lietuvos Kibernetinio saugumo įstatymas, kartu su ES NIS2 direktyva, sudaro teisinį pagrindą, skirtą DI sistemų apsaugai nuo išorinių grėsmių. Vis dėlto, šiame įstatyme nėra tiesioginių DI

[78] Lietuvos Respublikos Kibernetinio Saugumo Įstatymas (2014)

[79] EUROPOS PARLAMENTO IR TARYBOS DIREKTYVA (ES) 2022/2555 2022 m. gruodžio 14 d. dėl priemonių aukštam bendram kibernetinio saugumo lygiui visoje Sąjungoje užtikrinti, kuria iš dalies keičiamas Reglamentas (ES) Nr. 910/2014 ir Direktyva (ES) 2018/1972 ir panaikinama Direktyva (ES) 2016/114

reguliavimo nuostatų, todėl praktiškai DI saugumo reikalavimai taikomi per bendrą IT infrastruktūros apsaugos sistemą.

3.2.8. Informacinės visuomenės paslaugų įstatymas ir jo sąsaja su DI rizikų valdymu

Lietuvoje priimtas Informacinės visuomenės paslaugų įstatymas [80]reglamentuoja skaitmeninių paslaugų teikimą bei jų teikėjų atsakomybę. Šis teisės aktas yra itin svarbus dirbtinio intelekto (DI) reguliavimui, nes nustato teisinius principus, ribojančius DI naudojimą neteisėtai informacijos sklaidai, manipuliacijoms bei netinkamam vartotojų duomenų tvarkymui.

Pagrindiniai įstatymo aspektai, susiję su DI:

- **Skaitmeninių paslaugų teikėjų atsakomybė.** Paslaugų, kurios naudoja DI sprendimus (pvz., turinio rekomendacijų algoritmai, automatizuoti klientų aptarnavimo asistentai), teikėjai privalo užtikrinti, kad DI sistema nepažeistų teisės aktų, nediskriminuotų vartotojų ir neskatintų neteisėtos veiklos.
- **Atsakomybė už DI generuojamą turinį.** Jei DI pagrįstos sistemos sukuria ar platina klaidinančią, neteisėtą ar apgaulingą informaciją, platformų operatoriai gali būti laikomi atsakingais už jos poveikį vartotojams. Tai turi didelę reikšmę socialinių tinklų algoritmams ir automatiškai generuojamam turiniui.
- **Vartotojų teisės ir skaidrumo reikalavimai.** Įstatymas numato, kad DI naudojimas skaitmeninėse paslaugose turi būti skaidrus ir aiškiai paaiškinamas vartotojams. Tai reiškia, kad jei platforma ar paslauga naudoja DI priimant sprendimus dėl informacijos pateikimo, naudotojai turi būti informuoti apie šią aplinkybę.

Lietuvoje galiojantis Informacinės visuomenės paslaugų įstatymas yra glaudžiai susijęs su ES Skaitmeninių paslaugų aktu [81] (Digital Services Act, DSA), kuris nustato bendras skaitmeninių paslaugų taisykles visoje Europos Sąjungoje. DSA įpareigoja didžiausias interneto platformas laikytis papildomų skaidrumo, atskaitomybės ir turinio moderavimo reikalavimų, kuriuos Lietuvos įstatymas papildo nacionaliniu lygmeniu.

3.2.9. Nutarimas dėl dirbtinio intelekto technologijų naudojimo viešajame sektoriuje

Lietuvos Respublikos Vyriausybė, siekdama užtikrinti skaidrų ir etišką DI taikymą, taip pat atitiktį teisiniams reikalavimams, 2021 m. priėmė nutarimą „Dėl dirbtinio intelekto technologijų naudojimo

[80] Lietuvos Respublikos Informacinės Visuomenės Paslaugų Įstatymas 2006 m. gegužės 25 d. Nr. X-614

[81] Europos Parlamento ir Tarybos Reglamentas 2022/2065 2022 m. spalio 19 d. dėl bendrosios skaitmeninių paslaugų rinkos, kuriuo iš dalies keičiama Direktyva 2000/31/EB (Skaitmeninių paslaugų aktas)

viešajame sektoriuje principų“. Šis dokumentas[82] nėra privalomas teisės aktas, tačiau jis pateikia gaires, kurių turėtų laikytis valstybės institucijos, diegdamos DI sprendimus. Nutarime siekiama užtikrinti, kad DI technologijos nebūtų naudojamos būdais, kurie galėtų kelti grėsmę žmogaus teisėms, pažeisti privatumą ar neskaidriai priimti sprendimus. Pagrindiniai nutarime nustatyti principai:

- **Žmogaus teisių apsauga.** Valstybinės institucijos, diegdamos DI technologijas, privalo užtikrinti, kad jos nepažeistų pagrindinių žmogaus teisių ir laisvių. Tai itin svarbu automatizuoto sprendimų priėmimo atvejais, kai DI gali būti naudojamas socialinių paslaugų skyrimui, teisėsaugos veikloje ar kitose srityse, darančiose tiesioginį poveikį piliečiams.
- **Skaidrumas ir atskaitomybė.** DI priimami sprendimai turėtų būti paaiškinami ir pagrįsti. Jei valstybės institucijos remiasi DI sistemomis, priimdamos sprendimus, turinčius teisinį ar ekonominį poveikį asmenims (pvz., socialinių išmokų skyrimas, teisėsaugos veiksmai), asmenys turi turėti galimybę suprasti sprendimo logiką ir prireikus jį apskųsti.
- **Etiškas ir atsakingas naudojimas.** DI algoritmai neturi būti naudojami taip, kad galėtų stiprinti diskriminaciją ar šališkumą. Pavyzdžiui, jei automatizuotos sistemos daro įtaką piliečių prieigai prie viešųjų paslaugų, turi būti įvertinta, ar jos veikia objektyviai ir teisingai.
- **Duomenų apsauga.** Valstybės institucijos, naudodamos DI technologijas, privalo užtikrinti aukščiausius duomenų apsaugos standartus. Tai reiškia, kad jų veikla turi atitikti Bendrojo duomenų apsaugos reglamento (BDAR) reikalavimus bei Lietuvos Asmens duomenų teisinės apsaugos įstatymą.
- **Viešojo administravimo efektyvumo didinimas.** DI gali padėti optimizuoti administracinius procesus, sumažinti biurokratinę naštą, tačiau technologijų naudojimas neturėtų riboti piliečių teisių ar mažinti administracinių sprendimų skaidrumo.

DI taikymo sritys viešajame sektoriuje Lietuvoje DI jau naudojamas įvairiose valstybinėse institucijose. Nutarime išskiriamos kelios pagrindinės sritys, kuriose DI gali būti taikomas:

- **Teisėsauga ir viešasis saugumas.** Automatizuotos sistemos gali padėti užtikrinti kibernetinį saugumą, tačiau jos negali pažeisti privatumo teisių ar būti naudojamos masinei stebėsenai.
- **Sveikatos apsauga.** DI gali būti naudojamas ligų diagnostikai, medicininių duomenų analizei, tačiau būtina užtikrinti pacientų duomenų konfidencialumą.
- **Mokesčių administravimas.** DI gali padėti aptikti galimus mokesčių pažeidimus, tačiau jis negali tapti įrankiu pertekliniam duomenų rinkimui ar analizės šališkumui.

[82] Rezoliucija Dėl Dirbtinio Intelekto Technologijų Naudojimo Viešajame Sektoriuje Principų 2024 Nr. Xiv-2620

Šis dokumentas laikomas pirmuoju oficialiu žingsniu siekiant apibrėžti, kaip dirbtinio intelekto sprendimai turėtų būti taikomi viešajame sektoriuje. Jis glaudžiai susijęs su platesniu Europos Sąjungos teisės aktų kontekstu, įskaitant ir AI Act, ir signalizuoja, kad Lietuvos viešajam sektoriui ateityje teks prisitaikyti prie vis griežtesnių DI reguliavimo reikalavimų.

Tiesa, kadangi nutarimas nėra teisiškai privalomas, jo įgyvendinimas iš esmės priklausys nuo pačių institucijų iniciatyvos bei pasirengimo taikyti atsakingo naudojimo principus. Tai reiškia, jog realūs pokyčiai priklausys ne tik nuo reglamentavimo, bet ir nuo pačių organizacijų požiūrio į DI diegimą.

Tai reiškia, kad šiuo metu nėra tiesioginių sankcijų už jo nesilaikymą, tačiau jis formuoja ilgesnio laikotarpio DI reguliavimo pagrindus, skatinant atsakingą ir etišką DI naudojimą valstybės sektoriuje.

Apibendrinant galima teigti, kad Europos Sąjungos ir Lietuvos teisinė sistema siekia užtikrinti suderintą, atsakingą dirbtinio intelekto reguliavimą, orientuotą į žmogaus teisių apsaugą, skaidrumą bei sąžiningą konkurenciją. Vis dėlto praktika rodo, kad teisė dažnai atsilieka nuo technologinės pažangos – daugelis spragų tampa matomos tik po to, kai DI sprendimai jau pradėti taikyti. Vienas aiškiausių to pavyzdžių – neapibrėžtumas dėl DI generuojamo turinio autorystės, kuriam iki šiol nėra suformuluotas vienareikšmis teisinis vertinimas. Todėl, nors dabartinė sistema jau rodo pažangą, jos tolesnė plėtra ir prisitaikymas prie dinamiškos technologijų raidos išlieka esminis uždavinys.

3.3. Tarptautiniai dirbtinio intelekto rizikų valdymo standartai

Ankstesniuose skyriuose aptartos pagrindinės dirbtinio intelekto (DI) technologijų keliamos grėsmės – nuo duomenų saugumo spragų iki teisinio reguliavimo neapibrėžtumo – paskatino tiek privačias, tiek viešąsias organizacijas ieškoti būdų, kaip efektyviau valdyti su šiomis technologijomis susijusias rizikas. Kai kur atsakomybės praktikos išsivystė natūraliai, kaip savireguliacijos iniciatyvos, grindžiamos vidiniu organizacijų vertybiniu pagrindu[83] (L. Floridi & J. Cowls, 2019). Vis dėlto vien geros valios nepakako – kai kuriais atvejais prireikė konkretesnių standartų ir griežtesnių teisinių priemonių, siekiant užtikrinti etišką ir skaidrų DI taikymą[84] (Europos Komisija, 2021).

Būtent šioje sąveikoje tarp technologijų reguliavimo ir institucinio atsakingumo vis didesnį vaidmenį įgauna tarptautiniai DI rizikos valdymo standartai. Kai kurie jų – kaip ISO 31000, skirtas rizikų valdymui, ar ISO 27001, orientuotas į informacijos saugumą – buvo sukurti dar iki DI proveržio, tačiau vėliau adaptuoti atsižvelgiant į naujus technologinius iššūkius. Šių standartų poveikis akivaizdus ir

[83] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

[84] European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).

naujausiuose teisės aktuose, įskaitant Europos Sąjungos AI Act, kuriame matyti tiek NIST AI Risk Management Framework[85] (2023), tiek naujų ISO gairių[86] (2023) įtaka.

Visa ši raidos kryptis rodo, kad tarptautiniai standartai veikia ne tik kaip pagalbiniai įrankiai organizacijoms, norinčioms tvariai valdyti DI rizikas, bet ir kaip esminiai elementai formuojant vis išsamesnį ir griežtesnį reguliavimo peizažą[87] (Brundage et al., 2020). Šiame skyriuje bus išsamiai analizuojami pagrindiniai DI rizikų valdymo standartai, jų taikymo galimybės bei praktiniai aspektai, padedantys atsakyti į svarbų klausimą: kaip organizacijos gali suderinti atsakingą DI naudojimą su nuolat griežtėjančiais teisiniais reikalavimais?

3.3.1. ISO standartų vaidmuo dirbtinio intelekto rizikų valdyme

Tarptautiniu mastu pripažinti standartai suteikia organizacijoms struktūrizuotą būdą identifikuoti, vertinti ir mažinti su dirbtiniu intelektu susijusias rizikas. Tarp jų ypač išskiriami keturi svarbiausi ISO standartai, kurie padeda užtikrinti rizikų valdymą ir atsakingą, bei saugų DI naudojimą:

ISO 31000 – nustato bendruosius rizikos valdymo principus, kuriuos organizacijos gali taikyti DI keliamų pavojų analizei ir valdymui.

ISO 27001 – skirtas informacijos saugumui, jis užtikrinana, kad DI sistemose naudojami duomenys būtų apsaugoti nuo neteisėtos prieigos ir kibernetinių grėsmių.

ISO 42001 – pirmasis tarptautinis DI valdymo standartas, kuris nustato gaires, kaip organizacijos turėtų kurti, diegti ir valdyti DI sistemas, laikydamosi skaidrumo, etikos ir reguliavimo reikalavimų.

ISO 23894 yra vienas pirmųjų tarptautiniu mastu priimtų standartų, skirtų būtent dirbtinio intelekto rizikų valdymui. Jis suteikia organizacijoms aiškią struktūrą, kaip nuosekliai atpažinti, įvertinti ir suvaldyti DI keliamas grėsmes. Dokumente ypatingas dėmesys skiriamas tokiems iššūkiams kaip algoritmų šališkumas, sprendimų nenuspėjamumas ar kibernetinės grėsmės, kurios ypač aktualios diegiant DI sprendimus praktinėje veikloje. Šio standarto taikymas padeda ne tik padidinti technologijų patikimumą, bet ir užtikrinti, kad jos būtų naudojamos etiškai bei socialiai atsakingai.

Kiekvienas iš šių standartų atlieka savitą vaidmenį organizacijos strategijoje – jie padeda kryptingai valdyti dirbtinio intelekto keliamas rizikas ir kartu prisideda prie atitikties tarptautiniams normatyvams užtikrinimo.

ISO standartai plačiai taikomi Europoje ir Azijoje, tačiau Jungtinėse Amerikos Valstijose organizacijos dažniau vadovaujasi Nacionalinio standartų ir technologijų instituto [88](NIST) parengta

[85] National Institute of Standards and Technology. (2023). NIST AI Risk Management Framework. NIST.
[86] International Organization for Standardization. (2023). ISO 42001: Artificial Intelligence Management System.

[87] Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., & Amodei, D. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. arXiv preprint arXiv:2004.07213.
[88] NIST Risk Management Framework https://csrc.nist.gov/projects/risk-management/about-rmf

rizikų valdymo sistema (Risk Management Framework, RMF), kuri siūlo labiau lokalizuotą požiūrį į DI rizikos kontrolę. Nepaisant šių regioninių skirtumų, tarptautiniu mastu ISO išlieka vienu iš pagrindinių standartų, suteikiančių organizacijoms nuoseklią struktūrą DI rizikų vertinimui, valdymui ir mažinimui.

3.3.1.1 ISO standartų vaidmuo dirbtinio intelekto rizikų valdyme

ISO 31000 yra tarptautinis rizikos valdymo standartas, kurį parengė Tarptautinė standartizacijos organizacija. Standartas pateikia bendrą metodologiją rizikų valdymui, nepriklausomai nuo jų pobūdžio – finansinių, technologinių, teisinių ar susijusių su DI. Nors šis standartas nėra skirtas konkrečiai IT ar DI sričiai, jis veikia kaip struktūrinis pagrindas organizacijoms, siekiančioms įgyvendinti nuoseklų rizikos valdymo procesą [89](Aven, 2016).

Pirmą kartą paskelbtas 2009 m., ISO 31000 buvo atnaujintas 2018 m. [90], siekiant integruoti modernesnes rizikos valdymo praktikas. Šis standartas nustato bendruosius rizikos valdymo principus, tačiau nepateikia detalių priemonių ar techninių reikalavimų, todėl organizacijos jį dažnai taiko kartu su kitais standartais, kurie suteikia konkretesnes gaires skirtingoms sritims – pavyzdžiui, ISO 27001 (informacijos saugumo valdymui) arba ISO 42001 (DI valdymui).

Pagrindiniai ISO 31000 principai:

ISO 31000 nustato keturis pagrindinius principus, kurie užtikrina veiksmingą rizikos valdymą:

- **Integracija į organizacinę veiklą** – rizikos valdymas turi būti neatsiejama organizacijos valdymo ir sprendimų priėmimo dalis.
- **Struktūruotas ir sistemingas požiūris** – organizacijos turėtų naudoti aiškią, metodologiškai pagrįstą rizikos valdymo strategiją.
- **Prisitaikymas prie konteksto** – rizikos valdymas turi būti pritaikomas įvairiems sektoriams, įskaitant DI naudojimą.
- **Nuolatinis tobulinimas** – Organizacijos turi reguliariai peržiūrėti ir atnaujinti savo rizikos valdymo strategijas, atsižvelgdamos į kintančius iššūkius ir technologijų raidą.

Kadangi dirbtinio intelekto technologijos gali kelti įvairaus pobūdžio rizikas, pavyzdžiui, autonominius sprendimus ar duomenų apsaugos pažeidimus, organizacijos dažnai pasitelkia ISO 31000 standartą kaip pagrindinį modelį rizikoms atpažinti, įvertinti ir valdyti [91] (Hussain et al., 2022). Vis dėlto šio standarto dažnai nepakanka, norint suvaldyti su DI specifines grėsmes. Dėl to jis turėtų būti

[89] Aven, T. (2016). Risk assessment and risk management: Review of recent advances on their foundation. European Journal of Operational Research, 253(1), 1-13.

[90] International Organization for Standardization. (2018). ISO 31000:2018 Risk management – Guidelines. ISO.

[91] Hussain, W., Hussain, O. K., Chang, E., & Dillon, T. (2022). Risk management in AI-driven decision support systems: Challenges and future research directions. Artificial Intelligence Review, 55(4), 3291-3315.

derinamas su kitais išsamesniais dokumentais, tokiais kaip ISO 27001, ISO 42001 ar ISO 23894, kurie padeda spręsti saugumo, valdymo ir etiško technologijų naudojimo klausimus.

3.3.1.2 ISO 27001 ir ISO 27002 reikšmė dirbtini intelekto rizikų valdymui

Dirbtinio intelekto sistemų saugumas glaudžiai susijęs su informacijos apsauga – jei duomenys nėra tinkamai apsaugoti, modeliai tampa pažeidžiami, o jų priimami sprendimai gali būti šališki ar net klaidingi. ISO/IEC 27001 – tai pagrindinis informacijos saugumo standartas, kuris nustato reikalavimus organizacijoms, kaip valdyti rizikas, susijusias su duomenų apsauga. Tuo tarpu ISO/IEC 27002 šį standartą papildo praktinėmis gairėmis, padedančiomis diegti veiksmingas saugumo priemones organizacijų viduje. Kadangi DI modeliai apdoroja didelius duomenų kiekius, informacijos apsauga tampa esmine rizikos valdymo dalimi – saugumo pažeidimai gali lemti ne tik technines klaidas, bet ir teisinę atsakomybę.

ISO 27001 yra sertifikuojamas standartas, apibrėžiantis, kaip organizacijos turėtų apsaugoti informaciją ir valdyti saugumo rizikas, kad tai leistų:

- Užtikrinti DI modelių mokymosi duomenų patikimumą.
- Apsaugoti duomenis nuo neteisėtos prieigos, nutekėjimo ar manipuliacijos.
- Įdiegti nuoseklią informacijos saugumo politiką, pritaikytą kibernetinių grėsmių kontekste.

ISO 27002 – Praktinės gairės saugumo užtikrinimui ISO 27002 pateikia konkrečias saugumo priemones, kurias organizacijos gali taikyti pagal ISO 27001. DI sistemų kontekste išskiriami keli pagrindiniai aspektai:

- **Prieigos kontrolė** – ribojama, kas gali pasiekti DI modelius ir duomenis.
- **Duomenų vientisumas** – užtikrinama, kad informacija nebūtų pakeista ar sugadinta.
- **Incidentų valdymas** – numatomos reakcijos į kibernetinius pažeidimus, įskaitant modelių nuodijimą (data poisoning) ar adversarines atakas.
- **Privatumo apsauga** – įdiegiami duomenų nuasmeninimo ir šifravimo sprendimai asmens duomenims apsaugoti.

ISO 27001 ir ISO 27002 yra esminiai informacijos saugumo standartai, kurie padeda organizacijoms apsaugoti dirbtinio intelekto modelius ir laikytis teisinių reikalavimų. ISO 27001 apibrėžia bendruosius informacijos saugumo valdymo principus, o ISO 27002 juos papildo praktinėmis gairėmis, kaip šiuos principus taikyti kasdienėje veikloje. Teisiniu požiūriu, šie standartai padeda organizacijoms laikytis tarptautinių normų, tokių kaip BDAR ar AI Act, ir taip sumažina riziką susidurti su sankcijomis.

Be to, jų įgyvendinimas stiprina organizacijos reputaciją tiek viduje, tiek išorėje – tai aiškus signalas partneriams, klientams ir visuomenei, kad informacijos saugumas vertinamas atsakingai. Kadangi DI ekosistema nuolat kinta ir susiduria su vis naujomis grėsmėmis, ISO 27001 ir ISO 27002 tampa ne tik atsaku į esamus iššūkius, bet ir prevencine priemone, padedančia organizacijoms būti pasiruošusioms ateičiai.

3.3.1.3 ISO 42001 Dirbtinio Intelekto Valdymo Standartizavimas

Skirtingai nuo standartų, kurie reglamentuoja technologinius aspektus, ISO/IEC 42001 nukreiptas į organizacinį dirbtinio intelekto sprendimų valdymą[92] – t. y. kaip institucijos turėtų planuoti, diegti ir prižiūrėti DI taikymą, siekdamos ne vien technologinio efektyvumo, bet ir atsakingo rizikų bei jų poveikio kontrolės. Dokumente pabrėžiama sprendimų atsekamumo, skaidrumo ir jų suderinamumo su teisės reikalavimais bei etikos normomis svarba. Pažymėtina, kad tai yra sertifikuojamas standartas – organizacijos gali formaliai patvirtinti atitiktį ISO/IEC 42001 per akredituotas sertifikavimo institucijas. Standarto įgyvendinimas remiasi cikliniu veiklos valdymo modeliu (PDCA – *Plan Do Check Adjust*: planavimas, įgyvendinimas, vertinimas, tobulinimas), kuris sudaro sąlygas nuosekliai valdyti procesus ir juos tobulinti atsižvelgiant į išorinius ir vidinius pokyčius.

Praktiškai tai apima šiuos pagrindinius aspektus:

- **Konteksto ir poveikio vertinimą.** Organizacijos turi aiškiai nustatyti, kur ir kaip bus naudojamos DI sistemos, kokį poveikį jos gali turėti vidinei veiklai bei išorinėms suinteresuotosioms šalims, ir kartu įvertinti tiek galimas rizikas, tiek potencialią naudą.
- **Atsakomybių ir politikos nustatymą.** Svarbu tiksliai apibrėžti, kas atsakingas už sprendimų priėmimą DI srityje, kaip nustatomos etikos gairės bei sprendimų prižiūrėjimo mechanizmai, ir užtikrinti, kad šios nuostatos būtų įgyvendinamos praktikoje.
- **Kontrolės priemonių įgyvendinimą.** Tai apima technines ir organizacines priemones, kuriomis siekiama sumažinti klaidų, šališkumo ar duomenų pažeidimų riziką. Taip pat būtina numatyti priemones incidentų valdymui ir nuolatinei atitikties priežiūrai.
- **Veiklos rezultatų analizę ir tobulinimą.** Standartas numato, kad organizacijos nuolat vertintų, kaip veikia jų DI sistemos, atliktų vidaus auditus ir, jei reikia, koreguotų veiklos kryptį, atsižvelgdamos į kintančias sąlygas bei naujus reikalavimus.

Viena iš pagrindinių **ISO/IEC 42001** savybių – tai, kad šis standartas leidžia organizacijoms nuosekliai valdyti visą dirbtinio intelekto sprendimų gyvavimo ciklą, kartu stiprinant pasitikėjimą šiomis technologijomis. Tai tampa itin svarbu tais atvejais, kai dirbtinis intelektas taikomas jautriose srityse,

[92] ISO/IEC 42001 (2023) Information technology - Artificial intelligence - Management system https://www.iso.org/standard/81230.html

tokiose kaip sveikatos apsauga ar teisėsaugos institucijų veikla. Naudojant šį standartą, DI sprendimų diegimas tampa ne tik skaidresnis ir atsakingesnis, bet ir grindžiamas išankstiniu poveikio vertinimu, kuris padeda iš anksto įžvelgti galimas rizikas ir numatyti jų valdymo priemones.

Svarbu ir tai, kad ISO/IEC 42001 lengvai derinamas su kitais valdymo standartais, pavyzdžiui, ISO 27001 ar ISO 31000, todėl jis gali būti natūraliai integruotas į bendrą organizacijos valdymo sistemą. Iš esmės, šis dokumentas padeda ne tik tvarkytis su technologiniais iššūkiais, bet ir suteikia metodinį pagrindą spręsti teisinius, etinius bei reputacijos klausimus, kylančius diegiant DI į praktinę veiklą.

3.3.1.4 ISO 23894 Dirbtinio intelekto rizikų valdymo gairės

Skirtingai nei ISO/IEC 42001, kuris apibrėžia bendrą dirbtinio intelekto valdymo sistemą organizacijų mastu, ISO/IEC 23894 susitelkia išskirtinai į su DI susijusių rizikų valdymą[93]. Šio standarto esmė – suteikti aiškias gaires, kaip organizacijos turėtų nustatyti, įvertinti, stebėti ir kontroliuoti rizikas, kylančias taikant DI technologijas. Jis aktualus viso sprendimų gyvavimo laikotarpiu – nuo koncepcijos formavimo iki įgyvendinimo ir vėlesnės sistemos priežiūros.

Šios rekomendacijos stipriai siejamos su ISO 31000 standartu, kuris nustato universalius rizikos valdymo principus, tačiau ISO/IEC 23894 šiuos principus pritaiko būtent DI kontekste. Akcentuojama, kad su DI susijusios rizikos dažnai yra mažiau nuspėjamos, apima automatizuotą sprendimų priėmimą, o jų pasekmės gali turėti reikšmingą įtaką žmogaus teisėms ar duomenų apsaugai. Todėl jų valdymui būtina nuosekliai planuota, dokumentuota ir su organizacijos veikla integruota prieiga.

Pažymėtina, kad ISO/IEC 23894 nėra sertifikuojamas – organizacijos negali oficialiai patvirtinti atitikties šiam standartui per sertifikavimo procesą. Vis dėlto, jo taikymas laikomas gerosios praktikos pavyzdžiu ir gali būti naudingas tiek kaip savarankiškas metodas, tiek kaip papildymas kitiems standartams, tokiems kaip ISO/IEC 42001 ar ISO/IEC 27001.

Standartas apima šiuos pagrindinius rizikų valdymo etapus:

- **Rizikos identifikavimas** apima galimų grėsmių, kylančių dirbtinio intelekto taikymo metu, nustatymą.
- **Rizikos analizė ir vertinimas** – nustatoma tikimybė, kada rizika realizuosis, ir koks gali būti jos poveikis organizacijos veiklai bei suinteresuotosioms šalims.
- **Rizikos mažinimo priemonių planavimas** – numatomos techninės, organizacinės ir teisinės priemonės, kurios padėtų sumažinti rizikos pasekmes ir jos pasireiškimo tikimybę.

[93] ISO/IEC 23894:202 Artificial Inteligence Guidance on risk management (žiūrėta internetu 2025-03-17 https://www.iso.org/standard/77304.html

- **Stebėsena ir peržiūra** – nuolatinė rizikų stebėsena apima vertinimą, kaip veikia taikomos priemonės, atsižvelgiant į technologinius pokyčius ir teisės aktų kaitą.

ISO/IEC 23894 taip pat akcentuoja, kad veiksmingas rizikų valdymas turi remtis skaidrumu ir nuosekliu procesų dokumentavimu. Organizacijos turėtų aiškiai registruoti priimtus sprendimus, taikytas priemones bei rizikų stebėsenos rezultatus. Toks metodinis požiūris ne tik padeda efektyviau valdyti grėsmes, bet ir sustiprina organizacijos atsakomybę bei išorės pasitikėjimą.

Taip pat standarte pabrėžiama, jog į rizikos vertinimo procesus verta įtraukti suinteresuotąsias šalis, ypač tada, kai dirbtinio intelekto sprendimai gali turėti apčiuopiamą poveikį vartotojams ar plačiajai visuomenei.

Standartų palyginimas

Kadangi dirbtinio intelekto rizikų valdymas dažnai glaudžiai siejasi su kitomis organizacinio valdymo sritimis – tokios kaip informacijos sauga, bendrasis rizikos valdymas ar sisteminė veiklos priežiūra – svarbu palyginti aptartus ISO standartus, išskiriant jų paskirtį bei taikymo kontekstą.

4 Lentelė. ISO/IEC Standartų palyginimai

Standartas	Paskirtis	Sertifikavimas	Pritaikymo sritis
ISO/IEC 42001	DI valdymo sistemos diegimo ir priežiūros standartas	Taip	Organizacijos, kurios kuria / naudoja DI
ISO/IEC 23894	DI rizikų identifikavimo, analizės ir valdymo gairės	Ne	Bet kokia organizacija, taikanti DI
ISO/IEC 27001	Informacijos saugos valdymo sistemos standartas	Taip	Visos organizacijos
ISO 31000	Bendrieji rizikų valdymo principai ir gairės	Ne	Visos organizacijos, visų rūšių rizikos

Šaltinis: parengta autorės, remiantis ISO/IEC 42001, ISO/IEC 23894, ISO/IEC 27001, ISO/IEC 31000

3.3.2. NIST standartų vaidmuo dirbtinio intelekto rizikų valdyme JAV

Nacionalinis standartų ir technologijų institutas [94] (National Institute of Standards and Technology, NIST) – tai JAV federalinė agentūra, kurios pagrindinė funkcija – rengti techninius standartus ir gaires, užtikrinančias informacinių technologijų, įskaitant dirbtinį intelektą (DI), patikimumą, saugumą bei veiksmingumą. Nors dauguma NIST dokumentų turi rekomendacinį pobūdį,

[94] National Institute of Standards and Technology https://www.nist.gov/

kai kurie jų, tokie kaip NIST SP 800 serijos gairės, parengtos remiantis Federaliniu informacinių sistemų saugumo valdymo aktu (FISMA[95]), yra privalomi visoms JAV federalinėms institucijoms.

Be to, privačios organizacijos, siekiančios bendradarbiauti su JAV federaliniu sektoriumi, taip pat turi vadovautis tam tikrais standartais – pavyzdžiui, taikyti NIST SP 800–53 reikalavimus, susijusius su informacinių sistemų saugumu. Dėl šios priežasties net ir neprivalomi NIST dokumentai, ypač susiję su dirbtiniu intelektu, dažnai tampa plačiai pripažįstami ir faktiškai privalomi tiek JAV, tiek tarptautiniu mastu. Organizacijoms, norinčioms teikti paslaugas federalinėms institucijoms ar veikti reguliuojamose srityse, tenka taikyti NIST principus net neturint kitos alternatyvos.

NIST gairės išsiskiria praktiniu pritaikomumu ir lankstumu, todėl jos dažnai tampa pagrindu tiek viešojo, tiek privataus sektoriaus politikai – ypač kibernetinio saugumo, rizikų valdymo ir duomenų apsaugos srityse. Šiame kontekste vis daugiau reikšmės įgyja 2023 m. paskelbtas dokumentas – Dirbtinio intelekto rizikų valdymo sistema (AI Risk Management Framework, AI RMF), kuris tampa plačiai taikomu metodiniu pagrindu DI rizikų valdymui, įtvirtinant aiškius ir praktiškai pritaikomus veiklos principus.

3.3.2.1 AI Risk Management Framework, AI RMF

AI Risk Management Framework[96](AI RMF) – tai 2023 m. Nacionalinio standartų ir technologijų instituto (NIST) parengta metodinė sistema, kuri skirta padėti organizacijoms efektyviai valdyti rizikas, susijusias su dirbtinio intelekto sprendimais. Šio dokumento esmė – skatinti atsakingą, pagrįstą ir socialiai orientuotą DI technologijų kūrimą bei jų naudojimą, mažinant galimus neigiamus padarinius tiek organizacijų viduje, tiek plačiajai visuomenei.

AI RMF neturi privalomo ar sertifikuojamo statuso – tai lanksti gairių sistema, skirta pritaikymui įvairiose organizacijose, nepriklausomai nuo jų dydžio ar veiklos srities, taip pat nuo to, ar jos pačios kuria, ar tik naudoja DI technologijas. Dokumente laikomasi principo, kad rizikų valdymas turi būti įtvirtintas visoje organizacijos veikloje kaip nuoseklus, tęstinis procesas, o ne laikinas ar izoliuotas veiksmas.

AI RMF struktūra paremta keturiomis pagrindinėmis funkcinėmis sritimis, kurios padeda organizacijoms sistemingai valdyti su dirbtiniu intelektu susijusias rizikas:

1. **Valdyti** *(angl. Govern)* – organizacija suformuoja dirbtinio intelekto rizikų valdymo politiką, apibrėžia atsakomybes ir įtraukia suinteresuotąsias šalis.

[95] Federal Information Security Modernization Act (2014) https://www.cisa.gov/topics/cyber-threats-and-advisories/federal-information-security-modernization-act

[96] AI Risk Management Framework 2024 https://www.nist.gov/itl/ai-risk-management-framework

2. **Nustatyti** *(angl. Map)* – identifikuojama DI sistemos taikymo sritis, potencialūs rizikos šaltiniai ir jų galimas poveikis organizacijos veiklai ar naudotojams.
3. **Matuoti** *(angl. Measure)* – naudojami aiškūs rodikliai ir vertinimo metodikos, leidžiančios įvertinti DI sprendimo veikimą ir jam būdingas rizikas.
4. **Valdyti rizikas** *(angl. Manage)* – taikomos konkrečios rizikų mažinimo priemonės, vertinamas jų veiksmingumas, o procesai periodiškai atnaujinami atsižvelgiant į pokyčius.

Siekiant palengvinti šių principų taikymą praktikoje, NIST papildomai parengė dokumentą – AI *RMF Playbook*, kuriame pateikiamos detalios gairės, kaip praktiškai taikyti AI RMF principus organizacijų veikloje. Ši metodinė priemonė atlieka panašią funkciją kaip ISO 27002, kuris padeda įgyvendinti ISO 27001 reikalavimus. AI *RMF Playbook* leidžia organizacijoms pasirinkti jų specifikai tinkamiausias priemones, siekiant efektyviai valdyti DI keliamas rizikas.

AI RMF sistema remiasi pagrindinėmis vertybinėmis nuostatomis – **patikimumu**, **skaidrumu**, **saugumu**, **teisingumu** ir **atsakomybe**. Šios vertybės užtikrina, kad dirbtinio intelekto technologijos būtų kuriamos ir taikomos ne tik atitinkant techninius standartus, bet ir laikantis etinių bei teisinių principų. Toks vertybinis pagrindas padeda stiprinti visuomenės pasitikėjimą DI sprendimais ir skatina jų atsakingą diegimą įvairiose srityse.

Apibendrinant galima teigti, kad AI RMF sudaro lankstų ir sistemingą pagrindą organizacijoms, kurios siekia valdyti DI rizikas, užtikrinti atsakingą technologijų taikymą bei prisidėti prie saugios ir patikimos skaitmeninių inovacijų plėtros. Tai vienas pirmųjų tarptautiniu mastu pripažintų dokumentų, skirtų būtent praktiniam dirbtinio intelekto rizikų valdymui.

3.3.3. Organizacinės DI rizikų valdymo praktikos: GRC modelio integravimas

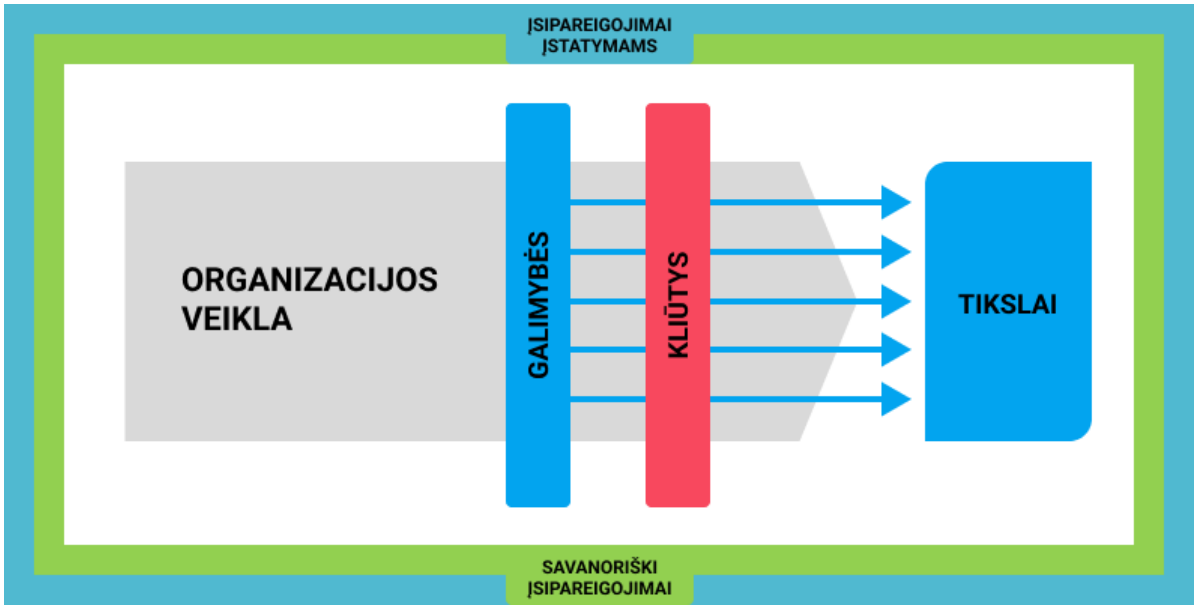
Analizuojant dirbtinio intelekto (DI) rizikų valdymo priemones, matyti, kad organizacijos šiuo metu susiduria su itin plačia standartų, gairių ir teisinių reikalavimų įvairove. ISO/IEC 42001, ISO/IEC 23894, NIST AI RMF, Europos Sąjungos AI aktas – kiekvienas iš jų pateikia savitą principų, procesų ir atsakomybės modelių visumą. Kai kurie standartai, tokie kaip ISO/IEC 42001 ar ISO/IEC 27001, yra sertifikuojami, todėl jų įgyvendinimas dažnai pareikalauja reikšmingų finansinių ir žmogiškųjų resursų. Tuo tarpu tokie dokumentai kaip **NIST AI RMF** siūlo lankstesnį, rekomendacinį požiūrį, kuris gali būti ypač naudingas, tačiau neretai pasirodo pernelyg abstraktus ir sudėtingai pritaikomas praktikoje, ypač mažesnėms ar mažiau specializuotoms organizacijoms. Taip pat, kai organizacija vienu metu taiko kelis skirtingus standartus, kyla rizika, kad tam tikri procesai ims dubliuotis, trūks koordinacijos tarp padalinių, o pats rizikų valdymas taps padrikas ir mažiau efektyvus.

GRC (***angl. Governance*, *Risk* and *Compliance***) ***(liet.k valdymas, rizikų valdymas ir atitiktis)*** koncepcija iškilo kaip atsakas į vis sudėtingesnį standartų, teisinių reikalavimų ir vidinių taisyklių

sąveikos lauką organizacijose. 2002 m. ją suformulavo Atviros atitikties ir etikos grupė (Open Compliance and Ethics Group, OCEG), pasiūliusi vieningą sistemą, padedančią ne tik efektyviau valdyti rizikas ar užtikrinti teisinę atitiktį, bet ir išlaikyti strateginį organizacijos kryptingumą Šis požiūris buvo kurtas kaip struktūrinis karkasas, sujungiantis skirtingus valdymo metodus – ISO, NIST, AI Act ar kitus – į vientisą, lengvai pritaikomą struktūrą.

Naujausia šios koncepcijos versija – GRC Capability Model 3.5[97] (2024) - tęsdama integracinį principą, ji aiškiai nurodo, kaip organizacijos turėtų planuoti, įgyvendinti ir vertinti valdymo, rizikų bei atitikties funkcijas. Dokumentas remiasi plačia ekspertų patirtimi ir įvairiapusėmis praktikomis iš skirtingų sektorių, todėl pateikia universaliai pritaikomą, tačiau praktiškai pagrįstą požiūrį, atitinkantį realius organizacijų poreikius. Nors pats modelis nėra specifiškai orientuotas į DI valdymą, jo lanksti, modulinė struktūra suteikia galimybę tikslingai įtraukti DI keliamas rizikas į bendrą GRC sistemą – nuo etinių iššūkių iki reguliacinės atitikties.

6 pav. GRC modelio principai



Šaltinis: Sudaryta autorės pagal OCEG Red Book 3.5 (2024)

Modelis ne tik aiškiai apibrėžia, kaip planuoti, koordinuoti ir vertinti valdymo, rizikos bei atitikties procesus, bet ir suteikia reikšmingų praktinių patobulinimų. Atnaujintoje versijoje:

- Suvienodinama terminologija tarp skirtingų organizacinių disciplinų.
- Apibrėžiami bendrieji komponentai ir elementai, būtini efektyviai GRC sistemos veiklai.
- Nustatomi vieningi informacijos valdymo reikalavimai, užtikrinant nuoseklų ir koordinuotą informacijos srautą.

[97] Scott Mitchell GRC Capability Model™ 3.5 (OCEG™ Red Book)https://www.oceg.org/grc-capability-model-red-book/

- Standartizuojamos tokios praktikos kaip politikų formavimas, darbuotojų mokymai, atitikties stebėsena.
- Įvardijami komunikacijos poreikiai, apimantys visus suinteresuotus dalyvius – nuo operatyvinio lygmens iki strateginių sprendimų priėmėjų.

Pagrindiniai GRC modelio komponentai apima:

- **Valdymą** *(angl. governance)* – tai apima organizacinės struktūros formavimą, aiškų atsakomybės paskirstymą bei sprendimų priėmimo mechanizmus, kurie padeda užtikrinti, kad dirbtinio intelekto iniciatyvos būtų tiesiogiai susietos su bendrais organizacijos strateginiais tikslais ir vertybinėmis nuostatomis. Esminis aspektas čia – ne tik koordinuoti techninius sprendimus, bet ir vertinti jų platesnį poveikį, aiškiai apibrėžiant atsakomybę tarp skirtingų padalinių ir suinteresuotųjų grupių. Toks integruotas požiūris padeda išvengti padriko DI diegimo ir sudaro sąlygas kryptingam, strategiškai pagrįstam jo taikymui.
- **Rizikų vertinimą** *(angl. risk assessment)* – tai nuoseklus procesas, skirtas atpažinti, analizuoti ir valdyti su DI susijusias grėsmes. Jos gali būti įvairios: nuo techninių (pvz., šališki algoritmai, nepatikimi duomenys) iki teisinių ar reputacinių padarinių. GRC modelyje akcentuojama būtinybė ne tik nustatyti rizikas, bet ir taikyti lankstų, nuolat atnaujinamą valdymą, reaguojantį į technologinius, teisinius bei veiklos pokyčius.
- **Atitiktį** *(angl. compliance)* reiškia organizacijos gebėjimą laikytis visų galiojančių teisinių, etinių ir vidinių reikalavimų, susijusių su dirbtinio intelekto kūrimu bei taikymu. Tai apima ne tik tokius teisės aktus kaip **AI Act ar BDAR**, bet ir platesnius atsakomybės klausimus, tokius kaip diskriminacijos prevencija ar naudotojų informavimas apie jų duomenų naudojimą. Veiksminga atitikties sistema nėra vien formali procedūrų visuma – ji remiasi organizaciniu sąmoningumu, kuriam būtini nuolatiniai darbuotojų mokymai, reguliarūs auditai, aiškios vidaus taisyklės ir skaidri komunikacija tiek viduje, tiek išorėje.

Integravus šiuos tris komponentus, GRC modelis sudaro vientisą sistemą, leidžiančią organizacijoms ne tik atitikti DI taikymui keliamus reikalavimus, bet ir formuoti tvarią, skaidrią bei strategiškai pagrįstą dirbtinio intelekto valdymo praktiką.

Organizacijų gebėjimas praktiškai taikyti DI rizikų valdymo standartus lemia ne tik atitiktį teisės aktams, bet ir visuomenės pasitikėjimą technologijomis. Vis dėlto neretai tarp formaliai priimtų dokumentų ir realios praktikos atsiranda spraga. Šį atotrūkį gali padėti užpildyti integruoti valdymo modeliai, tokie kaip GRC, kurie apjungia rizikų valdymą, atitiktį bei strateginį sprendimų priėmimą į vieną nuoseklią ir tarpusavyje suderintą sistemą. Tačiau net ir jie reikalauja brandžios organizacinės kultūros, todėl svarbiausias iššūkis lieka ne pasirinktose priemonėse, o gebėjime jas įgyvendinti atsakingai ir tvariai.

3.3.4. Teorinės analizės refleksija

Atlikta literatūros analizė atskleidė sudėtingą ir nuolat besikeičiančią DI rizikų valdymo aplinką, kurioje susikerta trys esminės dimensijos – technologinis progresas, teisinė reguliacija ir organizacinė praktika. Nors pažangūs DI modeliai jau daro reikšmingą poveikį sprendimų priėmimui įvairiuose sektoriuose – nuo viešojo administravimo iki verslo analitikos – teisėkūra ir instituciniai mechanizmai kol kas nespėja reaguoti pakankamai greitai. Dėl to atsiranda vadinamosios *„pilkosios zonos“*, kuriose tiek reguliavimo, tiek atsakomybės ribos lieka neaiškios.

Vienas iš akivaizdžiausių pavyzdžių – dirbtinio intelekto sukuriamo turinio autorystės klausimas, kuris iki šiol išlieka teisiškai neapibrėžtas, o praktika dar tik formuojasi. Panašus teisinis neaiškumas išlieka ir sprendžiant atsakomybės klausimus, kai autonominiai sprendimai priimami DI sistemų be tiesioginio žmogaus įsikišimo. Tokios situacijos kelia rizikų ne tik žmogaus teisių apsaugai, bet ir gali rimtai pakenkti pačių organizacijų reputacijai. Visa tai atskleidžia, kad teisės sistema dažnai nespėja žengti koja kojon su sparčia technologine pažanga ir dažniausiai reaguoja jau post factum (liet. „po fakto“) – tik tuomet, kai pasekmės tampa akivaizdžios arba kyla didesnis visuomenės spaudimas.

Tuo pat metu matoma reikšminga pažanga standartizacijos srityje – tokios institucijos kaip ISO ir NIST vis dažniau teikia išsamius metodinius modelius, skirtus dirbtinio intelekto rizikų identifikavimui, vertinimui ir valdymui. Vis dėlto, jų praktinis taikymas išlieka sudėtingas. Tam nepakanka vien techninių žinių – būtina turėti brandžią organizacinę struktūrą, pakankamai resursų ir aiškiai paskirstytas atsakomybes.

Pavyzdžiui, tokie standartai kaip ISO 42001 ar NIST AI RMF sudaro tvirtą teorinį pagrindą, tačiau jų įgyvendinimas reikalauja individualaus pritaikymo. Šių dokumentų negalima mechaniškai perkelti į bet kurią organizaciją, neatsižvelgiant į jos veiklos pobūdį, mastą ar konkretaus sektoriaus ypatumus.

Atlikta analizė rodo, kad nors GRC modelis iš pradžių nebuvo sukurtas būtent dirbtinio intelekto sričiai, šiandien jis tampa viena universaliausių ir praktiškai pritaikomų metodikų. Jo pagrindinis pranašumas – gebėjimas integruoti įvairius reguliavimo reikalavimus bei standartus į vieningą sistemą, kuri leidžia organizacijoms kryptingai valdyti rizikas, aiškiai paskirstyti atsakomybę ir suderinti valdymo, atitikties bei technologijų diegimo procesus. Vis dėlto ir šis modelis reikalauja aktyvios adaptacijos bei stipraus žmogiškojo kapitalo.

Apibendrinant galima teigti, kad nors teoriniai modeliai ir standartai egzistuoja, jų vien tik buvimas negarantuoja atsakingo dirbtinio intelekto taikymo – būtina nuosekli, kontekstui jautri jų integracija organizacijų praktikoje. Todėl tolimesniame darbo etape numatytas kokybinis tyrimas tampa ypač svarbus – jis padės suprasti, kaip organizacijos praktiškai vertina ir valdo su DI susijusias rizikas, kokias priemones naudoja bei kiek yra pasirengusios realiai įgyvendinti tarptautinius standartus ar jų adaptuotus modelius.

4. DIRBTINIO INTELEKTO RIZIKŲ VALDYMAS ORGANIZACIJOSE: TYRIMO METODOLOGIJA

Šiame skyriuje pateikiama empirinės dalies metodologija, integruojanti teorinę analizę su praktiniu tyrimu. Literatūros apžvalga parodė, kad nors egzistuoja tarptautiniai dirbtinio intelekto rizikų valdymo modeliai, jų perkėlimas į organizacijų praktiką nėra automatinis procesas – tam būtinas techninių žinių pagrindas, strateginis brandumas bei pakankami ištekliai.

Todėl tyrimo objektas empirinėje dalyje buvo tikslinamas – nuo bendro dirbtinio intelekto rizikų valdymo analizės pereita prie to, kaip šios rizikos yra atpažįstamos, vertinamos ir valdomos Lietuvos organizacijose. Toliau pristatoma pasirinkta tyrimo strategija, naudoti duomenų rinkimo ir analizės metodai, ekspertų atrankos kriterijai bei tyrimo patikimumą užtikrinantys veiksniai.

4.1 Tyrimo objektas, tikslas ir uždaviniai

Tyrimo objektas – dirbtinio intelekto rizikų valdymo procesai Lietuvos organizacijose.

Tyrimo tikslas – išanalizuoti, kaip šiose organizacijose praktiškai taikomi sprendimai, susiję su DI rizikų valdymu.

Tyrimo uždaviniai:

- Išanalizuoti mokslinę literatūrą, nagrinėjančią dirbtinio intelekto keliamas rizikas ir jų valdymo metodus.
- Įvertinti tarptautinius DI rizikų valdymo modelius (pvz., ISO, NIST AI RMF, GRC) bei jų pritaikomumą organizacijų kontekste.
- Atlikti ekspertinį tyrimą, siekiant suprasti, kaip Lietuvos organizacijos identifikuoja, vertina ir sprendžia su DI susijusias rizikas.
- Pasiūlyti praktines rekomendacijas ar modelio koncepciją, kuri galėtų padėti organizacijoms veiksmingiau valdyti DI rizikas.

4.2 Tyrimo strategija ir metodų pasirinkimas

Siekiant išanalizuoti, kaip Lietuvos organizacijos atpažįsta, vertina ir valdo su dirbtiniu intelektu susijusias rizikas, šiame tyrime **bus taikoma kokybinė tyrimo strategija**. Šis pasirinkimas pagrįstas tuo, kad kokybiniai tyrimai leidžia gilintis į sudėtingus reiškinius ir jų kontekstus, atskleisti prasmę, o ne vien kiekybiškai įvertinti duomenis. Pasak V. Žydžiūnaitės ir S. Sabaliausko (2017), kokybinis

tyrimas ypač tinkamas tuomet, kai siekiama suprasti subjektyvias patirtis bei socialinių reiškinių interpretacijas dalyvių akimis[98].

Kadangi DI rizikų valdymas dažnai nėra formalizuotas ar standartizuotas procesas ir priklauso nuo konkretaus organizacijos brandos lygio, sektoriaus ypatumų bei naudojamų technologijų, pasirinktas **pusiau struktūruoto interviu metodas**. Šis metodas leidžia išlaikyti aiškią tyrimo kryptį, tačiau kartu suteikia ir būtiną lankstumą. Tai sudaro galimybes reaguoti į pokalbio metu iškylančias temas bei gilintis į netikėtai atsiveriančius aspektus. Kaip pažymi K. Kardelis (2019), pusiau struktūruoti interviu nėra skirti vien faktiniams duomenims rinkti – jie taip pat padeda atskleisti vertybines nuostatas ir sprendimus, kurie kyla iš pašnekovų praktinės patirties[99].

Tyrimo metodo ribotumai:

Nepaisant ekspertinių interviu privalumų – galimybės gilintis į tiriamąją temą ir gauti vertingų kokybinių įžvalgų – būtina įvardinti ir keletą reikšmingų metodo apribojimų:

1. Subjektyvumas – ekspertų atsakymai dažnai remiasi asmenine patirtimi ir vertybėmis, todėl gauti rezultatai gali būti neapibrėžti ar šališki.
2. Ribotas taiklumas – net ir sudėtingų technologijų, tokių kaip dirbtinis intelektas, kontekste ekspertinės įžvalgos dažnai grindžiamos nuojauta, o ne objektyviai patvirtintais duomenimis
3. Maža imtis – dėl riboto respondentų skaičiaus gautos įžvalgos nėra statistiškai apibendrinamos visai organizacijų populiacijai.
4. Interpretacija – tyrimo rezultatai priklauso nuo tyrėjo gebėjimo suprasti, interpretuoti ir kategorizuoti atsakymus,
5. Konteksto priklausomumas – gautos įžvalgos gali būti būdingos ik konkrečiam laikotarpiui ar organizaciniam kontekstui.

Dėl šių priežasčių tyrimo duomenys vertinami kaip indikatyvūs, o ne kaip galutiniais sprendimai, tinkami visoms organizacijoms ar sektoriams.

Tyrimas bus grindžiamas **skerspjūvio strategija**, kuri, kaip nurodo K. Kardelis (2019), leidžia vienu metu rinkti duomenis iš įvairių šaltinių ir įvertinti tiriamo reiškinio būklę konkrečiu laiko momentu. Šis metodas ypač tinkamas situacijoms, kai siekiama suprasti, kokios praktikos ir požiūriai vyrauja organizacijose tam tikru laikotarpiu, o ne analizuoti jų pokyčius ilguoju laikotarpiu.

Tyrimo dalyviai bus atrenkami **pasitelkiant ekspertinę (tikslinę) atranką** – dalyvauti kviečiami specialistai, turintys praktinės patirties dirbant su dirbtinio intelekto diegimu, plėtra ar valdymu. Tokia

[98] Žydžiūnaitė, V., & Sabaliauskas, S. (2017). Kokybiniai tyrimai: Principai ir metodai. Vilnius: Vaga.
[99] Kardelis, K. (2007). Mokslinių tyrimų metodologija ir metodai. Šiauliai: Lucilijus.

atrankos forma padės užtikrinti, kad surinkti duomenys būtų empiriškai pagrįsti ir atspindėtų realią organizacijų praktiką.

Empirinė analizė bus grindžiama teorinėje dalyje aptartais tarptautiniais modeliais: ISO 31000, 42001, NIST AI RMF bei GRC principais. Šie modeliai bus pasitelkiami ne tik formuojant interviu struktūrą, bet ir kaip teorinis pagrindas surinktų duomenų interpretavimui. Toks teorinių nuostatų ir praktinių įžvalgų derinimas padės išlaikyti tyrimo nuoseklumą bei suteiks galimybę įvertinti, kiek organizacijose taikoma praktika atitinka tarptautinius dirbtinio intelekto rizikų valdymo principus.

4.3 Interviu struktūra ir tyrimo imtis

Empirinėje tyrimo dalyje bus taikomas pusiau struktūruoto interviu metodas, kuris leidžia derinti iš anksto parengtą klausimų struktūrą su galimybe lanksčiai prisitaikyti prie kiekvieno pašnekovo patirties. Interviu klausimai formuluojami remiantis teorinėje darbo dalyje aptartais DI rizikų valdymo aspektais bei tarptautiniais modeliais – ISO 31000, ISO/IEC 42001, NIST AI RMF ir GRC principais. Tokiu būdu siekiama atskleisti, kaip šios tarptautinės gairės pritaikomos organizacijų kasdienėje praktikoje.

Interviu klausimai suskirstyti į kelis teminius blokus:

- **Organizacinis kontekstas ir pašnekovo vaidmuo** – siekta suprasti organizacijos veiklos sritį, taikomus DI sprendimus ir pašnekovo atsakomybės ribas.
- **DI rizikų suvokimas ir identifikavimas** – kokios rizikos matomos diegiant DI technologijas ir kaip jos atpažįstamos.
- **Rizikų vertinimo ir valdymo praktikos** – ar laikomasi reguliavimo, taikomi standartai, atsižvelgiama į etinius veiksnius.
- **Atitiktis standartams, etiniai ir teisiniai aspektai** – ar organizacija taiko kokius nors standartus, laikosi reguliavimo, atsižvelgia į etinius veiksnius.
- **Iššūkiai ir tobulinimo poreikiai** – kas sudėtingiausia valdant DI rizikas, ko trūksta, kaip pašnekovai vertina integruoto rizikų valdymo modelio idėją.

Bendras interviu klausimų skaičius sieks 10–12, tačiau galutinis jų skaičius priklausys nuo pokalbio eigos ir atsirandančių temų. Interviu trukmė planuojama nuo 30 iki 60 minučių, priklausomai nuo dalyvio patirties ir pasirengimo dalyvauti pokalbyje.

Tyrimo imtį sudarys šeši ekspertai, atstovaujantys skirtingas sritis: finansines technologijas, klientų valdymą, asmens tapatybės nustatymą, IT sprendimus ir inovacijas. Toks dalyvių spektras suteiks galimybę pažvelgti į DI rizikų valdymo praktiką iš įvairių perspektyvų ir įvertinti, kaip šie procesai taikomi skirtinguose organizaciniuose kontekstuose.

Imtis formuojama taikant tikslinę (ekspertinę) atranką, kuri papildoma *„sniego gniūžtės“* metodu – tai reiškia, kad pradžioje atrinkti dalyviai rekomenduoja kitus potencialius tyrimo dalyvius, atitinkančius tyrimo kriterijus. Šis metodas ypač naudingas siekiant pasiekti uždaras profesionalų bendruomenes ir gauti giluminę, patirties pagrindu grindžiamą informaciją.

Dalyviai atrenkami remiantis šiais kriterijais:

- Turi tiesioginės patirties diegiant, prižiūrint ar valdant DI sprendimus organizacijoje;
- Yra dalyvavę DI projektuose arba rizikų valdymo procesuose;
- Atstovaujama organizacija naudoja bent vieną DI sistemą ar sprendimą.

Pasirinktas šešių ekspertų skaičius grindžiamas duomenų *sotumo* principu – kai tolesni interviu nebeatneša naujos esminės informacijos (Kardelis, 2019). Kaip teigia V. Žydžiūnaitė ir S. Sabaliauskas (2017), kokybinio tyrimo tikslas nėra kiekybinis apibendrinimas, o gilus analizuojamo reiškinio supratimas, kurį galima pasiekti pasitelkus informatyviausius atvejus. B. Bitinas, L. Rupšienė ir kt. (2008) taip pat pažymi, kad 5–9 ekspertų grupė gali būti pakankama reikšmingoms įžvalgoms formuoti, jei atranka atlikta tikslingai ir apgalvotai. Šiame tyrime šešių dalyvių interviu laikomi pakankamu pagrindu siekiant užsibrėžtų tikslų.

Visi interviu bus vykdomi nuotoliniu būdu. Gavus dalyvių sutikimą, pokalbiai bus įrašyti ir transkribuoti tolimesnei kokybinei turinio analizei. Bus laikomasi visų akademinės etikos reikalavimų – užtikrinamas dalyvių anonimiškumas, o surinkti duomenys naudojami išskirtinai moksliniais tikslais [100].

Taip pat, atsižvelgiant į taikomus reikalavimus ir laikantis aukščiausių tyrimų etikos standartų, buvo remtasi K. Kardelio (2019) bei V. Žydžiūnaitės ir S. Sabaliausko (2017) rekomendacijomis dėl dalyvių teisės į privatumą. Todėl nuspręsta kiekvienam ekspertui priskirti unikalų kodą. Tokia praktika padėjo užtikrinti respondentų anonimiškumą, išlaikyti tyrimo duomenų objektyvumą ir patikimumą, kartu apsaugant tiek ekspertų profesinę reputaciją, tiek jų atstovaujamų organizacijų interesus.

5 lentelė. Ekspertų unikalus kodai ir informacija apie interviu

Eksperto unikalusis kodas	Interviu trukmė (min)	Interviu data	Apklausos tipas
E1	**60**	**2025–04–07**	Žodžiu
E2	**45**	**2025–04–02**	Raštu
E3	**60**	**2025–04–11**	Žodžiu
E4	**30**	**2025– 04– 08**	Žodžiu
E5	**25**	**2025– 04– 14**	Raštu
E6	**22**	**2025– 04– 15**	Žodžiu

Šaltinis: Sudaryta autorės

[100] Bitinas, B., Rupšienė, L., & Žydžiūnaitė, V. (2008). Kokybinių tyrimų metodologija. Klaipėda: S. Jokužio leidykla-spaustuvė.

Tyrimo klausimyno sudarymas. Šiame tyrime bus taikomas pusiau struktūruotas interviu klausimynas, sudarytas iš 12 teminių klausimų (žr. 1 priedą). Klausimų rengimas rėmėsi mokslinės literatūros, teisės aktų ir praktinės rizikų valdymo patirties analize, siekiant atskleisti pagrindinius DI rizikų identifikavimo, vertinimo bei valdymo aspektus. Klausimai suformuluoti taip, kad skatintų dalyvius neapsiriboti vien faktų pateikimu, bet pasidalyti ir savo asmeninėmis įžvalgomis bei patirtimi. Jei pokalbio metu iškiltų poreikis, numatyta galimybė užduoti papildomų klausimų – tai suteikia lankstumo ir leidžia giliau panagrinėti temas, kurios natūraliai išryškėja pokalbio eigoje. Toks metodinis požiūris padeda geriau suprasti realią organizacijų praktiką ir kontekstą, kuriame priimami sprendimai. Surinkti duomenys bus analizuojami teminiu principu ir taps pagrindu praktinėms įžvalgoms bei rekomendacijoms, skirtoms veiksmingesniam DI rizikų valdymui organizacijų lygmeniu.

4.4 Duomenų analizės metodas

Interviu metu surinkti duomenys bus analizuojami taikant kokybinės turinio analizės metodą. Šis metodas suteikia galimybę struktūruotai išskirti pagrindines prasmes, reikšmines temas bei tendencijas, kurios atsiskleidžia tyrimo dalyvių atsakymuose. Kaip pažymi B. Bitinas, L. Rupšienė ir V. Žydžiūnaitė (2008), turinio analizė ypač tinkama socialinių ir organizacinių reiškinių tyrimui, kai siekiama suprasti, kaip tam tikros praktikos pasireiškia konkrečiuose kontekstuose ir aplinkose.

Duomenų analizė bus vykdoma keliais nuosekliais etapais:

- **Transkribavimas.** Interviu įrašai bus pažodžiui perrašyti, išsaugant kalbėsenos ypatumus, kurie gali padėti geriau suprasti pašnekovų intencijas ar emocinį krūvį.
- **Pradinė peržiūra.** Transkribuoti tekstai bus peržvelgti siekiant išskirti dažniausiai pasikartojančias sąvokas, temas, naratyvus ar požiūrius.
- **Teminių kategorijų kūrimas.** Pagal pradinę analizę bus suformuotos pagrindinės kategorijos (pvz. „rizikų identifikavimas“, „atsakomybės paskirstymas“, „teisinis neapibrėžtumas“, „kultūriniai barjerai“ ir pan.), kurios padės struktūruoti analizės eigą.
- **Kodavimas.** Kiekvienam pašnekovui bus priskirtas unikalus kodas (pvz. E1–E6), o jų atsakymai bus susieti su atitinkamomis temomis ir analizės kategorijomis.
- **Interpretacija.** Duomenys bus interpretuojami remiantis teorinėje dalyje aptartais modeliais (ISO 31000, ISO/IEC 42001, NIST AI RMF, GRC), siekiant įvertinti, kiek praktika atitinka siūlomas tarptautines nuostatas.

Analizė bus atliekama rankiniu būdu, taikant žymėjimo ir teminio išskyrimo metodus lentelėse. Toks būdas leidžia tyrėjui išlikti arti pirminių duomenų ir lanksčiai reaguoti į pokalbių metu išryškėjusius kontekstinius niuansus, galinčius turėti reikšmingą įtaką prasmių interpretavimui. Šio metodo tikslas nėra kiekybiškai nustatyti, kaip dažnai pasikartojo tam tikras požiūris, bet identifikuoti

pagrindines prasmines kryptis, kurios atskleidžia, kaip organizacijos suvokia ir sprendžia dirbtinio intelekto keliamas rizikas.

Toks analizės metodas padės atsakyti į pagrindinį tyrimo klausimą – kaip organizacijų viduje yra suvokiamos, vertinamos ir valdomos su dirbtiniu intelektu susijusios rizikos. Taip atskleidžiamos ne tik oficialios strategijos ar dokumentuoti procesai, bet ir realūs veiklos modeliai, sprendimų priėmimo logika bei vertybiniai pasirinkimai, formuojantys praktinį požiūrį į rizikų valdymą.

5. TYRIMO REZULTATAI IR ANALIZĖ

Analizuojant ekspertinių interviu metu gautus duomenis, jie buvo susisteminti taikant teminį išskyrimą ir interpretuoti remiantis teoriniais modeliais, išsamiai aptartais darbo antroje dalyje – ISO 31000, ISO/IEC 42001, NIST AI RMF bei GRC principais. Tyrimo struktūra rėmėsi iš anksto parengtu interviu klausimynu, kurio tikslas buvo ne tik įvertinti DI rizikų valdymo būklę Lietuvos organizacijose, bet ir nustatyti jų brandos lygį, dažniausiai pasitaikančius iššūkius bei identifikuoti esamas sistemines spragas.

Tyrimo dalyviai dalijosi ne tik faktine informacija apie taikomas praktikas, bet ir asmeninėmis įžvalgomis apie praktinius bei reguliacinius barjerus, kurie, jų nuomone, trukdo kryptingai ir sistemingai valdyti su dirbtiniu intelektu susijusias rizikas. Kai kurios organizacijos, ypač veikiančios pažangesniuose DI taikymo sektoriuose, nurodė jau turinčios tam tikras formalizuotas rizikų valdymo struktūras. Vis dėlto daugumai respondentų rizikų valdymas vis dar išlieka labiau intuityvus nei metodiškai apibrėžtas procesas.

Surinkti atsakymai leido išryškinti, kokios praktikos šiuo metu laikomos veiksmingomis, kokių priemonių organizacijos imasi, kad prisitaikytų prie nuolat kintančio teisinio reguliavimo ir tarptautinių standartų, bei kurios DI taikymo sritys joms kelia daugiausia neaiškumo ar iššūkių. Ekspertai buvo skatinami neapsiriboti tik esama patirtimi, bet ir pažvelgti į priekį – įvardyti, kaip jų manymu DI rizikų valdymas turėtų atrodyti idealiu atveju.

Apibendrinant, šiame skyriuje pateikti tyrimo duomenys padeda atskleisti tiek esamus DI rizikų valdymo spragos taškus, tiek galimus jų tobulinimo kelius. Gautos įžvalgos taip pat taps pagrindu formuojant praktines rekomendacijas, kurios gali padėti Lietuvos organizacijoms atsakingiau ir labiau struktūruotai integruoti dirbtinį intelektą į savo veiklos procesus.

5.1. Organizacinis kontekstas ir DI taikymo sritys.

Ekspertų įžvalgos apie dirbtinio intelekto (DI) diegimą organizacijose ir su tuo susijusias rizikas buvo apibendrintos remiantis tyrimo struktūra. Šio skyriaus tikslas – atskleisti, kaip įvairių sričių įmonės taiko dirbtinio intelekto sprendimus ir kokį vaidmenį ši technologija atlieka jų veikloje. Toks vertinimas leidžia ne tik geriau suprasti DI svarbą bei organizacijos brandos lygį, bet ir padeda identifikuoti spragas rizikų valdymo srityje.

Tyrime dalyvavo skirtingų sektorių atstovai – nuo finansinių paslaugų, klientų identifikavimo ir dokumentų analizės iki logistikos bei net kosmoso duomenų apdorojimo srityse veikiančių organizacijų.

Nepaisant šių skirtumų, DI visose organizacijose vis giliau įsitvirtina pagrindiniuose procesuose. Tapatybės atpažinimo sprendimus siūlančios įmonės atstovas E1 pažymėjo: „*Be DI mūsų verslas tiesiog*

nebūtų įmanomas – Dirbtinis intelektas mums būtinas tiek veido atpažinimui, tiek jo sulyginimui su dokumentuose esančiu atvaizdu. Taip pat naudojame DI nuskaityti dokumentų tekstą ir jį patikrinti su mūsų vidinę duomenų baze“.

Tuo tarpu **E2**, atstovaujantis pažangių kalbinių modelių kūrimo įmonei, išskyrė kitokį iššūkį – vertybinių ir etinių rizikų valdymą. Jų organizacijoje dirbtinis intelektas neapsiriboja atsakymų generavimu ar sprendimų siūlymu – jis veikia kaip realaus laiko sąveikos partneris, gebantis interpretuoti kontekstą ir prisitaikyti prie naudotojo elgsenos. Tačiau būtent toks dinamiškumas kelia specifinių grėsmių. „*Atliekant mokymo duomenų auditą pastebėjome, kad kai kurios socialinės grupės mūsų duomenyse buvo atstovaujamos nepakankamai, todėl atsirado rizika šališkiems atsakymams. Rizika nebuvo matoma iškart – ją identifikavome ne dėl išorinio incidento, o dėka vidinio etinio vertinimo*“, – pabrėžė **E2**. Kūrybiniuose modeliuose rizika neretai glūdi ne tiek techniniuose aspektuose, kiek pačiuose duomenyse – ypač ten, kur įsismelkia kultūriniai ar socialiniai šališkumai, dažnai nepastebimi iš pirmo žvilgsnio.

Kai kurios organizacijos DI diegimą pradėjo nuo paprastų automatizavimo užduočių, kurios ilgainiui virto platesniais sprendimais. E3, atstovaujantis tarptautinių pervedimų sektorių, pasakojo, kad dirbtinis intelektas iš pradžių veikė kaip pagalbininkas vadybininkams, o vėliau tapo svarbiu įrankiu klientų tapatybės tikrinimui: „*Iš pradžių integravome DI į savo vidinę dokumentaciją – jis tiesiog padėdavo vadybininkams atsakyti į klientų klausimus, veikė kaip savotiškas sufleris. Bet kai įsitikinome, kad tai veikia tikrai sklandžiai, panašią logiką pritaikėme ir autentifikacijai – leisti sistemai padėti atpažinti mūsų klientus*“.

Technologijų bendrovėje E4, kuriančioje mokėjimų platformas, DI buvo diegiamas palaipsniui – nuo darbuotojų iniciatyva pradėtų eksperimentų iki reikšmingų vartotojo patirties pokyčių. Pradžioje sprendimai nebuvo strateginiai, tačiau ilgainiui DI ėmė formuoti ir pačios organizacijos kryptį. Kaip pavyzdį ekspertas pateikė naujausią iniciatyvą: „*Mūsų įmonė teikia mokėjimų paslaugas, o viena iš naujų krypčių – įprastos vartotojo sąsajos pakeitimas DI grįsta pokalbių platforma. Klientai gali atlikti finansines operacijas tiesiogiai bendraudami su DI, ir tai iš esmės keičia jų patirtį su mūsų sistema*“.

Logistikos sektoriuje veikianti E5 organizacija DI sprendimus taiko keliuose svarbiuose procesuose, kurie laikomi esmine jų operacijų dalimi. DI integruotas į transporto stebėsenos sistemas, kurios realiuoju laiku leidžia sekti transporto priemonių buvimo vietą, stebėti vairuotojų veiklą – įskaitant greičio laikymąsi, poilsio laikotarpius ir kitus reglamentuojamus aspektus. Sistema neapsiriboja vien duomenų fiksavimu – ji aktyviai dalyvauja vairuotojo kasdienybėje. Pavyzdžiui, gali įspėti apie būtinybę pailsėti, pasiūlyti sumažinti greitį ar pasirinkti tinkamesnį maršrutą, atsižvelgdama į tuo metu galiojančias eismo sąlygas. Kai kuriose transporto priemonėse dirbtinis intelektas jau geba atpažinti kelio ženklus, kliūtis ar staigius situacijos pokyčius, tokiu būdu padėdamas vairuotojui greičiau reaguoti ir priimti saugesnius sprendimus. Panašūs sprendimai taikomi ir sandėliavimo srityje – čia DI

pasitelkiamas apkrovų balansavimui, srautų planavimui ir visos logistikos grandinės efektyvinimui. Nors ši technologija dar nėra įdiegta visose veiklos grandyse, ekspertas pabrėžė, kad jos vaidmuo jau dabar yra reikšmingas ir vis labiau įsiintegruojantis į organizacijos kasdienybę. „*Mes sekame vairuotojo veiklą, transporto judėjimą, sandėlio apkrovą – visa tai automatiškai stebi DI. Nors sprendimus vis dar tikrina žmogus, sistema pati jau geba įspėti apie nukrypimus*“, – paaiškino ekspertas.

E6 organizacija, veikianti kosmoso technologijų srityje, kuria dirbtiniu intelektu grįstas sistemas, skirtas palydovų paleidimo planavimui ir jų trajektorijų modeliavimui. Tokiuose kontekstuose dirbtinis intelektas nebeveikia vien kaip pagalbinė technologija – jis tampa pagrindiniu sistemos veikimo centru. Modeliai geba analizuoti vaizdinius duomenis, atpažinti objektus ir prognozuoti jų judėjimo trajektorijas, o tai itin svarbu siekiant išvengti susidūrimų kosmose ar koreguojant palydovų orbitas. Šie sprendimai remiasi nuolat pildoma duomenų baze, kurioje fiksuojama informacija apie tūkstančius Žemės orbitoje esančių objektų. Toks integruotas DI veikimas leidžia generuoti tikslias prognozes, pagrįstas realiuoju, nuolat kintančiu duomenų srautu. Tokios funkcijos ne tik padeda reaguoti į galimas grėsmes, bet ir iš anksto planuoti sudėtingus misijos parametrus. Kaip pažymėjo ekspertas, be DI tokia veikla tiesiog būtų neįmanoma: „*Kiekvienas paleidimas turi būti apskaičiuotas iki smulkmenų. DI mums leidžia matyti galimus scenarijus iš anksto ir realiu laiku juos adaptuoti, jei keičiasi sąlygos*“.

6. lentelė. Organizacinis kontekstas ir DI taikymo sritys

Eksperto kodas	**Veiklos sritis**	**DI taikymo lygis**	**DI funkcijos organizacijoje**	**DI brandos požymiai**
E1	Klientų identifikacija / fintech	Aukštas	Autentifikacija, elgsenos analizė	DI kaip pagrindas paslaugai
E2	NLP sprendimai	Aukštas	Generatyviniai modeliai, rizikų analizė	Struktūruotas taikymas, rizikų matrica
E3	Tarptautinės finansų paslaugos	Vidutinis–žemas	Vidiniai procesai, sprendimų palaikymas	Nėra atskiros DI komandos, bet yra metodikos
E4	Technologijos / mokėjimų platforma	Vidutinis–aukštas	Testavimas, klientų sąveika, generatyvinis DI	Neformalus valdymas, individualios iniciatyvos
E5	Logistika	Vidutinis–vidutinis	Maršrutų planavimas, eismo duomenų analizė	DI taikomas atskirai, spragų identifikacija
E6	Kosmoso duomenų analizė	Aukštas	Objektų analizė, orbitos planavimas	DI integruotas į sprendimus

Kaip matyti iš lentelės, apklaustose organizacijose DI taikymo apimtis ir brandos lygis labai skiriasi – vienur ši technologija jau tapusi esmine paslaugos dalimi, kitur ji veikia tik kaip pagalbinis įrankis konkrečioms funkcijoms automatizuoti. Tokie skirtumai kyla ne vien iš sektoriaus ypatumų ar įmonės dydžio, bet ir iš to, kokiu būdu vyko pati DI integracija – ar tai buvo nuosekliai suplanuotas strateginis žingsnis, ar natūraliai susiformavęs atsakas į specifinius veiklos poreikius ar komandų iniciatyvą.

Nepaisant veiklos sektoriaus ar įmonės dydžio, DI brandos lygis organizacijose ne visada atspindi technologinę pažangą. Vienur ši technologija naudojama plačiai, bet jos rizikos valdomos intuityviai, be aiškių mechanizmų. Kitur – net taikant DI tik tam tikrose srityse – rizikų valdymui jau taikomi struktūruoti sprendimai. Tai išryškina svarbią tendenciją: DI brandumas neapsiriboja vien tik technologijos mastu, o priklauso ir nuo gebėjimo užtikrinti sąmoningą jos valdymą.

Kaip teigė E3 ekspertas: „*Nors mūsų DI sprendimai prasidėjo nuo labai paprastų funkcijų, po truputį ėmėme galvoti – o kas, jei padarytų klaidą? Ką tai reikštų? Tai ir paskatino mus žiūrėti plačiau – ne vien funkcionaliai, bet ir atsakingai*“. Ši patirtis parodo, kad supratimas apie rizikas dažnai ateina su laiku, o ne kartu su pačia technologija.

Panašią mintį išsakė ir E4 atstovas, pabrėžęs, kad pokyčius dažnai inicijuoja patys darbuotojai: „*Pirmieji DI sprendimai pas mus atsirado ne dėl iš viršaus nuleistos strategijos, o dėl to, kad kažkas iš komandos narių sugalvojo išbandyti naują įrankį. Tik kai tai pradėjo duoti rezultatų, ėmėme svarstyti, kaip tai integruoti į visą sistemą*“ . Tai atskleidžia, kad inovacijos neretai prasideda nuo smalsumo ir eksperimentų, o ne nuo formalios politikos.

Kita vertus, kai kur technologinė pažanga lenkia valdymo procesus. E1 ekspertas pastebėjo: *„Mes taikome DI gana plačiai, bet apie rizikas kalbėti pradėjome gerokai vėliau, kai atsirado reguliaciniai reikalavimai. Iki tol tiesiog stengėmės, kad viskas veiktų kuo tiksliau“.* Tokiais atvejais rizikų valdymas ima vystytis tik tuomet, kai tampa būtinybe.

Tuo tarpu E2 organizacijoje atsakomybė už DI sprendimus jau integruota į platesnį vertybinį kontekstą. „*Pas mus DI nėra tik įrankis. Jis veikia kartu su mūsų vertybėmis – jeigu pastebime, kad modelis ima generuoti šališkus atsakymus, net jei jie statistiškai teisingi, stabdome procesą ir vertiname, ką tai reiškia etikos požiūriu*“, – sakė ekspertas. Toks požiūris atspindi brandesnį santykį su technologijomis - dėmesys skiriamas ne tik jų funkcionalumui ar efektyvumui, bet ir galimoms pasekmėms, kurias šie sprendimai gali sukelti platesniame socialiniame, etiniame ar aplinkos kontekste. Galima daryti išvadą, kad organizacijos skirtingai supranta ne tik DI galimybes, bet ir su ja susijusią atsakomybę. Vienose DI yra pagrindinis verslo variklis, kitose – dar tik besiformuojantis įrankis. Tačiau visų apklaustų ekspertų pasisakymuose sutapo viena idėja: DI neišvengiamai tampa vis svarbesne organizacijos dalimi, o kartu ir sritimi, kuri reikalauja kryptingo, atsakingo valdymo.

5.2 Rizikų atpažinimas ir vertinimo praktikos

Visi tyrime dalyvavę ekspertai pripažino: dirbtinio intelekto integracija neišvengiamai atneša rizikas.

E1, atstovaujantis klientų tapatybės sprendimų įmonei, pateikė konkrečią situaciją: „*Pradėjome kurti naują produktą, svarstėme naudoti ChatGPT. Tačiau paaiškėjo, kad įkeliami duomenys apdorojami serveriuose, kurių saugumo mes negalėjome garantuoti – ypač kalbant apie jautrią informaciją. Tik išnagrinėję dokumentaciją supratome, kad šitas sprendimas mums netinka*“. Šis atvejis parodo, kad net plačiai pripažinti ir technologiškai pažangūs sprendimai gali turėti paslėptų rizikų, kurios išryškėja tik tuomet, kai technologija pradedama taikyti realiomis sąlygomis.

Kalbėdamas apie rizikų identifikavimą, E1 pabrėžė, kad jų organizacija vadovaujasi situaciniu, kontekstiniu požiūriu. Jie taiko ISO 27001:2013 standartą ir rizikų vertinimo matricą, bet, kaip pastebi pats ekspertas, „*visa analizė pas mus labai kontekstinė – orientuota į tai, kur ir kaip paslauga bus naudojama*“. ES atveju daugiau dėmesio skiriama BDAR, o JAV – vietiniams teisės aktams. Tai leidžia lanksčiai prisitaikyti prie konkretaus konteksto, tačiau kartu rodo sąmoningą sprendimą nekurti universalaus, viską apimančio modelio: „*Mes nesistengiame sukurti viską apimančios sistemos – vietoje to remiamės praktika ir sveiku protu*“.

Rizikų prioritetai jų organizacijoje taip pat aiškūs – svarbiausios laikomos tos, kurios daro tiesioginį poveikį klientui ar verslo tęstinumui. Duomenų sauga, tapatybės patvirtinimas, paslaugų patikimumas – tai sritys, kurioms skiriamas didžiausias dėmesys. Kitoms, mažesnio poveikio grėsmėms taikomas stebėjimo principas: „*Į kai kurias rizikas žiūrime kaip į teorinį galvosūkį – pažymime, bet kol jos neturi aiškaus poveikio ar neatitinka mūsų veiklos specifikos, jos išlieka stebėjimo režime*“.

Šis požiūris iliustruoja, kad DI rizikų valdymas dažnai remiasi balansu tarp strateginio mąstymo ir operatyvaus prisitaikymo. Organizacijos siekia ne tiek iš anksto numatyti kiekvieną galimą riziką, kiek turėti pakankamai lanksčias sistemas, leidžiančias greitai reaguoti, kai jos tampa realios.

Šias įžvalgas patvirtina ir konkrečios praktinės situacijos. E5 ekspertas iš logistikos sektoriaus dalijosi atveju, kai DI modelis netiksliai interpretavo eismo duomenis, dėl ko buvo pasirinkti netinkami maršrutai. Technologijų įmonėje E4 paaiškėjo, kad testavimo metu formuojantis modelis galėjo inicijuoti mokėjimą be aiškios vartotojo komandos – rizika kilo dėl vadinamojo „haliucinavimas“ (*angl.* Ai BIAS) reiškinio, kai DI sistema sukuria nerealius veiksmų scenarijus.

Nors dauguma organizacijų taiko tam tikras rizikų identifikavimo priemones, ekspertai atkreipia dėmesį į tai, kad šie metodai turi savo ribas. Kai kur įmonės remiasi standartizuotais modeliais, pavyzdžiui, ISO 27001 ar rizikų vertinimo matricomis (E1), kurios leidžia įvertinti grėsmių tikimybę ir galimą poveikį. Vis dėlto tokie metodai ne visuomet geba aptikti sudėtingesnes, socialines ar etines rizikas. Tai ypač svarbu, nes galutinis vertinimas priklauso nuo žmonių, kurių sprendimus neišvengiamai

veikia jų asmeninės patirtys bei specifinis kontekstas, kuriame jie veikia. Dalies respondentų vertinimu, svarbų vaidmenį čia atlieka technologinė stebėsena – automatizuotos sistemos, galinčios realiu laiku fiksuoti DI modelių elgsenos pokyčius (E2). Tačiau net ir pažangiausios priemonės neapsaugo nuo vadinamųjų „nematomų“ rizikų, kurios išryškėja tik laikui bėgant – pavyzdžiui, dėl šališkų ar nepakankamų duomenų, kultūrinių neatitikimų ar konteksto nesupratimo. Tai rodo, kad efektyvus rizikų atpažinimas reikalauja ne tik technologinių sprendimų, bet ir žmogiškosios įžvalgos – gebėjimo reflektuoti, analizuoti, abejoti. Finansų sektoriaus atstovas E3 atkreipė dėmesį į decentralizuoto požiūrio reikšmę: „*Kiekvienas darbuotojas turi žinoti, ką laikyti rizika ir kur, ir kaip apie tai pranešti*“. Tokia kultūra leidžia greitai reaguoti į iššūkius, tačiau kartu kelia iššūkių darbuotojų sąmoningumui ir kompetencijų ugdymui.

Remiantis tyrimo interviu, galima išskirti keturias pagrindines DI rizikų grupes, kurias ekspertai įvardijo dažniausiai:

- **Techninės rizikos** – tai modelių netikslumai, „haliucinacijos“ ar sistemų veikimo nestabilumas. Tokios rizikos dažniausiai išryškėja testavimo arba realaus naudojimo metu (E4, E5, E7).
- **Reguliacinės ir atitikties rizikos** – susijusios su BDAR ar kitų teisės aktų interpretavimu, kai reglamentai ne visada dera su technologinių sprendimų specifika (E1, E2, E3). Kaip pažymėjo vienas ekspertas: „*Didžiausią painiavą kelia tie teisės aktai, kurie iš Europos lygmens dokumentų yra perteikti ar „adaptuoti“ nacionaliniu mastu – dažnai prarandamas pirminis kontekstas*.“
- **Organizacinės rizikos** – kyla dėl neaiškių atsakomybių, sprendimų automatizavimo be žmogaus priežiūros ar aklo pasitikėjimo DI sistemomis (E3, E4).
- **Šališkumo rizikos** – aktualios ypač generuojančiams ir NLP modeliams, kur ribotas ar neįvairus duomenų kiekis lemia iškreiptus rezultatus, stereotipus ar net diskriminacinius sprendimus (E2).

Svarbu pažymėti, kad šios rizikų grupės dažnai persidengia. Techninė klaida gali peraugti į teisinį iššūkį, o netiksliai veikiantis algoritmas – sukelti organizacines pasekmes ar pakenkti reputacijai. Dėl to viena rizika neretai išprovokuoja kitą, o visa rizikų sistema tampa dinamiška ir glaudžiai tarpusavyje susijusi.

DI rizikų atpažinimas šiandien organizacijose vyksta kaip mišrus procesas – jis priklauso ne tik nuo technologinių sprendimų, bet ir nuo žmonių gebėjimo stebėti, kritiškai vertinti, kelti klausimus ar net suabejoti pačios sistemos sprendimais. Tokiu atveju rizika tampa ne vien trukdžiu, bet ir vertingu signalu – ji išryškina tas vietas, kur dar trūksta brandos, kontrolės ar sąmoningo požiūrio.

5.3 DI rizikų valdymo strategijos ir atsakomybes pasiskirstymas

Nustačius su DI susijusią riziką, kyla praktinis klausimas – kaip į ją reaguojama organizacijų viduje? Kas priima sprendimus, kokių veiksmų imamasi, kaip paskirstomos atsakomybės? Tyrimo duomenys atskleidžia nevienalytį vaizdą: vienose įmonėse taikomi formalūs, aiškiai apibrėžti procesai, tuo tarpu kitose sprendimai dažnai priklauso nuo konkrečios situacijos ar darbuotojų iniciatyvos.

Dauguma organizacijų rizikų valdymą grindžia ankstyvose DI diegimo stadijose pasirinktais principais.

E1 ekspertas pabrėžė, kad jų organizacijoje remiamasi kontekstiniu požiūriu: „*Visa analizė pas mus labai kontekstinė – orientuota į tai, kur ir kaip paslauga bus naudojama*“. Tai leidžia greitai reaguoti į skirtingus reguliacinius reikalavimus, tačiau kartu rodo, kad sprendimai dažnai formuojami iš praktikos, o ne remiantis vientisa strategija.

E2 atvejis išsiskyrė labiau išgryninta, paskirstyta atsakomybių sistema. Pasak eksperto, pirmasis riziką dažniausiai pastebi priklausomai nuo jos pobūdžio: „*Jeigu kalbame apie netikėtą sistemos elgseną ar veikimo sutrikimą, jį dažniausiai fiksuoja automatinės stebėsenos priemonės. Tačiau jei kyla etinio pobūdžio klausimas – pavyzdžiui, dėl šališko atsakymo ar neadekvataus turinio – tai pirmiausia pastebi žmogus: naudotojas, moderatorius ar vidinis audito komandos narys“.*

Tokia dviejų lygių sistema leidžia apimti tiek techninius, tiek kultūrinius aspektus. Jei rizika reikšminga, sprendimų priėmimas perduodamas tarpdisciplininei grupei, kurioje dalyvauja technologijų, etikos ir teisės ekspertai. „*Mažesnio masto atvejais sprendimai priimami decentralizuotai – komandose, atsakingose už konkretų produktą ar sistemą*“, – paaiškino ekspertas. Kiekvienas DI komponentas turi savo „savininką“, kuris atsakingas už jo funkcionavimą ir saugumą, tačiau egzistuoja ir kolegialūs mechanizmai, užtikrinantys, kad rizika neliktų „*niekieno*“ atsakomybe.

E3 organizacijoje rizikų valdymas grindžiamas aiškiai decentralizuotu, bet koordinuotu principu. Atsakomybė už rizikų atpažinimą čia nėra priskirta tik vienai komandai – kiekvienas darbuotojas laikomas aktyvia šio proceso dalimi. Kaip paaiškino ekspertas, „*pas mus rizikos valdymas nėra paliktas tik vienai atsakingai komandai – pagal įmonės politiką kiekvienas darbuotojas yra atsakingas už tai, kad atpažintų galimą riziką ir apie ją praneštų*“.

Tam, kad šis modelis veiktų, organizacijoje sukurta speciali mokymų programa, padedanti darbuotojams suprasti, kur baigiasi paprastas trikdis ir prasideda tikra rizika. Visos identifikuotos rizikos galų gale keliauja į krizės valdymo skyrių – struktūrą, kuri, pasak eksperto, veikia kaip „organizacijos „nervų sistema“, kur viskas suvedama, analizuojama ir koordinuojami tolesni veiksmai“. Tokia schema užtikrina, kad rizikos neužstrigtų pavienėse grandyse, o pasiektų sprendimų priėmėjus laiku: „*Toks modelis leidžia greitai reaguoti, nes rizikos nėra „užstrigusios“ pas vieną žmogų – jos turi aiškų kanalą, kaip pasiekti sprendimų priėmėjus*“.

Kiek kitaip rizikų valdymo struktūra organizuota E4 atveju – mokėjimų bendrovėje. Čia DI rizikos nelaikomos vien techninės komandos atsakomybe – organizacijoje įdiegtas nuolatinės stebėsenos mechanizmas. „*Yra paskirtas atsakingas asmuo, atliekantis nuolatinį auditą. Aptikus pažeidžiamumą, informacija perduodama mokslinių tyrimų ir plėtros skyriui, kuris priima sprendimus dėl korekcijų*“, – teigė ekspertas. Toks modelis užtikrina ne tik greitą reagavimą, bet ir aiškų atsakomybės paskirstymą, taip sumažinant riziką, kad problema liks nepastebėta ar neperduota tinkamiems sprendimų priėmėjams.

Tuo tarpu E5 atveju, atstovaujančiame logistikos sektoriui, rizikos dažniausiai pirmiausiai pastebimos operacijų ar IT skyriuose. Nors organizacijoje yra paskirti atsakingi asmenys, realybėje informacijos perdavimas ne visada vyksta sklandžiai: „*Turime aiškiai apibrėžtus atsakingus asmenis, tačiau kartais informacija vėluoja pasiekti sprendimų priėmėjus*“. Ši patirtis atskleidžia, kad net turint formalizuotą struktūrą, rizikų valdymas gali susidurti su kliūtimis, ypač kai sprendimų efektyvumas priklauso nuo greito informacijos srauto.

E6 organizacijoje, kurioje kuriami palydovų paleidimo ir trajektorijų simuliacijos sprendimai, DI rizikos vertinamos itin atsakingai – atsižvelgiant į jų galimą kritinį poveikį misijų sėkmei. Pasak eksperto, dauguma techninių rizikų pirmiausia fiksuojamos pačios sistemos: „Kadangi mūsų sistema veikia realiuoju laiku ir apdoroja didžiulius duomenų srautus, turime automatizuotus indikatorius, kurie perspėja apie anomalijas – tiek trajektorijos skaičiavimuose, tiek objektų identifikavime.“

Vis dėlto net ir turint aukšto lygio technologinę bazę, žmogiškasis veiksnys išlieka būtinas. Ekspertas pabrėžė, kad sprendimų priėmimas visada vyksta bendradarbiaujant tarp skirtingų sričių specialistų: „*Kai iškyla rizika, pirmiausia ją užfiksuoja DI, bet sprendimas dėl veiksmų priėmimo visada pereina per žmones – mes vertiname, kokio masto poveikį tai gali turėti misijai ir ar reikia keisti skrydžio planą“.*

Šioje organizacijoje taikomas aiškus eskalavimo modelis – jei rizika atitinka „kritinio poveikio“ kriterijus, pavyzdžiui, galimą susidūrimą su kitu objektu orbitoje, sprendimas perduodamas „mobiliajai analizės grupei“. Tai tarpdisciplininė komanda, kurią sudaro ne tik inžinieriai, bet ir teisininkai, atsakingi už saugos ir atitikties užtikrinimą. Šis modelis leidžia vienu metu įvertinti tiek techninius parametrus, tiek teisinius aspektus.

Kaip pažymėjo ekspertas, tokia struktūra susiformavo neatsitiktinai – „*kosmose klaidų kaina yra labai didelė – todėl negalime sau leisti, kad sprendimas priklausytų nuo vieno žmogaus ar pavienės komandos*“. Tai atskleidžia, kad šiame sektoriuje DI rizikų valdymas yra neatsiejamas nuo platesnio saugos kultūros suvokimo, kuriame technologinis tikslumas visada derinamas su žmogiška kontrole ir atsakomybe.

Vertinant ekspertų pasisakymus apie DI rizikų valdymą organizacijose galima išskirti pagrindines šio skyriaus įžvalgas:

- **Atsakomybės paskirstymas nevienodas –** vienose organizacijose egzistuoja aiškūs procesai ir specializuoti rizikų valdymo padaliniai, kitur – sprendimai remiasi darbuotojų sąmoningumu ir neformaliomis iniciatyvomis. Didesnis veiksmingumas pasiekiamas ten, kur formalūs mechanizmai derinami su aktyvia vidine komunikacija ir nuolatiniu mokymu.
- **Standartai dažnai neatnaujinami –** nors plačiai naudojamas ISO/IEC 27001, tai bendras informacijos saugos standartas, kuris neapima specifinių DI keliamų rizikų. Naujesni standartai, tokie kaip ISO 42001, tyrime beveik neminėti – tai atskleidžia tarpą tarp teorinių gairių ir realios praktikos.
- **Žmogaus vaidmuo išlieka lemiamas –** kai rizikos fiksuojamos DI sistemų pagalba, galutinius sprendimus vis dar priima žmonės. Veiksmingiausiai veikia modeliai, kuriuose rizikų vertinime dalyvauja tarpdisciplininės komandos – nuo technologų iki teisininkų ar etikos ekspertų.
- **Reguliacijų vėlavimai –** DI technologijos vystosi greičiau nei reguliavimo priemonės. Organizacijos, kurios veikia pro aktyviai ir pačios kuria vidaus tvarkas dar iki išorinių reikalavimų atsiradimo, įgyja konkurencinį pranašumą.
- **Rizikų valdymas pereinamojoje stadijoje –** šiandieninė praktika dažnai svyruoja tarp tradicinių IT saugos metodų ir dar tik besiformuojančių DI rizikų valdymo principų. Tos organizacijos, kurios riziką suvokia kaip neatsiejamą inovacijos dalį, geriau prisitaiko ir kuria ilgaamžiškesnes, atsparesnes sistemas.

5.4 DI rizikų valdymo teisinio reguliavimo iššūkiai

Dirbtinio intelekto reguliavimas šiandien jau seniai peržengė vien tik atitikties uždavinius – tai tampa nuolat besikeičiantis procesas, kuriame organizacijos turi ne tik žinoti įstatymus, bet ir gebėti juos interpretuoti bei prisitaikyti prie dinamiškos technologijų raidos. Tyrimo metu paaiškėjo, kad daugelis organizacijų į teisines normas žiūri ne kaip į stabilią struktūrą, o kaip į nuolat kintantį lauką, kuriame būtinas nuolatinis stebėjimas ir lankstumas.

Skirtingi sektoriai, bendri iššūkiai

E1 atstovas, dirbantis su klientų tapatybės nustatymų, pabrėžė praktinio reguliavimo taikymo sudėtingumą: „*Reguliavimas dažnai atsiranda tik tada, kai problema jau išryškėjusi, o kol tai perauga į standartus ir realius technologinius pakeitimus – praeina nemažai laiko*“. Jų komanda nuolat seka teisės aktų pokyčius, tačiau dėl skirtingų jurisdikcijų (ES, JAV, kt.) šis procesas tampa sudėtingas. BDAR dažnai minima kaip esminis dokumentas, tačiau jo interpretacijos sukelia sunkumų: „*Kai kurie BDAR*

reikalavimai, ypač kai juos perrašo nacionaliniu lygiu, tampa dviprasmiški – nelabai aišku, ką jie realiai reiškia technologijų kontekste“.

E2, kurianti natūralios kalbos apdorojimo sprendimus, nurodė, kad teisiniai klausimai jų organizacijoje sprendžiami taikant tarpdisciplininį principą – juose dalyvauja ne tik teisininkai, bet ir technologijų bei etikos specialistai. Vienas iš pašnekovų pažymėjo, kad naujai kuriamos reguliavimo sistemos, tokios kaip AI Act ar DSA, šiuo metu vis dar atrodo gana abstrakčios. Jis komentavo: „*Kol tai tik principai – viskas aišku. Bet kai atsiranda konkrečios taisyklės, tai ir prasideda tikrieji iššūkiai. Ar mes tikrai suprantam, ką tai reiškia mūsų produktui?“.*

E3 atveju, dirbant tarptautinėje finansų srityje, teisinių žinių trūkumas realiai sukėlė riziką – vienas darbuotojas, nežinodamas galimų pasekmių, įkėlė kodą į viešą DI platformą. Ekspertas pabrėžė: *„Teisės aktai nepadės, jei žmonės nesuvokia, ką jie reiškia jų darbe“* – ši citata aiškiai atskleidžia, kad formali atitiktis savaime negarantuoja saugumo, jei nėra tikro supratimo, kaip taisyklės veikia praktikoje. E4, veikianti fintech sektoriuje, susiduria su praktinėmis BDAR interpretavimo problemomis. *„Kartais pats užklausos turinys neturi nieko bendro su asmens duomenimis, bet dėl techninės struktūros tai jau laikoma BDAR objektu“, –* sakė ekspertas, pabrėždamas, kaip techniniai sprendimai gali būti traktuojami kaip teisiniai pažeidimai dėl interpretacinių niuansų.

Pokalbiuose su ekspertais ne kartą nuskambėjo mintis, jog skirtingi teisės aktai neretai veikia nesuderintai, o tai kelia praktinių iššūkių organizacijoms. Vienas dažniausiai minėtų atvejų – įtampa tarp BDAR reikalavimų ir Skaitmeninių paslaugų akto (DSA) nuostatų. Kaip pažymėjo E2 ekspertas, įmonės, kurios anksčiau investavo į stiprią tapatybės apsaugą, dabar priverstos iš esmės peržiūrėti savo sprendimus: „*Iki šiol buvome įpratę, kad tapatybė turi būti paslėpta - dabar DSA reikalauja aiškiai identifikuoti prekybininką ar turinio teikėją. Tai ne tik komunikacijos iššūkis. Tai reiškia realius pokyčius – nuo duomenų valdymo iki to, kaip informacija pateikiama naudotojui*.“

Naujai priimami teisės aktai ne visada papildo anksčiau galiojusius – kai kuriais atvejais jie keičia esamų nuostatų prasmę arba įveda naujus, dažnai sudėtingus reikalavimus. Dėl to organizacijos neretai priverstos iš esmės peržiūrėti ir adaptuoti jau taikomas praktikas. **BDAR akcentavo duomenų minimizavimą ir identiteto apsaugą, o DSA reikalauja viešumo ir skaidrumo**. Tokia priešprieša verčia organizacijas spręsti kompleksinius klausimus: kaip pateikti informaciją nepažeidžiant anksčiau sukurtų apsaugos sistemų, kaip išlaikyti klientų pasitikėjimą, kai keičiasi pačios komunikacijos taisyklės.

E2 ekspertas pabrėžė, kad ši dilema ypač jautri DI paslaugų teikėjams, dirbantiems su verslo klientais: „*Kai pradedi viešinti duomenis apie savo klientus – net jei to reikalauja įstatymas – natūraliai kyla klausimas, ar jie jausis saugiai. Iki šiol mūsų pažadas buvo „mes jus saugosim“, o dabar turim pasakyti „bet tik tiek, kiek leidžia nauja tvarka“. Tokią žinutę reikia labai gerai suvaldyti*“.

Šis pavyzdys atskleidžia, kad teisiniai pokyčiai veikia ne tik dokumentuose, bet ir pačiuose organizaciniuose procesuose – nuo infrastruktūros iki reputacijos valdymo.

E5 ekspertas atvirai pripažino, kad apie naujus teisės aktus girdėta, bet giluminė analizė dar neatlikta: „*Žinau, kad kažkas buvo ruošiama, bet kol dar neįsigaliojo – sunku pasakyti, kaip tai paveiks mūsų veiklą*“. Tai rodo pasyvų laukimo režimą, kai sprendimai priimami tik tada, kai jau būtina.

Tuo tarpu E6, veikiantis kosmoso technologijų srityje, atskleidė pažangesnį požiūrį – reguliavimą jie suvokia kaip iš anksto planuojamą atsakomybę, o ne tik reakciją po įvykusio incidento. „*Kai kyla klausimas dėl saugumo orbitoje, turi būti tikras, kad tavo sprendimas ne tik techniškai teisingas, bet ir teisiškai pagrįstas – net jei tokio reguliavimo dar nėra, jis gali atsirasti po įvykio*“ – pabrėžė ekspertas.

Remiantis ekspertų įžvalgomis, galima išskirti penkis dažniausiai pasikartojančius teisinio reguliavimo iššūkius, su kuriais susiduria su dirbtiniu intelektu dirbančios organizacijos:

- **Reguliacinių aktų tarpusavio nesuderinamumas** – skirtingi teisės aktai, tokie kaip BDAR ir DSA, dažnai kelia prieštaringus reikalavimus: vieni akcentuoja duomenų apsaugą ir anonimiškumą, o kiti – skaidrumą bei naudotojų identifikavimą. Tokie konfliktai verčia organizacijas derinti nesuderinamus tikslus ir kurti kompromisines praktikas, kurios ne visada veikia aiškiai ar efektyviai.
- **Neaiškios ar dviprasmiškos interpretacijos –** ypač BDAR kontekste, teisės normų taikymas technologijoms, kurios vystosi greičiau nei teisė, dažnai sukelia neaiškumų. Organizacijos susiduria su situacijomis, kai ta pati techninė funkcija gali būti interpretuojama skirtingai skirtingose jurisdikcijose ar net institucijose.
- **Vėluojantis reguliavimas –** daugeliu atvejų teisiniai aktai atsiranda tik reaguojant į jau įvykusius incidentus. Tai reiškia, kad organizacijos, kurios nelaukia formalių reikalavimų ir pačios proaktyviai kuria vidaus taisykles, turi konkurencinį pranašumą.
- **Trūksta specifinių standartų DI kontekste –** dauguma organizacijų vis dar remiasi bendraisiais informacijos saugos standartais, tokiais kaip ISO 27001, tačiau jie nėra pritaikyti DI specifikai. Naujieji dokumentai, pavyzdžiui, ISO/IEC 42001:2023, dar mažai žinomi ir retai taikomi praktikoje.
- **Sąmoningumo ir kompetencijos spragos –** vien formalus taisyklių egzistavimas negarantuoja jų laikymosi – kaip pastebėjo keli ekspertai, darbuotojai dažnai nežino, kaip teisės aktai taikomi jų konkrečiame darbe. Todėl teisinis reguliavimas turi būti lydimas vidinių mokymų, aiškios komunikacijos ir praktinių vertimų į kasdienę veiklą.

Dirbtinio intelekto reguliavimo sritis vis dar formuojasi – ji reikalauja ne tik laikytis esamų taisyklių, bet ir aktyviai prisidėti prie jų kūrimo. Šiandien brandi organizacija – tai ne tik ta, kuri laikosi reikalavimų, bet ta, kuri geba juos suprasti, interpretuoti ir pritaikyti dinamiškai kintančiame technologijų pasaulyje.

5.5 Iššūkiai ir tobulinimo poreikiai DI rizikų valdyme

Viena svarbiausių tyrimo įžvalgų paaiškėjo tuomet, kai pokalbiai peržengė dabartinių praktikų ribas ir nukrypo į ateities vizijas – kaip organizacijos įsivaizduoja brandesnį, nuosekliau struktūruotą bei technologijomis pagrįstą DI rizikų valdymą. Ekspertų mintys atskleidė tiek nevienodą organizacijų brandos lygį, tiek bendrą norą pereiti nuo padrikų, epizodinių sprendimų prie vientisos, vertę generuojančios valdymo sistemos. Vis labiau įsitvirtina supratimas, kad rizika – tai ne vien formali atitiktis, o reikšmingas strateginis resursas, dėl kurio kinta organizacijų požiūris: jos pradeda mąstyti ne apie pavienius sprendimus, o apie ilgalaikę, apgalvotą ir į ateitį orientuotą valdymo kryptį.

Kai kuriose diskusijose nuskambėjo netikėti, bet itin ambicingi siūlymai. E1 ekspertas kėlė klausimą: „*Yra tūkstančiai sričių, kur DI jau keičia žmogaus darbą – tai kodėl negalėtume sukurti tokio kaip 'rizikų roboto'? T. y. DI, kuris pats automatiškai peržiūri sistemą, palygina ją su reguliavimu, nurodo spragas ir net generuoja rekomendacijas, ką keisti*“. Jo vizijoje toks įrankis veiktų ne vien kaip stebėsenos mechanizmas, bet kaip aktyvus sprendimų priėmimo dalyvis – nuolat besimokantis, prisitaikantis prie pokyčių ir veikiantis ne reaktyviai, o proaktyviai.

Šią mintį papildė idėja apie tarptautinį, centralizuotą rizikų valdymo agentą – sistemą, kuri „*turėtų galimybę suprasti konkrečią organizacijos veiklą, sulyginti ją su reguliaciniais reikalavimais ir pasiūlyti individualizuotą sprendimų žemėlapį*“. Tai būtų universalus „reguliacinis sluoksnis“, prie kurio jungtųsi organizacijos visame pasaulyje – „*tarsi bendras API, suderinantis teisę, technologijas ir organizacinę praktiką*“.

E2 ekspertas iškėlė kultūrinį rizikos suvokimo aspektą. Pasak jo, „*šiandien DI rizikos vis dar dažnai suprantamos kaip 'grėsmės, kurių reikia išvengti', tačiau brandesniame požiūryje rizika turėtų būti traktuojama kaip žinojimo ir strateginės orientacijos laukas*“. Jis pridėjo, kad „*jeigu būtų įrankis, kuris ne tik stebi, bet ir paaiškina – kodėl tai rizika, kaip ją suprasti socialiniame kontekste, koks būtų jos poveikis – tai keistų ir sprendimų kokybę, ir atsakomybės jausmą*“.

Kai kurie ekspertai iškėlė ir dar gilesnį barjerą – organizacinę kultūrą. E3 įžvalga buvo itin taikli: „*Jei žmonės nežino, kas yra rizika – jokia sistema nepadės.“* Tai parodo, kad net ir puikiai sukonstruota sistema neveiks, jei nebus žmogiškojo suvokimo ir motyvacijos ją taikyti. Rizikų valdymas čia tampa ne tik struktūra ar metodika, bet ir dalimi bendros organizacinės sąmonės – refleksijos, kuri turi būti integruota į kasdienį sprendimų priėmimą.

E4 ekspertas pasiūlė alternatyvią, labiau pragmatišką kryptį – rizikų valdymą derinant su finansiniais mechanizmais. Jis svarstė, kad ne visos rizikos turi būti suvaldytos vidiniais resursais – kai kuriais atvejais jos gali būti perkeltos, pavyzdžiui, draudimo paslaugoms: „*Mažesnio masto rizikos galėtų būti perkeltos į draudimo mechanizmus. DI rizikų valdymas turėtų būti lankstus ir neperkrautas pernelyg detaliomis 'biurokratinėmis' priemonėmis*“. Pasak jo, ypač *fintech* sektoriuje vis aiškiau

matomas poreikis turėti ne tik technologinį ar teisinį, bet ir finansinį saugumo tinklą, kuris padėtų sušvelninti galimų incidentų pasekmes. Tai rodo, kad brandesniame požiūryje rizika suvokiama ne tik kaip valdymo objektas, bet ir kaip strateginės investicijos klausimas.

E5 ekspertas, dirbantis logistikos sektoriuje, pastebėjo, jog rizikų valdymas dažnai atliekamas pernelyg retrospektyviai – sprendimai priimami jau įvykus pažeidimui, o ne proaktyviai siekiant jo išvengti: „*Kol kas mūsų reakcija yra tokia: įvyko → sureagavom. Bet būtų logiška, jei DI pats stebėtų, modeliuotų rizikos scenarijus ir siūlytų sprendimus prieš jiems tampant realybe*“. Ši mintis žymi pokytį – nuo pasyvaus reagavimo prie aktyvios prognozės ir pasiruošimo.

E6 ekspertas akcentavo labai žemišką, bet dažnai pamirštamą aspektą – duomenų kokybę. Pasak jo, „*net ir geriausias DI sprendimų modelis taps bevertis, jei jį maitinsim prastais, neaiškios kilmės ar pasenusiais duomenimis. Tai – tarsi bandyti pastatyti dangoraižį ant pelkės*“. Jis siūlė, kad rizikų valdymo sistemos pirmiausia turėtų gebėti įsivertinti savo informacinę bazę – mat rizika dažnai prasideda dar prieš analizę, būtent nuo duomenų nepatikimumo.

Šios įvairios, bet kryptingai vienas kitą papildančios vizijos rodo, kad dirbtinio intelekto rizikų valdymas nebegali būti suvokiamas vien kaip formalus atitikimo uždavinys. Tai jau nebe techninis procesas, o dinamiška erdvė, kurioje susikerta technologijos, teisė, etika, organizacinė kultūra ir strateginis mąstymas. Kad šios sritys galėtų veikti darniai, reikalingas ne tik struktūrinis pergalvojimas, bet ir gilus kultūrinis virsmas – nuo baimės ir gynybinių sprendimų link pasitikėjimo, atvirumo ir kūrybiškumo.

Atsižvelgiant į ekspertų įžvalgas, galima išskirti kelis pagrindinius bruožus, kurie, tikėtina, formuos brandesnį DI rizikų valdymą ateityje:

- **Proaktyvumas vietoje reaktyvumo:** organizacijos vis dažniau ieško sistemų, galinčių iš anksto identifikuoti rizikas ir pasiūlyti veiksmų kryptis dar prieš joms materializuojantis.
- **DI – ne tik objektas, bet ir įrankis:** vizijos apie „rizikų robotą“ ar automatizuotą reguliacinį agentą rodo, kad pats dirbtinis intelektas gali tapti aktyviu rizikų valdymo dalyviu.
- **Kultūrinis lūžis:** rizika pradeda būti suvokiama ne kaip grėsmė, o kaip strateginis žinojimo šaltinis – tai keičia organizacijų santykį su neapibrėžtumu ir skatina mokymosi kultūrą.
- **Rizikų diferencijavimas:** ne visos rizikos turi būti valdomos vienodai – dalis gali būti efektyviai perkeltos (pvz., per draudimą), taip išlaisvinant išteklius svarbiausiems iššūkiams.
- **Sisteminės integracijos poreikis:** vis labiau akcentuojamas poreikis kurti ne pavienes priemones, o visuminę ekosistemą, apimančią techninę, teisinę, socialinę ir finansinę rizikų dimensiją.

Šios įžvalgos leidžia suprasti, kad DI rizikų valdymas nebėra tik techninė ar teisinė procedūra. Jis tampa platesniu organizaciniu klausimu, atspindinčiu tai, kaip įmonė mąsto apie ateitį, kiek yra pasirengusi ne tik reaguoti į išorinius pokyčius, bet ir juos formuoti. Gebėjimas apjungti technologinį budrumą, etinį jautrumą ir strateginę viziją – tai vis dažniau tampa tuo, kas skiria šiuolaikinę organizaciją nuo tiesiog modernios.

5.6 Tyrimo interpretacija ir tolesnio taikymo kryptis

Tyrimo struktūra rėmėsi aiškiai suformuluotais klausimais: kokiomis aplinkybėmis organizacijos pradeda naudoti dirbtinį intelektą, kokie modeliai dominuoja atpažįstant ir vertinant rizikas, kaip paskirstoma atsakomybė bei kaip teisinis reguliavimas formuoja praktinius sprendimus. Visi šie aspektai buvo nuosekliai nagrinėti remiantis empiriniais duomenimis, ekspertų įžvalgomis ir konkrečiais pavyzdžiais iš skirtingų sektorių. Nors organizacijų patirtys skiriasi, viena tendencija pasikartojo visame tyrime – ryžtas judėti nuo epizodinių, fragmentuotų sprendimų prie sisteminio, integruoto ir strategiškai pagrįsto DI rizikų valdymo. Didėjantis supratimas, jog rizika yra ne vien grėsmė, bet ir strateginis išteklius, atveria kelią brandesnei valdymo logikai – tokiai, kurioje svarbus ne tik reagavimas, bet ir išankstinis prognozavimas, kontekstų interpretavimas bei atsakinga integracija.

Vienas iš labiausiai perspektyvių būdų šias įžvalgas perkelti į praktiką – tai integruoto GRC (*angl. Governance – Risk – Compliance*) modelio taikymas. Nors GRC koncepcija nėra nauja, dirbtinio intelekto kontekste ji įgyja papildomą aktualumą – kaip priemonė struktūruotai susieti skirtingus organizacijos lygmenis ir rizikų valdymą paversti ne izoliuotu, o visos strategijos bendra dalimi.

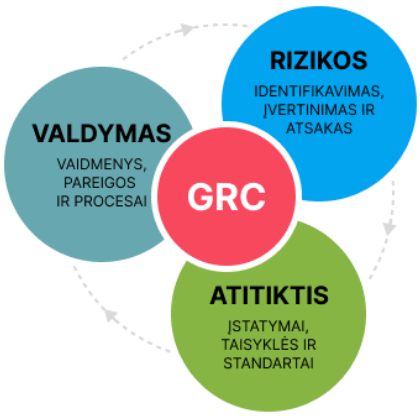
GRC modelis čia veiktų ne kaip dar viena formalumo norma, o kaip loginis karkasas, padedantis organizacijoms susivokti, kur jos yra, kokių principų laikosi ir kur link juda:

- **Valdymas *(angl. Governance):*** Ar organizacijoje aiškiai paskirstytos funkcijos, susijusios su DI rizikomis? Kas atsakingas už sprendimų priėmimą, kas – už stebėseną, o kas prižiūri etinius bei teisinius aspektus? Ar veikia mechanizmai, užtikrinantys nepertraukiamą DI sprendimų priežiūrą ir atskaitomybę? Šie klausimai padeda įvertinti, ar organizacija turi ne tik formalias struktūras, bet ir realius įgaliojimus valdyti DI procesus sąmoningai ir tvariai.
- **Rizikų valdymas *(angl. Risk):*** Ar organizacija geba laiku atpažinti technologines, etines, teisines ar reputacines rizikas, susijusias su DI sprendimais? Ar į šias rizikas žiūrima tik kaip į grėsmes, ar ir kaip į galimybes sustiprinti strateginį planavimą? Toks požiūris leidžia ne tik apsisaugoti, bet ir transformuoti rizikų valdymą į priemonę, kuriančią organizacinę vertę.
- **Atitiktis *(angl. Compliance):*** Ar organizacija neapsiriboja vien galiojančių teisinių reikalavimų (pvz., BDAR, AI Act, DSA) vykdymu, bet ir aktyviai seka jų pokyčius? Kaip

užtikrinama, kad DI sprendimai būtų nuosekliai derinami su nuolat kintančia reguliacine aplinka – tiek nacionaliniu, tiek tarptautiniu mastu? Efektyvus atitikties valdymas padeda ne tik išvengti pažeidimų, bet ir kurti pasitikėjimą DI taikymu organizacijos viduje ir išorėje.

Šio modelio stiprybė – jo pritaikomumas: brandesnės organizacijos gali jį naudoti kaip strateginį kompasą, o mažesnėms jis gali veikti kaip aiškus rėmas, padedantis susiorientuoti tarp dažnai prieštaringų reikalavimų. Vienas ekspertų taikliai pažymėjo: „*Mums nebereikia dar vieno popierinio reglamento – mums reikia įrankio, kuris augtų kartu su mumis*“. **GRC** galima vertinti ne tik kaip taisyklių laikymosi mechanizmą, bet ir kaip priemonę, skatinančią gilesnį rizikų supratimą, kritinę refleksiją, gebėjimą prognozuoti galimus scenarijus bei veikti atsakingai. Atliktas tyrimas parodė, kad dirbtinio intelekto rizikų valdymas dažniausiai nėra izoliuota funkcija – jis natūraliai susipina su platesniais organizaciniais procesais, tokiais kaip sprendimų priėmimo logika, darbuotojų kompetencijų plėtra ar inovacijų valdymas. Todėl svarbu kalbėti ne tik apie pasirinktus standartus, bet ir apie jų veikimą praktikoje: ar jie yra suprantami, ar aiškiai paskirstyta atsakomybė, ir ar tikrai pritaikyti prie konkretaus organizacijos veiklos konteksto.

7 pav. GRC Modelis DI rizikų valdymo pritaikymui



Šaltinis: Sudaryta autorės pagal OCEG Red Book 3.5 (2024)

Pagrindinė tyrimo žinia – DI rizikų valdymas yra daugialypis, evoliucinis procesas. Vienose organizacijose DI jau yra integruotas į strategines veiklos ašis, kitose – vis dar eksperimentuojama. Tačiau bendras vektorius aiškus: ieškoma sisteminio, reflektuoto ir prisitaikančio požiūrio į rizikas. GRC modelis čia galėtų atlikti tilto funkciją – ne kaip galutinis sprendimas, bet kaip struktūruotas pagrindas, leidžiantis kurti brandesnį, į ateitį orientuotą rizikų valdymą.

6. IŠVADOS IR REKOMENDACIJOS

Literatūros analizės išvados.

1. Atlikta literatūros analizė parodė, kad dirbtinio intelekto rizikų valdymo aplinka yra sudėtinga ir dinamiška. Joje susilieja spartūs technologiniai pokyčiai, tebesivystančios teisinio reguliavimo iniciatyvos ir organizacijų priimami praktiniai sprendimai. Vis dėlto teisėkūros procesai neretai nespėja reaguoti į greitą DI raidą, todėl atsiranda vadinamosios „pilkosios zonos“ - sritys, kuriose atsakomybės ribos lieka miglotos, o taikomi reguliaciniai instrumentai dar tik formuojami.
2. Dirbtinis intelektas kelia daugiapuses rizikas - nuo techninių ar sisteminių klaidų ir duomenų pažeidžiamumo iki gilesnių socialinių bei etinių problemų, tokių kaip šališkumas, diskriminacija ar žmogaus kontrolės silpnėjimas. Vien technologinių priemonių šiems iššūkiams suvaldyti nepakanka. Reikalingas kompleksinis, tarpdisciplininis požiūris, apimantis ne tik technologinius, bet ir teisinius bei etinius aspektus.
3. Europos Sąjungos dokumentuose pabrėžiama, kad DI sistemų sprendimai turi būti atsekami, skaidrūs ir atsakingi. Praktikoje tai reikalauja naujų organizacinių priemonių, kurios dar nėra iki galo išvystytos.
4. DI rizikų valdymui siūlomi įvairūs modeliai – ISO 31000, NIST, GRC. Literatūroje pripažįstama, kad šie modeliai yra vertingi rizikų valdymo įrankiai, tačiau jų veiksmingumas priklauso nuo gebėjimo juos pritaikyti konkrečiam DI kontekstui ir sektoriaus ypatumams. Etiniai ir kultūriniai aspektai literatūroje įvardijami kaip būtini DI valdymo komponentai – be žmogaus priežiūros automatizuoti sprendimai kelia grėsmę atsakomybei ir pasitikėjimui.

Atlikto tyrimo (ekspertų interviu) išvados:

1. Dirbtinio intelekto diegimas dažniausiai prasideda nuo praktinių, lokalių iniciatyvų, o ne nuo strategiškai suplanuotų sprendimų. Dėl to DI integracija vyksta netolygiai – pati technologija dažnai įsitvirtina greičiau nei jos keliamos rizikos sulaukia atitinkamų valdymo priemonių.
2. Rizikos paprastai išryškėja ne planavimo stadijoje, o realiame veiklos procese. Socialinio, etinio ar organizacinio pobūdžio rizikos retai numatomos iš anksto – jos tampa akivaizdžios tik veikiant praktikoje.
3. Rizikų valdymo būdai organizacijose labai skirtingi – nuo pavienių, atsitiktinių veiksmų iki formalizuotų sistemų. Net ir taikant tarptautinius standartus, tokius kaip ISO 27001, dažnai pasigendama specifinio jų pritaikymo DI kontekstui.
4. Atsakomybės paskirstymas ne visuomet aiškus, ypač kai DI diegimas dar laikomas eksperimentiniu. Tuo tarpu labiau subrendusiose organizacijose jau galima matyti aiškesnį vaidmenų pasiskirstymą ir koordinavimą.

5. Technologinės stebėsenos sistemos yra svarbios, tačiau be žmogaus įsitraukimo – nepakankamos. Etiniai ir kultūriniai klausimai reikalauja gilesnio konteksto suvokimo – tokio, kurį gali užtikrinti tik žmogaus patirtis ir vertybinis mąstymas.
6. Teisinis reguliavimas dažnai nespėja su DI raida, o tokie teisės aktai kaip BDAR ar DSA kartais kelia prieštaravimų, verčiančių organizacijas peržiūrėti jau veikiančius sprendimus. ISO/IEC standartai organizacijose taikomi selektyviai.
7. Naujesni, DI specifiką geriau atliepiantys dokumentai, tokie kaip ISO/IEC 42001:2023, kol kas dar nėra plačiai žinomi ar taikomi.
8. Tarpdisciplininis bendradarbiavimas – pasirodė ypač veiksmingas. Technologų, teisininkų ir etikos specialistų dialogas leidžia išsamiau suprasti su DI susijusias rizikas ir priimti labiau subalansuotus, kontekstui jautrius sprendimus.
9. Organizacinė kultūra turi esminę reikšmę rizikų valdymo kokybei. Net turint aiškiai apibrėžtus procesus ir įrankius, be darbuotojų sąmoningumo, vidinės komunikacijos ir atsakomybės pasidalijimo efektyvus valdymas išlieka ribotas.
10. Holistinis požiūris tampa vis aktualesnis. Organizacijos siekia apjungti valdymo, rizikų ir atitikties sritis į vieningą sistemą – GRC modelis čia tampa vis dažnesniu pasirinkimu.
11. Rizika vis dažniau suvokiama kaip planuojamas ir valdomas procesas, o ne tik grėsmė. Toks požiūris leidžia ją vertinti ir kaip galimybę sistemingam organizacijos tobulėjimui.
12. Ateities vizijos siejamos su didesniu automatizavimu, pažangiu duomenų valdymu ir rizikų draudimo mechanizmų integracija. Šios tendencijos rodo augantį organizacijų siekį veikti proaktyviai – ne tik reaguoti į iškilusias problemas, bet ir numatyti jas iš anksto.

REKOMENDACIJOS

1. Skatinti tarpdisciplininį bendradarbiavimą DI rizikų valdyme. Norint veiksmingai įvertinti su dirbtiniu intelektu susijusias rizikas, būtinas glaudus skirtingų sričių specialistų – technologų, teisininkų, etikos ekspertų ir strateginio valdymo atstovų - bendradarbiavimas. Tik toks integruotas požiūris leidžia rizikas suvokti ne vien techniniu lygmeniu, bet ir įvertinti jų vertybinius, organizacinius bei socialinius aspektus.
2. Įdiegti arba pritaikyti GRC modelį kaip DI rizikų valdymo pagrindą. GRC (Governance – Risk – Compliance) struktūra suteikia organizacijai galimybę nuosekliai susieti sprendimų priėmimą, rizikų analizę ir atitiktį teisiniams reikalavimams. Tai lankstus, tačiau kryptingas karkasas, pritaikomas prie skirtingų organizacinio brandos lygių.
3. Reguliariai peržiūrėti ir pritaikyti taikomus standartus atsižvelgiant į dirbtinio intelekto specifiką yra itin svarbu. Kad išliktų veiksmingos, organizacijos turėtų remtis naujesnėmis metodinėmis gairėmis, tokiomis kaip ISO/IEC 42001:2023, skirtomis DI valdymui. Tai padėtų išvengti automatizuoto, realios situacijos nebeatitinkančio standartų taikymo ir geriau prisitaikyti prie nuolat besikeičiančios technologinės bei teisinės aplinkos.
4. Skirti dėmesio darbuotojų švietimui ir vidinei komunikacijai apie rizikas yra esminė sąlyga veiksmingam rizikų valdymui. Ši sritis neturėtų būti traktuojama vien kaip vadovybės atsakomybė – kiekvienas darbuotojas turi gebėti atpažinti galimas grėsmes, suprasti jų svarbą ir tinkamai reaguoti. Tam būtini ne tik formaliai organizuojami mokymai, bet ir nuosekli, atvira vidaus komunikacijos kultūra, skatinanti sąmoningumą bei atsakomybę visais organizacijos lygmenimis.
5. Įtraukti socialinį ir etinį rizikų kontekstą į DI sprendimų kūrimą. Techniniai sprendimai būtini, tačiau jie negali būti vienintelė priemonė – būtina įvertinti, kokį poveikį DI sprendimai turės žmonėms, pasitikėjimui bei organizacijos reputacijai.

LITERATŪROS SĄRAŠAS

1. A.Balkevičius (2018) „Biudžeto rizikos Valdymas“
2. A.Sidorenko „Risk – Academy‘s Guide on ISO 31000 https://www.researchgate.net/publication/369916561 (žiūrėta 2025– 01– 10)
3. AI Risk Management Framework 2024 https://www.nist.gov/itl/ai– risk– management– framework
4. Anderson, H. S., Woodbridge, J., & Filar, B. (2018). DeepDGA: Adversarially– Tuned Domain Generation and Detection. Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security
5. Aven, T. (2016). Risk assessment and risk management: Review of recent advances on their foundation. European Journal of Operational Research, 253(1), 1– 13.
6. B.Biggio & F.Roli. (2018). Wild patterns: Ten years after the rise of adversarial machine learning.
7. Baesens, B. (2014). Analytics in a Big Data World: The Essential Guide to Data Science and its Applications. Wiley.
8. Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. First Monday, 21(11).
9. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency.
10. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
11. Bitinas, B., Rupšienė, L., & Žydžiūnaitė, V. (2008). Kokybinių tyrimų metodologija. Klaipėda: S. Jokužio leidykla– spaustuvė.
12. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.
13. Brown University Office of Information Technology. (2023). Beware AI– Enhanced Phishing Attempts. Skaityta internetu 2025– 03– 02 https://it.brown.edu/phish– bowl– alerts/beware– ai– enhanced– phishing– attempts it.brown.edu
14. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv preprint arXiv:1802.07228.
15. Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., & Amodei, D. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. arXiv preprint arXiv:2004.07213.
16. Brynjolfsson, E., & McAfee, A. (2017). Machine, Platform, Crowd: Harnessing Our Digital Future. W. W. Norton & Company.
17. Cadwalladr, C., & Graham– Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian.
18. CCPS (Center for Chemical Process Safety) 2018 Bow Ties in Risk Management ISBN: 9781119490388
19. Chesney, R., & Citron, D. K. (2019). Deepfakes and the new disinformation war: The coming age of post– truth geopolitics. Foreign Affairs, 98(1), 147– 155.
20. COSO ERM modelis (2004) „ Enterprise Risk Management – Integrated Framework“
21. D. Hillson, R Murray– Webster (2007) „Understanding and Managing Risk Attitude“
22. D. Nelsonas (2024 Unite.AI) „Kas yra Backpropagation?“ https://www.unite.ai/lt/kas– yra– backpropagation/ (žiūrėta 2025– 01– 09)
23. D.V. Carvalho, E.M. Pereira, & J.S. Cardoso (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832.

24. DENDRAL: A case study of the first expert system for scientific hypothesis formation R. K. Lindsay , B. G. Buchanan , E. A. Feigenbaum , J. Lederberg Volume 61, Issue 2, 1993, 209–261psl.
25. De Ruijter, F. Guldenmund (2016) „The bowtie method: A review“ 211– 218psl
26. European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).
27. EUROPOS PARLAMENTO IR TARYBOS DIREKTYVA (ES) 2022/2555 2022 m. gruodžio 14 d. dėl priemonių aukštam bendram kibernetinio saugumo lygiui visoje Sąjungoje užtikrinti, kuria iš dalies keičiamas Reglamentas (ES) Nr. 910/2014 ir Direktyva (ES) 2018/1972 ir panaikinama Direktyva (ES) 2016/114
28. Europos Parlamento ir Tarybos Reglamentas (ES) 2016/679
29. „EUROPOS PARLAMENTO IR TARYBOS REGLAMENTAS (ES) 2024/1689 2024 m. birželio 13 d. kuriuo nustatomos suderintos dirbtinio intelekto taisyklės.
30. Europos Parlamento ir Tarybos Reglamentas 2022/2065 2022 m. spalio 19 d. dėl bendrosios skaitmeninių paslaugų rinkos, kuriuo iš dalies keičiama Direktyva 2000/31/EB (Skaitmeninių paslaugų aktas)
31. Europos Parlamento ir Tarybos Reglamentas dėl nesutartinės civilinės atsakomybės taisyklių pritaikymo dirbtiniam intelektui (Atsakomybės už dirbtinį intelektą direktyva) (ES) 2022/0303
32. Europos Parlamento ir Tarybos Reglamentas(ES) 2022/1925 (2022) dėl atvirų konkurencijai ir sąžiningų skaitmeninio sektoriaus rinkų.
33. F.Knight (1921) „Risk, Uncertainity and Profit“
34. Federal Information Security Modernization Act (2014) https://www.cisa.gov/topics/cyber– threats– and– advisories/federal– information– security– modernization– act
35. Feigenbaum, E. A. (1977). The Art of Artificial Intelligence: Themes and Case Studies of Knowledge Engineering. IJCAI, 77, 1014–1029.
36. Ferrara, E. (2020). Disinformation and social bot operations in the run up to the 2020 US election. Harvard Kennedy School Misinformation Review.
37. Ferrara, E. (2020). Social Bots and Social Media Manipulation in 2020: The Year in Review. arXiv preprint arXiv:2102.08436.
38. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).
39. Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.
40. Fredrikson, M., et al. (2015). Model inversion attacks that exploit confidence information and basic countermeasures.
41. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. Prieiga per: http://www.deeplearningbook.org (žiūrėta 2025– 01– 09)
42. Google DeepMind (2024) „AlphaGo mastered the ancient game of Go, defeated a Go world champion, and inspired a new era of AI systems. https://deepmind.google/research/breakthroughs/alphago (žiūrėta 2025– 01– 02)
43. H.Markowithz (1952) „Portfolio Selection“ The Journal of Finance: Volume 7, Issue 1
44. Hitaj, B., Ateniese, G., & Pérez– Cruz, F. (2019). PassGAN: A Deep Learning Approach for Password Guessing. International Conference on Applied Cryptography and Network Security, 217– 237.
45. Hussain, W., Hussain, O. K., Chang, E., & Dillon, T. (2022). Risk management in AI– driven decision support systems: Challenges and future research directions. Artificial Intelligence Review, 55(4), 3291– 3315.

46. Hintze 2016 „Understanding the Four Types of Artificial Intelligence“ https://www.govtech.com/computing/understanding– the– four– types– of– artificial– intelligence.html (žiūrėta 2025– 01– 02)
47. IBM 2024 DEEPBLUE „IBM’s computer checkmated a human chess champion in a computing tour de force“ https://www.ibm.com/history/deep– blue (žiūrėta 2024– 01– 02)
48. IBM Data and AI Team 2024 „The four types of AI based on functionalities“ prieiga per internetą https://www.ibm.com/think/topics/artificial– intelligence– types (žiūrėta 2025– 01– 02)
49. International Organization for Standardization. (2018). ISO 31000:2018 Risk management – Guidelines. ISO.
50. International Organization for Standardization. (2022). ISO 27001: Information security management systems – Requirements.
51. International Organization for Standardization. (2023). ISO 42001: Artificial Intelligence Management System.
52. ISO/IEC 23894:202 (2023) Artificial Inteligence Guidance on risk management (žiūrėta internetu 2025– 03– 17 https://www.iso.org/standard/77304.html
53. ISO/IEC 42001 (2023) Information technology – Artificial intelligence – Management system https://www.iso.org/standard/81230.html
54. J. Neuman, O.Morgensten (1944) „Theory of games and economic behavoir“
55. K. P. Murphy (2012) „Machine learning a probabilistic perspective“
56. Kardelis, K. (2007). Mokslinių tyrimų metodologija ir metodai. Šiauliai: Lucilijus.
57. Lietuvos Dirbtinio Intelekto Strategija | Ateities Vizija (2018)
58. Lietuvos Respublikos Asmens Duomenų Teisinės Apsaugos Įstatymas Nr. I– 1374
59. Lietuvos Respublikos Informacinės Visuomenės Paslaugų Įstatymas 2006 m. gegužės 25 d. Nr. X– 614
60. Lietuvos Respublikos Kibernetinio Saugumo Įstatymas (2014)
61. M. T. Ribeiro, S. Singh, C. Guestrin – Machine Learning „Why Should I Trust You?": Explaining the Predictions of Any Classifier“
62. M.Douglas, A. Wildavsky (1983) Risk and Culture An Essay on the Selection of Technological and Environmental Dangers
63. Marian Rejewski Rankraščiai publikuoti tik 1980m. „An Application of the Theory of Permutations in Breaking the Enigma Cipher Applicaciones Mathematicae. 16, No. 4, Warsaw 1980.
64. McCarthy, J. (1956). Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Dartmouth College.
65. McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, 5(4), 115–133
66. Miller, J., Smith, R., & Robinson, T. (2020). AI– Driven Cyber Attacks on Financial Institutions: Assessing the Threat Landscape. Journal of Financial Crime, 27(1), 123– 137.
67. MYCIN artificial intelligence program B.J. Copeland „Encyclopedia Britannica“
68. N. Srivastava, G.Hinton, A.Krizhevsky, I.Sutskever & R.Salakhutdinov (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929– 1958.
69. N.Carlini, et al. (2021). Extracting Training Data from Large Language Models.
70. National Institute of Standards and Technology https://www.nist.gov/
71. National Institute of Standards and Technology. (2023). NIST AI Risk Management Framework. NIST.

72. NIST Risk Management Framework https://csrc.nist.gov/projects/risk– management/about–rmf
73. Kurakin (2017). „Adversarial examples in the physical world“.
74. P. Bromiley , M. McShane , A. Nair , E. Rustambekov 2014 „Enterprise Risk Management: Review, Critique, and Research Directions“
75. Paulauskaitė– Tarasevičienė, A., & Šutienė, K. (2020). Intelektikos pagrindai. KTU leidykla.
76. Poole, D., Mackworth, A., & Goebel, R. (1998). Computational Intelligence: A Logical Approach. Oxford University Press.
77. R. Shokri, et al. (2017). Membership inference attacks against machine learning models.
78. Rezoliucija Dėl Dirbtinio Intelekto Technologijų Naudojimo Viešajame Sektoriuje Principų 2024 Nr. Xiv– 2620
79. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408.
80. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back–propagating errors. Nature, 323, 533–536.
81. Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.
82. S. Kaplan, B. John Garrik 1981 “On the quantitative definition of risk” Risk Analysis journal 11– 27psl
83. Scott Mitchell GRC Capability Model™ 3.5 (OCEG™ Red Book)https://www.oceg.org/grc–capability– model– red– book/
84. Searle, J. R. (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences, 3(3), 417–424.
85. Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low– credibility content by social bots. Nature Communications, 9, 4787.
86. Smith, H., & McKeen, J. D. (2009). Developments in Practice XXXIII: A Holistic Approach to Managing IT– based Risk. Communications of the Association for Information Systems, 25, pp– pp. https://doi.org/10.17705/1CAIS.02541
87. T. Hastie, R. Tibshirani, J Friedman „The Elements of Statistical Learning Data Mining, Inference, and Prediction Second Edition“ (2017)
88. T.Miller, P.Howe & L.Sonenberg, L. (2020) Explainable AI: Understanding, trust, and control. Artificial Intelligence, 290, 103385.
89. Turing, A. (1950). Computing Machinery and Intelligence. Mind, 59(236), 433–460.
90. U.Beck (1992) „Risk Society towards a new modern“
91. V. Stasytytė, L.Aleksiene (2015) „Operational Risk Assessment and Management in Small and Medium– sized Enterprises“ DOI: 10.3846/btp.2015.568
92. W.H. Inmon (2002) „Building the data warehouse“ 35psl
93. Waymo (2024) „Self– driving cars – Autonomuos Vehicles“ https://waymo.com (žiūrėta 2025– 01– 02)
94. Wechsler, D. (1944). The Measurement of Adult Intelligence. Williams & Wilkins. – 3psl.
95. Wooldridge, M. (2009). An Introduction to MultiAgent Systems (2nd ed.). John Wiley & Sons.
96. Y. LeCun, Y.Bengio & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436– 444. DOI: 10.1038/nature14539
97. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending Against Neural Fake News. Advances in Neural Information Processing Systems (NeurIPS).
98. Žydžiūnaitė, V., & S.Sabaliauskas, S. (2017). Kokybiniai tyrimai: Principai ir metodai. Vilnius: Vaga.

MYKOLO ROMERIO UNIVERSITETAS
VIEŠOJO VALDYMO IR VERSLO FAKULTETAS
VERSLO IR EKONOMIKOS INSTITUTA

IVETA GASPARAVIČIENĖ
Kibernetinio saugumo valdymo magistrantūros studijų programa
Magistro baigiamasis darbas
Tema: „Dirbtinio Intelekto Rizikų Valdymas“
Vadovas: Prof. Dr. Mindaugas Kiškis
Vilnius, 2025

SANTRAUKA

Darbo tema – „Dirbtinio intelekto rizikų valdymas“. Pagrindinis tikslas – išanalizuoti, kaip Lietuvos organizacijos identifikuoja ir valdo su dirbtiniu intelektu (DI) susijusias rizikas, įvertinant, kiek taikomi sprendimai atitinka šiuolaikinius tarptautinius standartus ir galiojančius teisės aktus.

Dirbtinio intelekto technologijos sparčiai keičia organizacijų veiklos logiką – jos ne tik atveria naujas galimybes, bet ir iškelia sudėtingus klausimus, susijusius su sprendimų skaidrumu, atsakomybės ribomis, duomenų šališkumu bei etinėmis dilemomis. Šiame tyrime analizuojami pagrindiniai DI rizikų valdymo standartai ir modeliai – ISO/IEC 42001, ISO 27001, NIST AI RMF ir GRC struktūra – siekiant įvertinti jų sąveiką su Europos Sąjungos reguliavimo sistema (AI Act, GDPR, DSA) bei jų poveikį organizacijų vidaus kontrolės mechanizmams.

Teorinė analizė buvo derinama su kokybiniu tyrimu, grįstu pusiau struktūruotais interviu. Tyrimo dalyviai – ekspertai iš Lietuvos viešojo ir privataus sektoriaus – atskleidė, kad daugelyje organizacijų DI rizikų valdymas vis dar vyksta fragmentiškai. Standartai dažnai taikomi formaliai, neintegruojant jų į kasdienes praktikas. Tarp pagrindinių iššūkių išskirti sprendimų skaidrumo stoka, žmogiškojo faktoriaus sumažėjimas technologiniuose sprendimuose, ribota duomenų valdymo kokybė ir neaiškios etinės ribos.

Remiantis tyrimo rezultatais, siūlomas struktūruotas, tarpdisciplininis požiūris, apjungiantis technologinius, teisinius ir etinius komponentus į vieningą, praktikoje pritaikomą sistemą. Taip pat pateikiamos rekomendacijos, padedančios organizacijoms geriau prisitaikyti prie nuolat besikeičiančios technologijų ir reguliavimo aplinkos, didinant pasitikėjimą bei skatinant atsakingą DI diegimą.

Raktažodžiai: dirbtinis intelektas, ISO/IEC 42001, rizikų valdymas, GDPR, GRC, DSA

MYKOLAS ROMERIS UNIVERSITY
FACULTY OF PUBLIC GOVERNANCE AND BUSINESS
INSTITUTE OF BUSINESS AND ECONOMMICS

IVETA GASPARAVIČIENĖ

Master's Study Programme in Cybersecurity Management
Master's thesis
Title: "Artificial Intelligence Risk Management"
Supervisor: Prof. Dr. Mindaugas Kiškis
Vilnius, 2025

SUMMARY

The main objective of this thesis is to examine how Lithuanian organizations identify and manage risks associated with artificial intelligence (AI), assessing the extent to which their practices align with modern international standards and applicable legal regulations.

AI technologies are reshaping how organizations operate, offering significant opportunities while simultaneously introducing complex challenges—such as lack of transparency in decision-making, unclear lines of accountability, data bias, and unresolved ethical concerns. This study explores key AI risk management standards and models—ISO/IEC 42001, ISO 27001, the NIST AI Risk Management Framework, and the GRC model—focusing on how these frameworks interact with EU regulations (AI Act, GDPR, and DSA) and influence internal organizational controls.

The theoretical discussion is supported by qualitative research based on semi-structured interviews with experts from Lithuania's public and private sectors. The findings indicate that AI risk management remains fragmented across many organizations. Standards are frequently adopted at a surface level, without being meaningfully embedded in day-to-day operations. Key issues identified include limited decision-making transparency, reduced human involvement in automated processes, insufficient data governance, and ambiguous ethical parameters.

Drawing on these insights, the study proposes a structured, interdisciplinary approach that brings together technological, legal, and ethical perspectives into a cohesive and applicable framework. It also provides practical recommendations aimed at helping organizations adapt more effectively to the fast-changing technological and regulatory environment, while fostering trust and encouraging the responsible use of AI.

Keywords: artificial intelligence, ISO/IEC 42001, risk management, NIST AI RMF, GDPR, AI bias, GRC, DSA

PRIEDAI

1. Apklausos Anketa

Tema: Kvietimas dalyvauti tyrime apie DI rizikų valdymą organizacijose

Laba diena,

Mano vardas – Iveta Gašparavičienė, studijuoju kibernetinio saugumo valdymo magistrantūroje Mykolo Romerio universitete. Šiuo metu rengiu magistro baigiamąjį darbą, kuriame nagrinėju, kaip Lietuvos organizacijos praktinėje veikloje atpažįsta, vertina ir sprendžia rizikas, kylančias taikant dirbtinio intelekto technologijas.

Kviečiu Jus prisidėti prie šio tyrimo dalyvaujant pusiau struktūruotame interviu, kuriame būtų aptariami aktualūs klausimai apie DI rizikų valdymą. Pokalbis būtų konfidencialus – visa surinkta informacija bus naudojama tik moksliniams tikslams, o dalyvių tapatybės nebus viešinamos.

Interviu trukmė – apie 45 minutes, pasirinktu formatu (nuotoliniu būdu, telefonu ar susitinkant gyvai) ir Jums patogiu metu. Klausimus pridedu prie šio laiško, tad galėsite su jais susipažinti iš anksto. Jeigu sutiktumėte dalyvauti arba norėtumėte daugiau informacijos, mielai susisieksiu Jums patogiu laiku. Iš anksto nuoširdžiai dėkoju už Jūsų dėmesį ir skirtą laiką!

Pagarbiai, Iveta Gašparavičienė

INTERVIU KLAUSIMYNAS

Tyrimo tema: Dirbtinio intelekto rizikų valdymas organizacijose

1. Kokia jūsų organizacijos pagrindinė veiklos sritis? Kokius DI sprendimus jau naudojate arba šiuo metu diegiate?
2. Ar galėtumėte prisiminti konkretų atvejį, kai jūsų organizacijoje kilo su DI susijusi rizika? Kaip ši rizika buvo atpažinta?
3. Kaip vertinate savo dabartinius rizikų atpažinimo metodus – ar jie veiksmingi, ar turite abejonių dėl jų ribotumo?
4. Kokie kriterijai dažniausiai lemia, kurios DI rizikos vertinamos kaip svarbiausios? Kaip sprendžiate, kurioms teikti pirmenybę, ir kaip elgiatės su mažesnio prioriteto rizikomis?
5. Kai organizacijoje kyla su DI susijusi rizika – kas apie ją pirmasis sužino, kas inicijuoja veiksmus? Ar atsakomybės už rizikų valdymą aiškiai paskirstytos?

6. Kaip jūsų organizacija reaguoja į pasikeitimus reguliavime – ar turite mechanizmus, padedančius greitai prisitaikyti prie naujų standartų ar įstatymų?
7. Kurių teisės aktų ar standartų (pvz., DSA, GDPR, NIST, ISO) reikalavimai jums kelia daugiausia neaiškumo ar praktinių įgyvendinimo iššūkių?
8. Kurioje DI taikymo srityje jūsų organizacijoje, jūsų manymu, slypi didžiausia rizika? Kodėl būtent ši sritis? (pvz., klientų aptarnavimas, sprendimų automatizavimas, duomenų analizė)
9. Kaip vertinate dabartinius rizikų valdymo modelius ar sistemas, tokias kaip GRC ar ISO, NIST, COBIT ar kt. Ar esate apie jas girdėjęs?
10. Kurios DI rizikų valdymo situacijos jūsų organizacijoje šiuo metu kelia daugiausia neaiškumo ar reikalauja daugiausia pastangų?
11. Kaip manote, kaip artimiausiu metu keisis DI rizikų valdymo praktika jūsų srityje ar apskritai organizacijose?
12. Jei rytoj pradėtumėte kurti DI rizikų valdymo sistemą nuo nulio – nuo ko pradėtumėte, ką darytumėte kitaip nei dabar?