Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Skatinamojo mokymosi pritaikymas pasitikėjimo vertai rekomendacinei sistemai ir dinaminei kainodarai

Baigiamasis magistro studijų projektas

Žydrūnas Bautronis

Projekto autorius

Prof. dr. Robertas Alzbutas

Vadovas

Prof. dr. Vytautas Snieška

Vadovas

Kaunas, 2025

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Skatinamasis mokymasis pasitikėjimo vertai rekomendacinei sistemai ir dinaminei kainodarai

Baigiamasis magistro studijų projektas

Didžiųjų verslo duomenų analitika (6213AX001)

Žydrūnas Bautronis

Projekto autorius

Prof. dr. Robertas Alzbutas

Vadovas

Prof. dr. Vytautas Snieška

Vadovas

Doc. dr. Tomas Iešmantas

Recenzentas

Prof. dr. Aušra Rūtelionė

Recenzentė

Kaunas, 2025

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Žydrūnas Bautronis

Skatinamasis mokymasis pasitikėjimo vertai rekomendacinei sistemai ir dinaminei kainodarai

Akademinio sąžiningumo deklaracija

Patvirtinu, kad:

1. baigiamąjį projektą parengiau savarankiškai ir sąžiningai, nepažeisdama(s) kitų asmenų autoriaus ar kitų teisių, laikydamasi(s) Lietuvos Respublikos autorių teisių ir gretutinių teisių įstatymo nuostatų, Kauno technologijos universiteto (toliau – Universitetas) intelektinės nuosavybės valdymo ir perdavimo nuostatų bei Universiteto akademinės etikos kodekse nustatytų etikos reikalavimų;

2. baigiamajame projekte visi pateikti duomenys ir tyrimų rezultatai yra teisingi ir gauti teisėtai, nei viena šio projekto dalis nėra plagijuota nuo jokių spausdintinių ar elektroninių šaltinių, visos baigiamojo projekto tekste pateiktos citatos ir nuorodos yra nurodytos literatūros sąraše;

3. įstatymų nenumatytų piniginių sumų už baigiamąjį projektą ar jo dalis niekam nesu mokėjęs;

4. suprantu, kad išaiškėjus nesąžiningumo ar kitų asmenų teisių pažeidimo faktui, man bus taikomos akademinės nuobaudos pagal Universitete galiojančią tvarką ir būsiu pašalinta(s) iš Universiteto, o baigiamasis projektas gali būti pateiktas Akademinės etikos ir procedūrų kontrolieriaus tarnybai, nagrinėjant galimą akademinės etikos pažeidimą.

Žydrūnas Bautronis

Patvirtinta elektroniniu būdu

Bautronis, Žydrūnas. Skatinamasis mokymasis pasitikėjimo vertai rekomendacinei sistemai ir dinaminei kainodarai. Magistro studijų baigiamasis projektas / vadovas Prof. dr. Robertas Alzbutas , vadovas Prof. dr. Vytautas Snieška; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Taikomoji matematika (Matematikos mokslai).

Reikšminiai žodžiai: skatinamasis mokymasis, rekomendacinės sistemos, pasitikėjimo vertas DI, dinaminė kainodara, paaiškinamumas.

Kaunas, 2025. 70 p.

Santrauka

Per pastarąjį dešimtmetį dirbtinio intelekto (DI) plėtra e. komercijoje vis labiau rėmėsi skatinamuoju mokymusi (angl. *reinforcement learning*, RL), ypač, siekiant optimizuoti dinaminę kainodarą ir suasmenintas rekomendacijas. Tačiau realiose verslo situacijose kyla esminių iššūkių, susijusių su modelių skaidrumu, teisingumu ir sprendimų stabilumu – ypač kai algoritminiai sprendimai daro tiesioginę įtaką vartotojų patirčiai ir verslo rezultatams. Šiame tyrime pateikiama struktūruota metodika, kuri į RL pagrindu veikiančias kainodaros sistemas integruoja pasitikėjimo verto DI principus, orientuotus ne tik į našumą, bet ir į sprendimų paaiškinamumą bei socioekonominį teisingumą.

Naudojant realius transakcijų duomenis sukurta individualizuota RL aplinka, kurioje modeliuojama dinaminė kainodara, vartotojų paklausa, atsargų lygiai ir produktų kategorijos. Palyginti keli RL algoritmai (DQN, PPO, A2C), vertinant jų sprendimų priėmimo elgseną, mokymosi stabilumą ir atitikimą pasitikėjimo kriterijams. Aiškinamumui užtikrinti atlikta trajektorijų analizė bei SHAP (angl. *SHapley Additive exPlanations*) vertinimas, leidžiantys atsekti, kodėl RL agentai pasirinko konkrečius kainodaros veiksmus ir kaip šie dera su rinkos logika bei etikos principais.

Rezultatai parodė, kad dalis modelių rekomendavo trumpalaikį atlygio maksimizavimą, o kiti (pvz., DQN) taikė prisitaikančias, riziką įvertinančias strategijas, vienu metu optimizuodami pelningumą, atsargų valdymą ir paklausos elastingumą. Pasiūlytas DQN sprendimas bendrą pelną padidino 12,58 %, palyginti su bazine kainodaros strategija (4,71 mln. € prieš 4,18 mln. €), išlaikydamas nulinį neetiško kainų kėlimo atvejų rodiklį esant žemai paklausai. SHAP analizė atskleidė, kad didžiausią įtaką veiksmams turėjo sandėlio užimtumo lygis, paklausos pokytis ir produkto elastingumas, o neigiamus sprendimus lėmė atsargų kaupimas ir kainų kėlimas nepalankiomis sąlygomis.

Apjungdamas tradicinius RL našumo rodiklius su aiškinamumo ir teisingumo vertinimais, tyrimas patvirtina, jog pasitikėjimu grįsti DI principai gali lemti atsakingus, skaidrius ir verslo tikslus atitinkančius sprendimus. Siūloma metodika sudaro prielaidas patikimus RL agentus saugiai diegti ne tik dinaminės kainodaros, bet ir kitose e. komercijos srityse, kur reikalingas etiškas ir išorinį reguliavimą atitinkantis automatizuotas sprendimų priėmimas.

Bautronis, Žydrūnas. Reinforcement Learning for Trustworthy Recommendation System and Dynamic Pricing. Master's Final Degree Project / supervisor Prof. dr. Robertas Alzbutas, supervisor Prof. dr. Vytautas Snieška; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Study field and area (study field group): Applied Mathematics (Mathematical Sciences).

Keywords: Reinforcement learning, trustworthy AI, recommender systems, dynamic pricing, explainability.

Kaunas, 2025. 70 pages.

Summary

Over the past decade, the growth of Artificial Intelligence (AI) in e-commerce has increasingly relied on Reinforcement Learning (RL), especially, to optimize dynamic pricing and personalized recommendations. Yet real-world deployments face critical challenges linked to model transparency, fairness, and decision stability, especially when algorithmic choices directly influence users' experiences and business outcomes. This study presents a structured framework that explicitly embeds trustworthy AI principles into RL-based pricing systems, evaluating not only performance but also interpretability and socio-economic fairness.

Using real transactional data, we built a bespoke RL environment that simulates dynamic pricing, consumer demand, inventory levels, and product categories. Several RL algorithms (DQN, PPO, A2C) were compared in terms of decision-making behavior, learning stability, and compliance with trustworthiness criteria. To enhance explainability, we conducted trajectory audits and SHapley Additive exPlanations (SHAP) analyses, tracing why RL agents chose specific prices and how those choices aligned with market logic and ethical constraints.

Results show that while some models focused recommendations on short-term reward maximization, others—most notably the DQN agent—adopted more adaptive, risk-aware strategies that balanced profitability with inventory management and demand elasticity. The proposed DQN approach achieved a 12.58 % profit increase over the baseline pricing strategy (raising total profit from €4.18 million to €4.71 million) while recording zero unethical price hikes under low demand. SHAP analysis revealed that stock levels, demand shifts, and product elasticity had the largest positive impact on actions, whereas inventory hoarding and price increases in unfavorable conditions contributed negatively.

By combining conventional RL performance metrics with explainability and fairness assessments, this research demonstrates that integrating trustworthy-AI principles into agent design yields responsible, business-aligned outcomes. The proposed methodology offers a practical blueprint for safely deploying reliable RL agents not only in dynamic pricing but also across other e-commerce domains where ethical and regulatory compliance is paramount.

Turinys

Lentelių sąrašas **8**
Paveikslų sąrašas **9**
Santrumpų ir terminų sąrašas **10**
Įvadas **11**
1. Literatūros apžvalga **13**
1.1. Pasitikėjimo verto dirbtinio intelekto svarba šiuolaikiniams verslo sprendimams 13
1.1.1. Pasitikėjimo vertas dirbtinis intelektas 13
1.1.2. Dirbtinio intelekto vaidmuo sprendimų priėmime 15
1.1.3. Dirbtinio intelekto poveikis įmonės konkurencingumui 16
1.2. Sprendimų priėmimas ir patikimo dirbtinio intelekto integracija 17
1.2.1. Sprendimų priėmimas ir jo rūšys 17
1.2.2. Sprendimų priėmimui įtaką darantys veiksniai ir jų svarba organizacijose 19
1.2.3. DI privalumai sprendimo priėmime 19
1.3. Pasitikėjimo vertos rekomendacinės sistemos paremtos RL koncepcija 20
1.3.1. Duomenų kokybė ir prieinamumas 20
1.3.2. Dirbtinio intelekto paaiškinamumas 21
1.3.3. Etiniai/teisiniai klausimai ir atsakomybių pasidalijimas 22
1.4. Rekomendacinės sistemos koncepcija adaptuojant DI įmonės viduje 23
1.5. Dinaminė kainodara: skatinamojo mokymosi taikymas ir lyginamoji analizė 25
1.6. Patikimumo iššūkiai skatinamojo mokymosi modeliuose 26
1.6.1. Metodai didinantys RL paaiškinamumą, sąžiningumą ir stabilumą 27
2. Tyrimo metodai **29**
2.1. Markovo sprendimų procesas 29
2.1.1. Markovo sprendimų proceso samprata ir taikymas 29
2.1.2. Markovo sprendimų proceso komponentės 30
2.2. Skatinamasis mokymasis 31
2.2.1. Daugialypio lošimų automato problema 32
2.2.2. Tyrinėjimo ir išnaudojimo dilema 32
2.2.3. Vertės funkcija 33
2.2.4. Tikslo funkcija skatinamajame mokyme 33
2.3. Skatinamojo mokymosi algoritmai ir jų klasifikacija 34
2.3.1. Laiko skirtumo metodai 35
2.3.2. Strategijomis pagrįsti metodai 36
2.3.3. Aktoriaus-kritiko metodai 37
2.3.4. Giliojo Q-mokymosi tinklai (DQN) 38
2.4. Kiti naudoti metodai 39
2.4.1. Pagrindinių komponenčių analizė 39
2.4.2. Kryžminės validacijos metodas 40
2.5. Patikimumą didinantys sprendimų priėmimo modeliai grįsti dirbtiniu intelektu 40
2.6. Skatinamojo mokymosi bibliotekos pasirinkimas 41
2.6.1. *Stable baseline3* skatinamojo mokymosi biblioteka 42
2.6.2. *Stable baseline* algoritmai 43
3. Tyrimo rezultatai **45**
3.1. Pasirinkto duomenų rinkinio apžvalga 45
3.2. Duomenų paruošimas darbui 46
3.3. Skatinamojo mokymosi aplinkos kūrimas 48
3.4. Pasitikėjimo verto algoritmo paieška ir modelio apmokymas 49
3.5. Modelio optimizavimas ir geriausių hiperparametrų paieška 52

3.6. Dinaminės kainodaros pritaikius RL rezultatai 53
3.6.1. Dinaminė kainodara skirtingoms kategorijoms 53
3.6.2. Skatinamojo mokymosi agento elgsenos analizė 54
3.6.3. Skirtingų kainodaros strategijų palyginimas 59
3.7. Pasitikėjimo vertos rekomendacinės sistemos tikrinimas 60
3.8. Modelio stabilumo įvertinimas naudojant kryžminę validaciją 62
3.9. Pasitikėjimo verto rekomendacinės sistemos koncepcija 63
Išvados 65
Literatūros sąrašas 66
Priedai 71
1 priedas. Tyrimo rezultatų sklaida 71
2 priedas. „PricingEnv“ kodo fragmentas (Python) 72

Lentelių sąrašas

1 lentelė. Mokslinių tyrimų analizė skirta RL pritaikymui dinaminėje kainodaroje 26
2 lentelė. Pasirinktų Stable baseline 3 algoritmų palyginimas. 44
3 lentelė. Hiperparametrų rinkinys kiekvienam algoritmui. 49
4 lentelė. Geriausių hiperparametrų rinkinys 52
5 lentelė. Tyrimo rezultatų sklaida 71

Paveikslų sąrašas

1 pav. Pasitikėjimo verto dirbtinio intelekto apibrėžimas 14
2 pav. Sprendimų priėmimo etapai 18
3 pav. Duomenų kokybės rodikliai 21
4 pav. Dirbtinio intelekto apibrėžimas 24
5 pav. Statinės ir dinaminės kainodaros palyginimas 25
6 pav. Markovo sprendimo proceso pavyzdys 30
7 pav. Giluminio skatinamojo mokymosi pavyzdys 31
8 pav. Skatinamojo mokymosi algoritmų klasifikacija 34
9 pav. Q-mokymosi algoritmas 36
10 pav. Strategija pagrįsto algoritmo pavyzdys [63]. 37
11 pav. Aktoriaus-kritiko algoritmo pavyzdys [63]. 37
12 pav. DQN algoritmo pavyzdys [63]. 38
13 pav. 4 dalių kryžminė validacija 40
14 pav. Organizacijos sprendimo priėmimo modeliai paremti dirbtiniu intelektu. 41
15 pav. Skatinamojo mokymosi bibliotekų palyginimas 42
16 pav. Elektroninės parduotuvės duomenų modelis. 45
17 pav. Duomenų paruošimo darbui eiga. 46
18 pav. TOP 5 kategorijos pagal vidutinį pelningumą 47
19 pav. Gauti grafikai su TensorBoard kiekvienam modeliui 50
20 pav. Pasitikėjimo verto modelio parinkimas. 51
21 pav. Modelio hiperparametrų paieška. 52
22 pav. Dinaminė kainodara statybų ir elektronikos prekėms. 53
23 pav. Paklausa prieš kainos keitimą. 55
24 pav. Prekės kainos elastingumas pagal veiksmą. 55
25 pav. Paklausos palyginimas su sandėlio likučiu. 56
26 pav. Kainų keitimo veiksmų pasiskirstymas. 57
27 pav. PCA klasterizacija 58
28 pav. Kainos pokytis per laikotarpį. 59
29 pav. Skirtingų kainodarų palyginimas 60
30 pav. Atlygio komponenčių koreliacija su galutiniu atlygiu. 61
31 pav. SHAP analizė 62
32 pav. Modelio stabilumo įvertinimas. 63
33 pav. Pasitikėjimo vertos rekomendacinės sistemos koncepcija. 63

Santrumpų ir terminų sąrašas

Santrumpos:

DVDA – Didžiųjų verslo duomenų analitika.

GDI – Generatyvinis dirbtinis intelektas

DI – Dirbtinis intelektas

ŽI – Žmogiškasis intelektas

PDI – Patikimas dirbtinis intelektas

RL – Skatinamasis mokymasis

DQN- Giliojo Q-mokymosi tinklas

PPO – Proksimalus strategijos optimizavimas

A2C – Aktoriaus-kritiko metodas

SB3 – "Stable Baselines3" biblioteka

SHAP – Shapley reikšmių aiškinimas

Terminai:

Pasitikėjimo vertas – terminas, naudojamas apibūdinti sistemą ar technologiją, kuriai galima patikėti sprendimų priėmimą dėl jos skaidrumo, aiškinamumo, patikimumo ir etinių standartų laikymosi.

Skatinamasis mokymasis – mašininio mokymosi paradigma, kurioje agentas mokosi veiksmų seka optimizuoti atlygio funkciją, kuriuo formaliai aprašoma RL aplinka.

Rekomendacinė sistema – algoritmų ir metodų visuma, skirta analizuoti vartotojų elgseną bei sąveika su sistema tam, kad būtų prognozuojama ir pateikiama asmeniškai aktuali informacija – produktai, paslaugos ar turinys. Tokios sistemos dažnai naudojamos komercijoje ar socialiniuose tinkluose, siekiant pagerinti vartotojo patirtį ir padidinti įsitraukimą.

Hiperparametras – modelio ar mokymo proceso parametras, kuris nenustatomas mokymo metu, bet pasirenkamas eksperimentiškai (pvz., mokymosi greitis, sluoksnių skaičius).

Denormalizacija - duomenų bazės struktūros pertvarkymas, kai kelios lentelės jungiamos į vieną, siekiant optimizuoti užklausų našumą, dažniausiai aukojant perteklinį duomenų saugojimą.

Įvadas

Skaitmeninės ekonomikos plėtra ir didžiųjų duomenų prieinamumas lėmė augantį poreikį automatizuotoms sprendimų priėmimo sistemoms, kurios gebėtų greitai prisitaikyti prie dinamiškos rinkos aplinkos ir vartotojų elgsenos. Viena perspektyviausių technologijų šiam tikslui – skatinamasis mokymasis (angl. *reinforcement learning*, RL) leidžiantis modeliui savarankiškai mokytis iš sąveikos su aplinka ir optimizuoti ilgalaikę naudą. RL jau pasiekė įspūdingų rezultatų valdomųjų žaidimų (AlphaGo, OpenAI Five) bei automatizacijos srityse, tačiau verslo praktikose, pavyzdžiui, dinaminėje kainodaroje ar personalizuotose rekomendacinėse sistemose, iškyla papildomi reikalavimai – sprendimų paaiškinamumas, skaidrumas ir šališkumas.

Šių reikalavimų nepaisymas gali kelti finansinę riziką, siaurinti vartotojų pasitikėjimą, pažeisti reguliacinius reikalavimus (ES AI akto nuostatos) ir išprovokuoti nepageidaujamą šališkumą. Todėl pramonėje vis dažniau keliami klausimai: kaip padaryti RL modelius paaiškinamus? Kokių mechanizmų reikia, kad algoritmas išlaikytų sprendimų stabilumą ir elgtųsi šališkai, net kai optimizuoja pelną?

Dabartiniai RL algoritmai dažniausiai kuriami orientuojantis į grąžos optimizavimą (pvz., pelno, CTR ar konversijų). Tokia orientacija iš prigimties skatina „juodosios dėžės“ sprendimus: modelis pateikia veiksmą, bet jo motyvacija lieka neaiški. To neužtenka realiose verslo aplikacijose, kur būtina atsekti, kodėl modelis pasirinko konkrečią kainą ar pasiūlymą, ir įrodyti, kad sprendimas nepažeidžia vartotojo teisių bei bendrųjų sąžiningumo principų. Taigi pagrindinė mokslinė problema – sukurti RL sistemą, kuri ne tik optimizuotų veiklos rodiklius, bet ir patenkintų pasitikėjimo vertos kriterijus: paaiškinamumą, skaidrumą, sąžiningumą bei sprendimo stabilumą.

Šiame darbe pirmą kartą pasiūloma integruota metodologija, kuri viename modelyje sujungia pažangius skatinamojo mokymosi algoritmus (DQN, PPO, A2C), lokalaus ir agreguoto paaiškinamumo technikas (trajektorijų auditą, SHAP analizę) bei sąžiningumo ir stabilumo metrikas, pritaikytas dinaminės kainodaros ir rekomendacijų uždaviniams. Skirtingai nei ankstesniuose tyrimuose, kuriuose RL vertinamas daugiausia pagal pelno ar konversijų augimą, čia pateikiamas daugiasluoksnis vertinimo modelis: veiklos rodikliai analizuojami kartu su sprendimų paaiškinamumo, šališkumo, reputacinės rizikos ir reguliacinio atitikimo kriterijais. Be to, sukuriama realistiška simuliacinė aplinka, kurioje vartotojų elgsena modeliuojama per paklausos elastingumą ir sezoniškumą, taip suteikiant galimybę saugiai „ištestuoti“ patikimus RL sprendimus dar prieš jų diegimą verslo praktikoje. Tokia holistinė prieiga leidžia ne tik pasiekti verslo rodiklių pagerėjimą, bet ir įrodyti, kad RL agentai gali būti šališki, skaidrūs ir reguliacinių reikalavimų atitinkantys verslo pagalbininkai.

Praktiniu požiūriu ši metodologija sudaro prielaidas įmonėms diegti RL sprendimus be papildomo „etikos mokesčio“ – t. y. nereikia rinktis tarp pelno ir skaidrumo. Akademine prasme darbas plečia pasitikėjimo verto DI (angl. *trustworthy, AI*) diskursą, parodydamas, kad skaidrumas ir sąžiningumas gali būti ne išorinės „prilipdomos“ savybės, o organiška RL sistemos architektūros dalis.

Tyrimo objektas – pasitikėjimo verti skatinamojo mokymosi modeliai, taikomi verslo problemoms spręsti, ypatingą dėmesį skiriant skaidrumui, šališkumui, interpretacijai ir stabilumui.

Tyrimo tikslas – sukurti pasitikėjimo verto skatinamojo mokymosi modelį, skirtą dinaminei kainodarai ir personalizuotai rekomendacijų sistemai, kuris sąlygotų sprendimų skaidrumą, sąžiningumą ir stabilumą bei sumažintų šališkumą ir neapibrėžtumą priimant sprendimus.

Darbo uždaviniai:

1. išanalizuoti skatinamojo mokymosi metodų taikymą dinaminei kainodarai ir rekomendacinėms sistemoms, įvertinant Lietuvoje bei užsienyje atliktus tyrimus, taikomus matematinius metodus ir programines priemones bei suformuluojant reikalavimus modeliams ir algoritmams.
2. sukurti konceptualų ir matematinį modelį, pagrįstą pastiprinamojo mokymosi metodais, užtikrinant jo teisingumą, skaidrumą ir paaiškinamumą verslo kontekste bei parengti programinius sprendimus dinaminei kainodarai ir rekomendacijų sistemai generuoti.
3. atlikti eksperimentus su skatinamojo mokymosi modeliais, vertinant jų veikimą pagal patikimo dirbtinio intelekto koncepcijos pagrindinius aspektus. Taip pat patvirtinti rezultatų validumą tinkamais patikros metodais.
4. išanalizuoti gautus rezultatus ir pateikti išvadas, apibendrinant eksperimentų duomenis, įvertinant sukurtų modelių veikimą bei parengiant rekomendacijas dėl jų tobulinimo ir praktinio pritaikymo versle.

1. Literatūros apžvalga

Šiuolaikinėje verslo aplinkoje įmonių gebėjimas kaupti, analizuoti ir tikslingai taikyti informaciją tapo vienu iš pagrindinių konkurencinio pranašumo šaltinių. Didėjantis duomenų srautas, spartėjantis skaitmenizavimas ir pažangios technologijos leidžia organizacijoms geriau pažinti savo klientus, optimizuoti operacinius procesus bei priimti pagrįstus strateginius sprendimus dinamiškomis ir neapibrėžtomis sąlygomis. Dirbtinis intelektas (DI), kaip viena iš sparčiausiai besivystančių technologinių krypčių, suteikia naujas galimybes sprendimų priėmimo procesų tobulinimui, leidžiant įmonėms didinti veiklos efektyvumą, inovatyvumą ir kurti ilgalaikę pridėtinę vertę.

Pasitelkus pažangias duomenų analitikos priemones ir mašininio mokymosi metodus, organizacijos gali automatizuoti kompleksinius sprendimų priėmimo procesus, sumažinti žmogiškųjų klaidų riziką ir greičiau reaguoti į rinkos pokyčius. Skaitmeninė analitika, sentimentų analizė, didžiųjų duomenų apdorojimas bei rekomendacinės sistemos leidžia įmonėms ne tik geriau suprasti vartotojų elgseną, bet ir iniciatyviai numatyti jų poreikius bei personalizuoti pasiūlymus ir stiprinti klientų lojalumą.

Technologijų diegimas sudaro sąlygas ne tik vidinių procesų optimizavimui – nuo tiekimo grandinių valdymo iki dinaminės kainodaros sprendimų taikymo, bet ir prisideda prie naujų verslo modelių kūrimo bei inovacijų skatinimo. Tarp pažangiausių metodų, pritraukiančių vis daugiau dėmesio, išsiskiria skatinamasis mokymasis, leidžiantis agentams mokytis optimalios veiklos dinamiškai besikeičiančiose aplinkose. Tuo pačiu auga susidomėjimas patikimo skatinamojo mokymosi (angl. *trustworthy reinforcement learning,* TRL) metodikomis, kurios siekia užtikrinti sprendimų skaidrumą, stabilumą ir etiškumą.

Tačiau kartu su augančiomis galimybėmis kyla ir nauji iššūkiai: sprendimų interpretacijos sudėtingumas, duomenų privatumo užtikrinimas, algoritmų šališkumo mažinimas ir poreikis užtikrinti etišką sprendimų priėmimą. Šie klausimai tampa ypač aktualūs, kai DI sprendimai tiesiogiai veikia vartotojų patirtis ar finansinius rezultatus.

Atsižvelgiant į šiuos aspektus, dirbtinio intelekto ir skatinamojo mokymosi metodų taikymas tampa neatsiejama šiuolaikinių organizacijų strategijos dalimi, siekiant ne tik trumpalaikių veiklos rezultatų, bet ir ilgalaikės, tvarios plėtros konkurencingoje globalioje rinkoje.

1.1. Pasitikėjimo verto dirbtinio intelekto svarba šiuolaikiniams verslo sprendimams

1.1.1. Pasitikėjimo vertas dirbtinis intelektas

Dirbtinio intelekto sąvoka apima technologijų ir metodų rinkinį, leidžiantį sistemoms atlikti užduotis, kurios tradiciškai siejamos su žmogaus intelektu – tokias kaip mokymasis, problemų sprendimas, sprendimų priėmimas ar kalbos supratimas. Russel'as ir Norvig'as DI apibrėžia kaip „žmogaus kognityvinių funkcijų simuliacija naudojant intelektualius agentus“ [1], o Kaplan'o ir Haenleins'o teigimu, „...yra laikoma sistema, gebanti teisingai interpretuoti išorinius duomenis, mokytis iš jų ir panaudoti mokymo metu įgytą informaciją, siekiant konkrečių tikslų ar užduočių” [2]. Tuo tarpu, Duan'as, Edwards'as ir Dwivedi's pažymi, kad modernios DI sistemos geba veikti su tiek struktūrizuotais, tiek nestruktūrizuotais duomenimis, adaptuodamos savo elgseną realiuoju laiku ir taip optimizuodamos rezultatus [2].

DI technologijų proveržį paskatino spartus skaičiavimo galios augimas, duomenų prieinamumo plėtra bei pažangesni algoritmai [3]. Šiuolaikinis DI grindžiamas keturiomis kertinėmis kryptimis: mašininiu mokymusi (angl. ML), giliuoju mokymusi (angl. *deep learning,* DL), natūralios kalbos apdorojimu (angl. *natural language processing,* NLP) ir kompiuterine rega (angl. *computer vision,* CV). Generatyviosios DI technologijos (GDI), tokios kaip *GPT* ar *DALL-E*, žymi naują etapą, kai DI geba ne tik interpretuoti duomenis, bet ir kurti originalų turinį [4].

1 pav. Pasitikėjimo verto dirbtinio intelekto apibrėžimas

Tačiau kartu su šiuo potencialu kyla klausimas – ar DI sistemomis galima pasitikėti? Pasitikėjimo vertas DI reiškia ne tik techninį tikslumą ar našumą, bet ir etišką, paaiškinamą bei socialiai atsakingą veikimą (žr. 1 pav.) [5] . Šiuo atžvilgiu svarbiausi aspektai yra:

- Paaiškinamumas (angl. *explainability*) – galimybė suprasti, kaip ir kodėl DI sistema priėmė konkretų sprendimą. Tai itin svarbu, kai DI taikomas tokiose srityse kaip finansai, medicina ar teisė, kur neteisingas sprendimas gali turėti rimtų pasekmių.
- Skaidrumas ir atskaitomybė – verslo subjektai, diegiantys DI, turi ne tik žinoti, kaip sistema veikia, bet ir būti pasiruošę atsakyti už jos sprendimus. Todėl reikalingi mechanizmai, leidžiantys audituoti ir stebėti DI veiklą [7].
- Šališkumo mažinimas – DI modeliai gali paveldėti šališkumus iš treniravimo duomenų. Tokie šališki sprendimai gali sukelti diskriminaciją arba nelygybę, todėl būtina naudoti metodus, identifikuojančius ir mažinančius šališkumą [9].
- Saugumas ir duomenų privatumas – DI turi būti diegiamas taip, kad būtų užtikrintas jautrios informacijos konfidencialumas, o sprendimai nediskriminuotų vartotojų netiesioginiu būdu.
- Duomenų kokybė ir pasitikėjimas duomenimis – patikimas DI yra neatsiejamas nuo patikimų duomenų. Prasti ar neišsamūs duomenys gali sukelti klaidingus rezultatus, todėl būtinas dėmesys duomenų rinkimo, valymo, stebėsenos ir atnaujinimo kokybei [8, 9].

Nepaisant šių iššūkių, daugelis mokslinių tyrimų ir praktinių pavyzdžių rodo, kad tinkamai suprojektuoti ir valdomi DI sprendimai gali būti ne tik našūs, bet ir patikimi. Verslo kontekste

pasitikėjimu pagrįstas DI tampa esminiu elementu srityse, kur būtinas sprendimų objektyvumas – pavyzdžiui, rekomendacinėse sistemose, rizikų analizėje ar dinaminėje kainodaroje [7].

1.1.2. Dirbtinio intelekto vaidmuo sprendimų priėmime

Dirbtinis intelektas tapo esminiu šiuolaikinės verslo aplinkos elementu, kuris nuolat transformuoja įvairius veiklos procesus. Istoriškai tradicinės DI sistemos, tokios kaip neuroniniai tinklai ir mašininio mokymosi algoritmai, buvo naudojamos automatizuoti daugelį pasikartojančių ir laiko reikalaujančių užduočių. Tai ypač svarbu tokiose srityse kaip gamyba, logistikos valdymas, apskaita ir klientų aptarnavimas. Pavyzdžiui, robotizuota procesų automatizacija (RPA), pagrįsta DI, leidžia automatizuoti duomenų įvedimą, užsakymų apdorojimą ir kitus rutininius veiksmus taip sumažinant klaidų skaičių ir darbo apimtį [10]. Tačiau šios sistemos dažnai susidurdavo su iššūkiais, susijusiais su tiesiogine žmogaus sąveika ir prisitaikymu, ribojančiais jų apimtį ir efektyvumą sudėtingose situacijose, dėl ko tradicinės DI sistemos negalėjo efektyviai veikti.

Naujausios DI technologijos, tokios kaip *GDI*, pavyzdžiui, *ChatGPT*, keičia šią situaciją. Jos suteikia galimybę įgyvendinti natūralų žmogaus ir DI bendradarbiavimą, apdorojant ir interpretuojant ne tik tekstinius, bet ir vizualinius bei garso duomenis [7]. Tai ypač aktualu šiuolaikinėse organizacijose, kur reikia apdoroti didelius duomenų kiekius, tačiau tai daroma ribotais ištekliais. GDI gali veikti kaip neutralus sprendimų priėmimo tarpininkas, padedantis priimti labiau pagrįstus ir objektyvius sprendimus [4]. GDI generuojamos įžvalgos taip pat gali padėti organizacijoms greičiau prisitaikyti prie rinkos pokyčių ir išvengti potencialių nuostolių.

DI naudojimas versle yra ne tik technologinė naujovė, bet ir strateginė būtinybė. Šiuolaikiniai verslo lyderiai mato DI kaip įrankį, kuris gali padėti ne tik automatizuoti procesus, bet ir iš esmės pakeisti sprendimų priėmimo kultūrą. DI technologijos leidžia įmonėms greičiau prisitaikyti prie kintančių rinkos sąlygų, konkurencinio spaudimo ir technologinių pokyčių, suteikiant daugiau galimybių inovacijoms ir konkurenciniam pranašumui [11]. Viena to priežasčių yra ta, kad DI padeda geriau valdyti rizikas ir priimti informuotus sprendimus, kurie gali sumažinti neigiamą ekonominių, technologinių ar politinių pokyčių poveikį. Pavyzdžiui, DI modeliai gali analizuoti pasaulines rinkos tendencijas ir padėti įmonėms greitai reaguoti į pasikeitimus, užtikrinant veiklos tęstinumą ir stabilumą [7].

Mokslo bendruomenėje nuomonės apie DI išsiskiria. Kai kurie mokslininkai pabrėžia, jog duomenimis paremtu analitiniu mąstymu DI gali pakeisti žmogaus intuiciją ir patirtį ir leidžia sukurti naujus sprendimų priėmimo modelius, pritaikytus sudėtingoms verslo aplinkybėms [13], tačiau yra ir skeptikų, kurie teigia, kad DI naudojimas kelia įvairius iššūkius, ypač susijusius su duomenų kokybe, privatumu ir saugumu [13]. Šie aspektai turi būti sprendžiami taip, kad DI sprendimai būtų ne tik efektyvūs, bet ir patikimi bei etiškai atsakingi.

Atsižvelgiant į DI teikiamą naudą verslo sprendimų priėmimo procesuose, svarbu kurti struktūras ir metodikas, užtikrinančias atsakingą ir efektyvų DI naudojimą: tai apima ne tik technologijų diegimą, bet ir organizacinės kultūros keitimą, kuriame DI technologijos yra integruojamos į kasdienes verslo praktikas, užtikrinant optimalų rezultatų pasiekimą.

1.1.3. Dirbtinio intelekto poveikis įmonės konkurencingumui

Dirbtinio intelekto integravimas į verslo sprendimų priėmimo procesus yra vienas iš reikšmingiausių šiuolaikinės organizacijų konkurencingumą didinančių veiksnių. Dabartinėje sparčiai kintančioje verslo aplinkoje įmonės, sugebančios efektyviai panaudoti DI, įgyja pranašumą, kuris leidžia ne tik greičiau reaguoti į rinkos pokyčius, bet ir priimti labiau pagrįstus, duomenimis paremtus sprendimus. Šis pranašumas pasireiškia įvairiais aspektais. Pirma, DI gali pagerinti verslo operacijas ir padidinti efektyvumą, padėdamas pagerinti produkto kokybę ar pristatymo laikus, o tai gali teigiamai paveikti klientų pasitenkinimą ir lojalumą, kas gali įmonei suteikti konkurencinį pranašumą ir leisti padidinti verslo rinkos dalį [14]. Pavyzdžiui, „IBM“ *Watson* gali padėti mažoms ir vidutinėms įmonėms kurti asmeninius turinius rinkodarai ir didinti turinio pasiskirstymą per turimus kanalus.

Kitas svarbus taikymo aspektas - DI gali padėti verslams geriau analizuoti klientus ir juos klasifikuoti pagal jų poreikius ir įpročius, o tai leistų personalizuoti konsultavimo ar kitas paslaugas pagal kiekvieną klientą ar pirkėją. Toks sprendimas leistų sumažinti klientų išlaikymo kaštus ir daryti įtaką tiek naujų tiek buvusių klientų prisitraukimui [15]. Atsižvelgus į poreikius, bendrovė „Salesforce“ šiuo metu kuria naują DI sprendimą *Einstein*. Remiantis prognozine analize, šis įrankis gali pastebėti įžvalgas įmonės duomenyse, leidžiančias pasiūlyti asmeninius pasiūlymus ir veiksmus įmonėms.

Tarp kitų privalumų galima išskirti, jog DI gali padėti verslams atrasti naujas pajamų augimo galimybes, pagerinant rinkodaros strategijas ar sukuriant naujus produktus. Šios naujos galimybės leistų sumažinti išlaidas ir padidinti pajamas, o tai pagerintų pelningumą ir finansinius veiklos rezultatus [16, 17]. Taip pat yra tyrimų, rodančių, kad dirbtinis intelektas gali padėti verslams tiksliau prognozuoti ateities pajamų srautus ir nustatyti potencialius rizikos veiksnius, o tai leidžia geriau valdyti pinigų srautus ir sustiprinti įmonės likvidumą [18].

Verta paminėti, kad esant sudėtingomis įmonės transakcijomis, DI intelekto panaudojimas gali turėti neigiamą poveikį verslo veiklai [19]. Būtent dėl šios priežasties įmonės susiduria su sudėtingais ir personalizuotais produktais, kurie reikalauja aukšto lygio žmogiškosios ekspertizės ir sąveikos, kad būtų galima suprasti ir patenkinti klientų poreikius. Dėl to dirbtinio intelekto technologijos gali nepajėgti visiškai suprasti klientų poreikių ar suteikti to paties lygio pritaikytų paslaugų kaip žmogiškieji pardavimų atstovai [20]. Ir tai gali lemti mažesnį klientų pasitenkinimą ir išlaikymą, darant neigiamą poveikį verslo veiklai. Be to, dirbtinio intelekto technologijų diegimas įmonėms gali būti brangus ir laikui imlus procesas, reikalaujantis didelių pritaikymo ir integravimo su esamomis sistemomis sąnaudų [21]. Galimos naudos, kurias dirbtinio intelekto technologijos galėtų suteikti po įdiegimo, gali nepašalinti investicijų kaštų verslams, o tai gali turėti neigiamą poveikį verslo veiklai. Pažymėtina, kad dirbtinio intelekto technologijų diegimą įmonėse dažnai riboja silpna duomenų kultūra, dėl kurios daugelis organizacijų nesuvokia viso DI potencialo. Vis dėlto, įmonės, kurios sėkmingai įdiegė šias technologijas, pasiekė reikšmingų rezultatų. [22]:

- “Danone Group” naudodama ML pagerino savo paklausos prognozių tikslumą, pagerino planavimo koordinavimą tarp rinkodaros, pardavimų, klientų aptarnavimo, tiekimo grandinės ir finansų, dėl ko buvo gautos tikslesnės prognozės. Taip pat, naudojant ML 20% sumažėjo prognozių klaidų, 30% sumažėjo prarastų prekių pardavimai, 30% sumažėjo produkto atgyvenimo laikotarpiai ir net 50% sumažėjo paklausos planuotojų darbo krūvis.

- “Nokia” naudodama mašininį mokymąsi stebi ir analizuoja gamybos procesus, kurie praneša operatoriui apie nesuderinamumus procesuose, tam, kad problemos būtų išspręstos realiu laiku.
- “Caterpillar Marine Division” sutaupo apie 400 tūkst. JAV dolerių per metus analizuodama duomenis, kaip dažnai turi būti valomi laivai.

Kita reikšminga analizė buvo atlikta tyrime, kuriame buvo nagrinėjama, kaip B2B sektoriuje veikiančios mažosios ir vidutinės įmonės (angl. SME) gali praktiškai pritaikyti DI [25]. Tyrime analizuota Kinijos įmonė „CQ“, įkurta 2004 m., padedanti SME eksportuoti produkciją. Ši įmonė integravo DI sprendimus įvairiose veiklos srityse - nuo klientų segmentavimo ir kuponų personalizavimo iki rekomendacinių sistemų, reklaminių skydelių generavimo ir automatizuotos logistikos. Be to, pasitelkus NLP technologijas buvo sukurtas pokalbių robotas su kalbos atpažinimu ir vertimu, padidinantis pasiekiamumą tarptautinėje rinkoje. Taip pat diegtos klastočių atpažinimo priemonės, kurios padeda apsaugoti įmonės reputaciją bei vartotojų pasitikėjimą.

Taip pat buvo atlikti moksliniai tyrimai, norint išsiaiškinti DI poveikį verslui, viename tyrime buvo pastebėtas tiesioginis DI gebėjimų poveikis verslo veiklai, parodant, kad investicijos į DI ir jo panaudojimas gali suteikti konkurencinį pranašumą ir pagerinti verslo veiklą [23]. Šie rezultatai patvirtina ankstesnius tyrimus [24], kurie teigia, kad dirbtinis intelektas gali turėti teigiamą poveikį verslo veiklai, tačiau gauti rezultatai gali priklausyti nuo daugelio veiksnių, tokių kaip verslo tikslai, verslo aplinka ir kitos organizacinės charakteristikos.

Apibendrinant galima teigti, kad DI integravimas į įmonės veiklą turi reikšmingą teigiamą poveikį jos konkurencingumui. Tai leidžia įmonėms didinti produktyvumą, gerinti klientų aptarnavimą ir priimti duomenimis pagrįstus sprendimus, kurie prisideda prie jų sėkmės rinkoje. Tačiau svarbu pažymėti, kad sėkminga DI integracija reikalauja tinkamų žmogiškųjų išteklių ir investicijų į technologijas, todėl įmonės turėtų rūpestingai planuoti šį procesą.

1.2. Sprendimų priėmimas ir patikimo dirbtinio intelekto integracija

1.2.1. Sprendimų priėmimas ir jo rūšys

Norint išsamiau aptarti pasitikėjimo verto dirbtinio intelekto pritaikymą sprendimų priėmime, pirmiausia būtina suprasti, kas yra sprendimų priėmimas ir kokie yra jo pagrindiniai elementai. Sprendimų priėmimas yra pagrindinis organizacijos veiklos elementas, kuris tiesiogiai daro įtaką įmonės rezultatams, veiklos efektyvumui ir konkurencingumui rinkoje. Šis procesas apima galimų veiksmų alternatyvų pasirinkimą, siekiant skirtingų tikslų. Sprendimų priėmimo procesas gali būti paprastas, kai pasirenkama iš kelių aiškių alternatyvų, arba sudėtingas, kai reikia atsižvelgti į daugelį kintamųjų ir/ar nežinomųjų. Sprendimų priėmimo kokybė ir greitis dažnai lemia organizacijos gebėjimą prisitaikyti prie besikeičiančių sąlygų ir išlaikyti konkurencingumą [7], todėl supratimas apie sprendimų priėmimo struktūrą ir jo etapus yra būtinas, siekiant įvertinti, kaip DI gali padėti tobulinti šį procesą.

2 pav. Sprendimų priėmimo etapai

Sprendimų priėmimas paprastai susideda iš septynių etapų, kurie užtikrina, kad sprendimas būtų pagrįstas ir efektyvus:

1. **Tikslo identifikavimas.** Pirmas sprendimų priėmimo proceso žingsnis yra aiškiai apibrėžti problemą ir tikslą, kurį reikia pasiekti. Šiame etape reikia suprasti apie problemos pobūdį ir jo svarbą organizacijos veiklai [26].
2. **Informacijos surinkimas.** Norint pradėti ieškoti sprendimų, pirmiausia būtina surinkti visą reikalingą informaciją, kuri leistų pasirinkti optimalų sprendimo variantą. Tai reikalauja prieigos prie organizacijos viduje sukauptų duomenų Taip pat galimybės gauti duomenis apie išorinius veiksnius, tokius kaip rinkos sąlygos ir konkurentų veikla.
3. **Alternatyvų nustatymas.** Identifikavus problemą ir surinkus informaciją, būtina nustatyti galimus sprendimo būdus ir alternatyvas. Kiekviena alternatyva turėtų būti vertinama atsižvelgiant į tai, kaip ji padės pasiekti tikslą [27].
4. **Alternatyvų įvertinimas**. Alternatyvos vertinamos remiantis tam tikrais kriterijais, tokiais kaip kaštai, rizika, galimi poveikiai ir nauda. Tai padeda išsirinkti tinkamiausią sprendimo variantą [7].
5. **Sprendimo parinkimas.** Pasirinkus alternatyvą, priimamas galutinis sprendimas. Svarbu užtikrinti, kad sprendimas atitiktų nustatytą tikslą ir būtų tinkamas esamai situacijai.
6. **Įgyvendinimas**. Pasirinkto sprendimo įgyvendinimas vykdomas remiantis parengtu planu, kuriame būtina aiškiai paskirstyti užduotis ir atsakomybes bei nustatyti veiksmų seką.
7. **Rezultatų vertinimas**. Po sprendimo įgyvendinimo svarbu įvertinti, ar pasiekti norimi rezultatai, ir, prireikus, atlikti korekcijas. Šis žingsnis padeda tobulinti sprendimų priėmimo procesą ateityje.

Sprendimų priėmimas, atsižvelgus į sprendimo sudėtingumo lygį ir situacijos kontekstą, skirstomas į kelias pagrindines grupes:

- **Struktūruoti sprendimai**. Tai yra tokie sprendimai, kurie gali būti lengvai priimami pagal nustatytas taisykles ir logiką.
- **Nestruktūruoti sprendimai.** Tai sprendimai, kurie nėra aiškiai apibrėžti. Jie retai būna pasikartojantys, todėl reikalauja kūrybingumo, intuicijos ir detalios analizės. Vienas tokių pavyzdžių yra sprendimai, susiję su naujų produktų kūrimu ar įėjimu į naujas rinkas [28].
- **Strateginiai sprendimai.** Ilgalaikiai sprendimai, kurie dažniausia priimami aukščiausio lygio vadovų ir reikalaujantys plataus masto analizės.
- **Taktiniai sprendimai.** Šie sprendimai yra priimami trumpesniu laikotarpiu ir dažnai susiję su konkrečių užduočių arba projektų vykdymu. Jie remiasi strateginiais sprendimais ir padeda juos įgyvendinti. Pavyzdžiui, tai gali būti sprendimai dėl konkrečių rinkodaros kampanijų vykdymo arba operatyvinių gamybos procesų optimizavimo.

1.2.2. Sprendimų priėmimui įtaką darantys veiksniai ir jų svarba organizacijose

Efektyvus sprendimų priėmimas yra esminis organizacijos veiklos aspektas, nes nuo jo tiesiogiai priklauso organizacijos sėkmė. Gerai apgalvoti ir pagrįsti sprendimai leidžia organizacijai veikti efektyviau, pasiekti užsibrėžtus tikslus ir išlikti konkurencinga rinkoje. Priešingai, nepagrįsti ir skubotai priimti sprendimai gali sukelti rimtų pasekmių - nuo finansinių nuostolių iki reputacijos pablogėjimo ir darbuotojų moralės nuosmukio [29]. Norint užtikrinti, kad sprendimai būtų kuo efektyvesni, būtina atsižvelgti į įvairius veiksnius, kurie gali turėti įtakos sprendimų priėmimo procesui. Šie veiksniai gali kilti tiek iš organizacijos vidaus, tiek iš išorės, ir daryti esminę įtaką sprendimų priėmimo eigai bei rezultatams. Juos suvokdami, sprendimų priėmėjai gali geriau pasirengti, įvertinti galimas kliūtis ir pranašumus, galinčius paveikti galutinį sprendimą. Žemiau pateikiami pagrindiniai tokie veiksniai:

- **Informacijos prieinamumas.** Kuo daugiau turima informacijos, tuo tikslesnis gali būti sprendimas. Tačiau per didelis informacijos kiekis taip pat gali sukelti vadinamąjį "informacijos pertekliaus" efektą, kuris gali trukdyti priimti sprendimus. Tyrimai rodo, kad per didelis informacijos kiekis gali sukelti sprendimų priėmimo paradoksą, kai per daug pasirinkimo galimybių lemia mažiau efektyvius sprendimus [30].
- **Rizikos tolerancija.** Sprendimų priėmėjai turi skirtingą rizikos toleranciją. Kai kurie linkę priimti rizikingus sprendimus, tikėdamiesi didesnio atlygio, kiti – renkasi saugesnius variantus. Rizikos tolerancijos lygis dažnai priklauso nuo asmeninių ir organizacinių veiksnių, įskaitant ankstesnę patirtį, organizacijos kultūrą ir vadovų asmenines savybes [31].
- **Laiko apribojimai.** Laikas, per kurį reikia priimti sprendimą, taip pat gali daryti įtaką sprendimų priėmimo kokybei. Skubotai priimti sprendimai gali būti mažiau apgalvoti ir turėti neigiamų pasekmių. Tyrimai rodo, kad laiko spaudimas dažnai sumažina sprendimų priėmimo kokybę, nes sprendimų priėmėjai linkę sutrumpinti problemos analizę ir naudoti paprastesnes sprendimų taisykles [32].
- **Kultūriniai ir organizaciniai veiksniai.** Organizacijos kultūra ir struktūra taip pat turi įtakos sprendimų priėmimui. Pavyzdžiui, demokratinėse organizacijose sprendimai gali būti priimami kolektyviai, o hierarchinėse – sprendimų priėmimas gali būti centralizuotas. Skirtingos organizacijos struktūros gali daryti didelę įtaką sprendimų priėmimo procesams bei lemti sprendimų greitį, kūrybingumą ir įgyvendinimą [33].

1.2.3. DI privalumai sprendimo priėmime

Šiuolaikinės organizacijos, atsižvelgdamos į veiksnius, lemiančius sprendimų priėmimo efektyvumą, vis dažniau remiasi technologijomis, tokiomis kaip dirbtinis intelektas (DI), lyginant su tradiciniais sprendimo būdais, kurie pagrįsti žmogiškuoju intelektu (ŽI, angl. *human intelligence*, HI) ir intuicija. Šis DI integravimas į organizacijos sprendimų priėmimo procesą atveria daugybę galimybių, kurios gali iš esmės transformuoti verslo operacijas.

Viena tokių yra galimybė apdoroti didžiulius duomenų kiekius per labai trumpą laiką, pasinaudojus didžiųjų duomenų technologijom kaip „Apache Spark“ ar „Apache Hadoop“, kurios paskirstydamos duomenų apdorojimo užduotis tarp kelių procesorių ar branduolių, leidžia efektyviau ir greičiau juos apdoroti. Be to, mašininio mokymosi algoritmai, tokie kaip atsitiktiniai miškai ir gilusis mokymasis, yra optimizuoti dirbti su dideliais duomenų kiekiais, efektyviai apdorodami ir analizuodami sudėtingas duomenų struktūras. Ši savybė yra vis svarbesnė dabartinėje verslo aplinkoje, kur

sprendimo priėmimo greitis ir prisitaikymas prie nuolat kintančių rinkos sąlygų yra esminiai konkurencinio pranašumo elementai. Tuo tarpu tradiciniai, ŽI paremti sprendimo priėmimo metodai, nors ir vertingi dėl savo lankstumo ir kūrybiškumo, analizuojant sudėtingus kontekstus, dažnai nėra tokie greiti ir efektyvūs [27, 34]. Taip pat, pastaraisiais metais giluminis mokymasis pasistūmėjo į priekį, leisdamas mašinoms mokytis iš neapdorotų duomenų ir taip į sprendimų priėmimo procesą įtraukti didesnį duomenų kiekį, todėl tokiose situacijose, kai yra daug kintamųjų, žmogus nesugeba apdoroti visų duomenų ar rasti priežastingumo tarp jų, tačiau DI, pasitelkiant ML ir algoritmus, randa geresnį sprendimą.

Kitas svarbus aspektas yra DI gebėjimas padėti organizacijoms geriau valdyti rizikas ir prognozuoti galimus neigiamus įvykius. Naudojant DI, galima iš anksto identifikuoti galimas grėsmes ir nedelsiant imtis prevencinių veiksmų, kurie sumažina tiek finansines, tiek operacines rizikas. To pavyzdys - bankų naudojamas ML, siekiant aptikti sukčiavimus mokėjimų sistemose. Kitas reikšmingas rizikos valdymo aspektas yra natūralios kalbos apdorojimas, leidžiantis analizuoti tekstinius duomenis, tokius kaip naujienų straipsniai ar klientų atsiliepimai. Tinkamai apdorojus šiuos duomenis, galima identifikuoti svarbią informaciją apie galimas rizikas ir kylančias problemas [7].

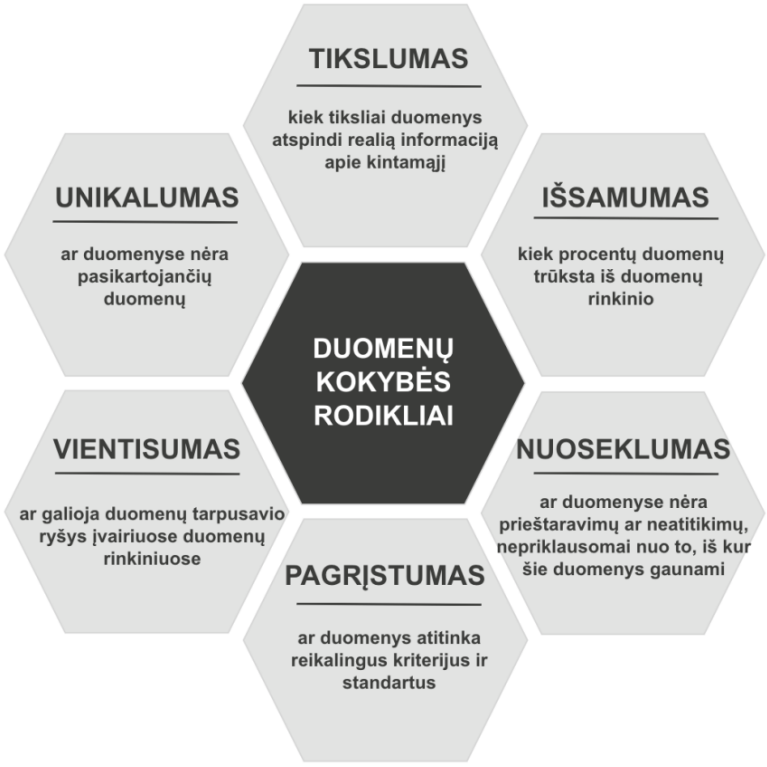
Be to, DI suteikia sprendimų priėmimui objektyvumo ir mažina šališkumo riziką. Tradiciniai sprendimų priėmimo būdai, pagrįsti žmogaus intuicija ir patirtimi, nors ir gali būti labai efektyvūs, dažnai būna veikiami subjektyvių veiksnių, tokių kaip emocijos, šališkumas ar išankstinis nusistatymas. DI, kita vertus, priima sprendimus remdamasis griežtomis duomenų analizėmis ir statistika, o tai sumažina šališkumo riziką ir padidina sprendimų nuoseklumą bei patikimumą [35].

1.3. Pasitikėjimo vertos rekomendacinės sistemos paremtos RL koncepcija

Nors DI integravimas į sprendimų priėmimo procesus atveria daug galimybių, būtina nepamiršti, kad ši technologija taip pat kelia nemažai iššūkių, kuriuos reikia įveikti norint pasiekti maksimalią naudą.

1.3.1. Duomenų kokybė ir prieinamumas

Vienas didžiausių iššūkių, nuo kurio priklauso DI pritaikymo ir integravimo tikslumas, yra duomenų kokybė ir jų prieinamumas. DI algoritmai yra veiksmingi tik tada, kai jie veikia su tiksliais ir reprezentatyviais duomenimis, tačiau daugelis organizacijų susiduria su problemomis, kai duomenų kiekis yra nepakankamas, netikslūs ar šališki. Netinkamai sukonstruoti duomenų rinkiniai gali lemti neteisingus sprendimus, kurie ne tik nesuteikia naudos, bet gali atnešti ir nuostolių. Be to, organizacijos dažnai patiria sunkumų dėl duomenų saugojimo, apsaugos ir prieinamumo, ypač kai kalbama apie didžiulius duomenų kiekius, reikalingus efektyviam DI veikimui [35].



3 pav. Duomenų kokybės rodikliai

Duomenų kokybė yra esminis veiksnys, lemiantis DI sistemų sėkmę, nes nuo jos priklauso ne tik mokymosi procesas, bet ir taikymo etape gaunami rezultatai. Duomenų kokybė apima tokius aspektus kaip tikslumas, aktualumas ir nuoseklumas. Pavyzdžiui, ISO/IEC 25012:2008 standartas nurodo, kad duomenų kokybė priklauso nuo įvairių charakteristikų, kurias reikia pritaikyti, atsižvelgiant į konkretų DI taikymo atvejį. Tikslumas yra vienas svarbiausių kriterijų – duomenys turi tiksliai atspindėti realybę, o klaidingi duomenys turi būti atskirti nuo teisingų.

Nuoseklumas yra dar vienas svarbus duomenų kokybės aspektas, nes duomenys turi būti suderinti tiek viduje, tiek tarp skirtingų duomenų rinkinių, kad būtų išvengta prieštaringų rezultatų. Be to, duomenų prieinamumas ir skaidrumas yra būtini, siekiant užtikrinti, kad duomenys būtų lengvai pasiekiami ir patikimi. Aiškūs duomenų aprašymai ir dokumentacija padeda užtikrinti duomenų kilmės atsekamumą ir padidina vartotojų pasitikėjimą DI priimamais sprendimais [36].

Norint pasiekti aukštą duomenų kokybės lygį, organizacijos turi tiksliai apibrėžti duomenų reikalavimus ir griežtai valdyti visą duomenų apdorojimo procesą – nuo duomenų surinkimo ir valymo iki jų analizės ir interpretavimo. Tik užtikrinus kokybišką duomenų valdymą, galima tikėtis patikimų ir naudingų DI sprendimų [37].

1.3.2. Dirbtinio intelekto paaiškinamumas

Kitas svarbus iššūkis yra DI modelių paaiškinamumas. Daugelis DI modelių, ypač tie, kurie naudoja giluminio mokymosi algoritmus, veikia kaip „juodosios dėžės“, kuriose galutinis sprendimas yra gaunamas per sudėtingus, žmonėms sunkiai suprantamus procesus. Šis neaiškumas gali kelti problemų, kai sprendimų priėmėjai turi pasitikėti DI priimtais sprendimais arba juos pagrįsti prieš

kitus suinteresuotus asmenis. Be aiškumo ir paaiškinimo, kaip DI priima sprendimus, gali kilti pasitikėjimo krizė, trukdanti visapusiškai išnaudoti DI potencialą organizacijoje [38].

Paskutinius dešimtmečius yra daug koncentruojamasi į tai, kaip pagerinti DI modelių prognozavimo tikslumą. Vienas tokių pavyzdžių yra įvairių šaltinių teksto analizė tam, kad būtų galima identifikuoti potencialias verslo klaidas, pasinaudojant DL paremtu NLP modeliu. Siekis pasiekti maksimalų prognozavimo tikslumą turi dvi medalio puses. Norint pagerinti modelio tikslumą, dažniausia kenčia jo paaiškinamumas ir tai sprendimų priėmėjams nekelia pasitikėjimo, o kartais net priveda prie pačio DI atmetimo [39]. Kaip galima pasitikėti tuo, ko negali pilnai suprasti? Ypač tai aktualu valstybiniame ar kitame strateginiame sektoriuje, kur skaidrumas yra itin svarbus. Vienas iš pagrindinių sunkumų interpretuoti sudėtingus DI modelius yra jų struktūros, tokios kaip giluminis mokymasis, kuris susideda iš daugybės neuronų sluoksnių, kurie atlieka skaičiavimas ar operacijas, remdamiesi ankstesnių sluoksnių rezultatais. Kiekvienas sluoksnis geba atlikti sudėtingas netiesines transformacijas, todėl tampa sudėtinga interpretuoti galutinius rezultatus, o tai sukelia problemų analizuojant galimus modelio šališkumus ar padarytas klaidas. Būtent dėl šių priežasčių "juodosios dėžės" principu veikiantys modeliai kelia etikos ir atsakomybės klausimus, ypač tada, kai sprendimas gali turėti itin reikšmingų pasekmių. Vienas iš tokių pavyzdžių - 2018m. skandalas dėl lyčių diskriminacijos naudojant DI darbuotojų atrankoje. Įmonė sukūrė ir naudojo DI pagrįstą darbuotojų atrankos sistemą, kuri turėjo padėti įmonės personalo skyriui greičiau ir efektyviau atrinkti kandidatus į darbo vietas, tačiau sistema pradėjo rodyti lyčių šališkumą ir sistemingai diskriminavo moteris, rodydama prioritetą vyrams. Būtent dėl tokių priežasčių reikia atkreipti dėmesį į galimus pavojus, susijusius su DI šališkumu, interpretavimu ir būtinybe užtikrinti modelių skaidrumą ir supratimą.

Atsižvelgiant į šį iššūkį, atsirado poreikis paaiškinamajam dirbtiniam intelektui - PDI (angl. *explainable artificial intelligence,* XAI). Šis yra DI sritis, kurios tikslas yra sukurti modelius ir sistemas, kurios ne tik priima sprendimus, bet ir gali paaiškinti, kaip ir kodėl buvo priimtas konkretus sprendimas. PDI siekia padidinti pasitikėjimą DI, suteikdama vartotojams galimybę suprasti modelių logiką, priimtus sprendimus ir sprendimų pagrindimą. Vienas paprasčiausių pavyzdžių yra sprendimų medis, kuris yra priskiriamas prie PDI, kuris naudojamas klasifikavimo ir prognozavimo užduotimis. Šis yra lengvai paaiškinimas, nes kiekvienas sprendimas yra aiškiai matomas per nuoseklų taisyklių rinkinį.

Daugybė mokslininkų teigia, kad PDI yra raktas į DI panaudojimą ir integravimą tokiose verslo srityse, kaip e-prekyba, bankininkystė ir finansinės paslaugos [40]. Taip pat reikia pabrėžti, kad žinojimas, kaip veikia modelis, yra toks pat svarbus kaip jo tikslumas, nes tai padeda geriau suprasti pačius modelio hiperparametrus. Tai savo ruožtu suteikia vartotojams galimybę geriau pagrįsti prognozes suinteresuotoms šalims.

1.3.3. Etiniai/teisiniai klausimai ir atsakomybių pasidalijimas

Dirbtinio intelekto sistemų kūrimas, diegimas ir vertinimas kelia sudėtingus etinius ir teisinius klausimus, ypač kai kalbama apie jų taikymą socialiai jautriose srityse, tokiose kaip sveikatos priežiūra, finansai, teisėsauga ir švietimas. Įvairūs dalyviai, dalyvaujantys DI sistemų kūrimo procese, turi skirtingus reikalavimus, susijusius su etiškumo pritaikymu sistemoms, todėl būtina suderinti šiuos reikalavimus kuo paprasčiau ir efektyviau. Tai tampa dar sudėtingiau, kai didelės sistemos yra kuriamos kaip atskiros funkcinės posistemės, kurių sąveika dažnai sukelia papildomą

kompleksiškumą, apsunkinantį galimybę įvertinti, kiek bendra sistema atitinka etinius ir teisinius reikalavimus [41].

Norint šį kompleksiškumą valdyti, būtina sukurti procesą, kuris leistų įvertinti tiek atskiras posistemes, tiek visą sistemą. Tai reiškia, kad ne tik pirminis DI sistemos kūrimas, bet ir nuolatinis jos veikimo stebėjimas ir vertinimas yra esminiai, siekiant užtikrinti, jog sistema atitiktų visus reikiamus etinius ir teisės standartus. Ypač svarbu atsižvelgti į tai, kad DI sistemos, kurios mokosi ir tobulėja jau po jų diegimo, reikalauja papildomos priežiūros ir nenutrūkstamo vertinimo. Besikeičiančios sąlygos gali sukelti netikėtus rezultatus, jei treniravimo metu į šias sąlygas nebuvo atsižvelgta, todėl algoritmo peržiūrėjimas ir koregavimas tampa būtinybe [42].

Kitas svarbus aspektas yra atsakomybių pasidalijimas. Kai DI sistema priima sprendimus, kurie gali turėti rimtų pasekmių, būtina aiškiai nustatyti, kas atsakingas už šių sprendimų pasekmes. Atsakomybės grandinė dažnai apima daugelį dalyvių – nuo DI sistemų kūrėjų, kurie yra atsakingi už algoritmo kūrimą ir pradinį treniravimą, iki organizacijų, kurios diegia ir naudoja šias sistemas. Be to, svarbu, kad atsakomybės klausimai būtų aiškiai reglamentuoti, ypač tais atvejais, kai DI sistema veikia autonomiškai arba kai sprendimai priimami, remiantis sudėtingais ir sunkiai paaiškinamais modeliais [43].

Norint užtikrinti, kad DI sistemos būtų etiškai priimtinos, būtina į procesą įtraukti kuo daugiau įvairių suinteresuotų šalių. Tai ypač svarbu sistemų, susijusių su svarbiais etiniais klausimais, kūrimo procese. Suinteresuotųjų šalių dalyvavimas ne tik padeda geriau suprasti socialinį ir kultūrinį kontekstą, kuriame veikia DI sistemos, bet ir užtikrina, kad šios sistemos būtų priimtinos visuomenei. Ilga atsakomybės grandinė, apimanti visus proceso dalyvius, padeda sumažinti riziką ir užtikrinti, kad sprendimai būtų priimti atsakingai ir etiškai [44].

Apibendrinant galima teigti, jog nors dirbtinio intelekto integravimas į sprendimų priėmimo procesus suteikia reikšmingų galimybių ir papildo žmogiškąjį intelektą, šis procesas susiduria su nemažais iššūkiais, kuriuos būtina įveikti siekiant pagerinti sprendimų priėmimo procesą organizacijos viduje. Organizacijos turi būti pasirengusios spręsti duomenų kokybės ir prieinamumo problemas, užtikrinti DI modelių aiškinamumą, atsižvelgti į etinius ir teisinius aspektus bei spręsti pasipriešinimo pokyčiams ir kvalifikacijos stokos klausimus. Tik tokiu būdu organizacijos galės visapusiškai išnaudoti DI potencialą ir užtikrinti, kad sprendimai, kuriuos priima DI, būtų tikslūs, objektyvūs ir etiški.

1.4. Rekomendacinės sistemos koncepcija adaptuojant DI įmonės viduje

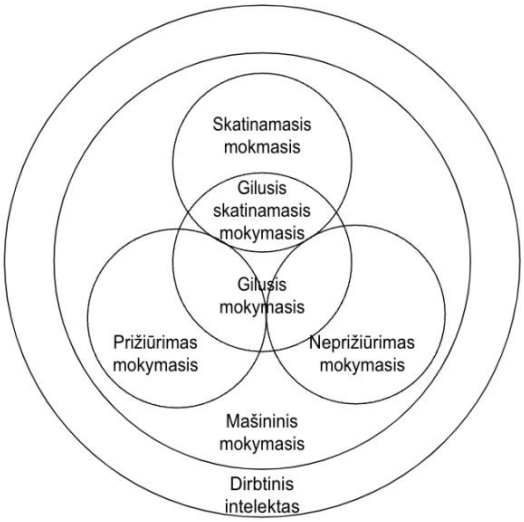
Atsižvelgiant į vis labiau kompleksiškus verslo iššūkius bei augantį reikalavimą kurti etiškas, paaiškinamas ir adaptyvias sprendimų priėmimo sistemas, organizacijoms neužtenka vien tradicinių mašininio mokymosi modelių. Sudėtingose ir dinamiškai besikeičiančiose aplinkose reikia metodų, kurie galėtų ne tik reaguoti į pasikeitimus, bet ir nuolat tobulinti savo veikimo strategijas. Viena iš tokių pažangių technologijų yra skatinamojo mokymosi metodika, kuri tampa vis dažniau taikoma verslo sprendimų priėmimo srityje.

RL metodai, ypač integruoti su giliojo mokymosi algoritmais (angl. *deep reinforcement learning*, DRL), leidžia kurti sprendimų sistemas, kurios geba mokytis iš patirties – t. y. iš sąveikos su aplinka – ir pritaikyti savo strategijas pagal gaunamą grįžtamąjį ryšį. Tokie modeliai nesiremia iš anksto

nustatytomis taisyklėmis ar fiksuotais sprendimų rinkiniais – vietoje to jie nuolat eksperimentuoja, mokosi ir koreguoja savo elgesį, siekdami maksimizuoti ilgalaikį atlygį.

Vienas pagrindinių šių sistemų privalumų – dinamiškas prisitaikymas prie aplinkos: sprendimų priėmimo strategija nėra fiksuota, o tobulinama, remiantis naujai surinktais duomenimis ir besikeičiančiomis sąlygomis [45]. Toks požiūris leidžia modeliams ne tik efektyviau reaguoti į pokyčius, bet ir priimti labiau personalizuotus, organizacijos specifiką atitinkančius sprendimus.

Be to, ilgalaikio atlygio siekimas, būdingas RL metodams, leidžia orientuotis į tvarų verslo rezultatų gerinimą, o ne vien trumpalaikę naudą. Tai ypač svarbu kontekstuose, kur sprendimai turi ilgalaikį poveikį organizacijos veiklos rezultatams ar klientų pasitenkinimui. Naujausi tyrimai rodo, kad giliojo ir skatinamojo mokymosi metodikų derinys leidžia pasiekti itin aukštą sprendimų adaptaciją bei patikimumą net ir sudėtingose situacijose, kuriose sprendimai priimami atsižvelgiant į daugelį kriterijų [46].



4 pav. Dirbtinio intelekto apibrėžimas

Papildomai verta pažymėti, kad skatinamasis mokymasis išaugo į savarankišką dirbtinio intelekto paradigmą, kurią galima taikyti ne tik rekomendacinėse sistemose ar verslo procesuose, bet ir kur kas platesniame taikomųjų sričių spektre. Spartūs DI ir ML pasiekimai sudarė sąlygas RL įsitvirtinti kaip esminei interaktyvaus mokymosi metodikai. Skirtingai nuo prižiūrimo ar neprižiūrimo mokymosi, RL remiasi agento ir aplinkos sąveika, kai agentas per bandymus ir klaidas mokosi strategijų, siekdamas maksimizuoti kaupiamą atlygį [47].

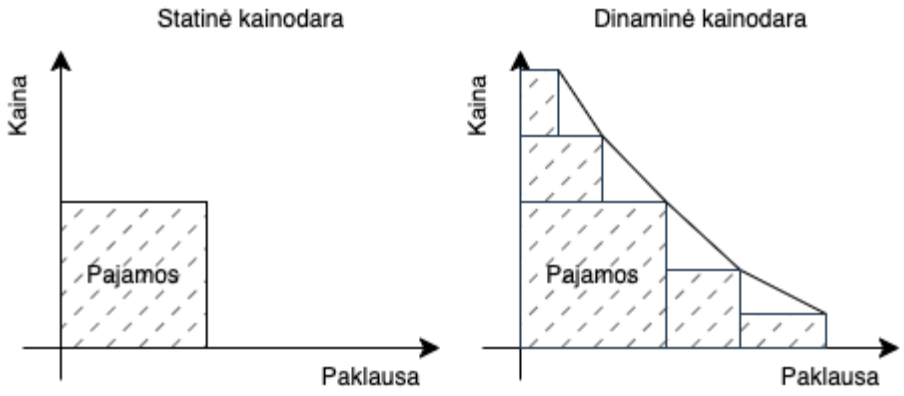
Skatinamojo mokymosi universalumas lėmė jo sėkmingą taikymą įvairiose srityse. Robotikoje ir autonominėse sistemose RL leidžia agentams įgyti sudėtingus valdymo įgūdžius, tokius kaip navigacija ar koordinacija. Komunikacijų ir tinklų valdymo srityje RL algoritmai optimizuoja resursų paskirstymą ir gerina tinklų efektyvumą. Natūralios kalbos apdorojimo ir kompiuterinės regos srityse RL gebėjimas spręsti nuoseklaus sprendimų priėmimo problemas leidžia pasiekti geresnių rezultatų esant fragmentiškam grįžtamajam ryšiui ir didelio matmenų skaičiaus aplinkose. Žaidimų ir kitų sistemų srityje RL metodai jau yra pralenkę žmogaus lygio strateginius gebėjimus [49].

Nepaisant didelės pažangos, RL vis dar susiduria su iššūkiais, tokiais kaip būtinybė valdyti tyrinėjimo ir išnaudojimo (angl. *exploration–exploitation*) balansą, darbas su didelės dimensijos erdvėmis ir užtikrinti pakankamą mokymosi efektyvumą. Pastaraisiais metais gilusis skatinamasis mokymasis (angl. *deep reinforcement learning*, DRL), integruotas su neuroniniais tinklais, išplėtė RL galimybes - tai leidžia agentams veikti sudėtingose aplinkose su dideliais duomenų kiekiais ir sparčiai adaptuotis prie naujų sąlygų [48].

1.5. Dinaminė kainodara: skatinamojo mokymosi taikymas ir lyginamoji analizė

Aptarus bendrąsias skatinamojo mokymosi taikymo galimybes sprendimų priėmimo procesuose, natūralu išskirti specifinę, bet ypač svarbią sritį – dinaminę kainodarą. Dinaminė kainodara tampa vis aktualesnė dėl nuolat kintančios paklausos ir pasiūlos verslo aplinkoje, todėl pažangios, adaptuojančios sprendimų sistemos, pagrįstos RL metodikomis, įgauna reikšmingą vaidmenį.

Dinaminė kainodara yra vienas iš pagrindinių būdų, leidžiančių įmonėms optimizuoti pajamas ir efektyviai valdyti rinkos svyravimus. Tradicinės kainodaros strategijos dažnai nesugeba greitai reaguoti į aplinkos pokyčius, todėl jos praranda aktualumą dinamiškose rinkose. Pastaruoju metu daug dėmesio skiriama dirbtinio intelekto metodų taikymui šioje srityje, ypač skatinamojo mokymosi algoritmams, kurie leidžia modeliams nuolat mokytis ir prisitaikyti prie naujų aplinkybių [46].



5 pav. Statinės ir dinaminės kainodaros palyginimas

Tradiciniai dinaminės kainodaros metodai dažniausiai apima šias tris kategorijas:

1. **Taisyklėmis grįsta kainodara.** Kainos nustatomos pagal iš anksto apibrėžtas taisykles, pavyzdžiui, nuolaidos artėjant galiojimo pabaigai ar kainų didinimas padidėjus paklausai.
2. **Paklausos pagrindu nustatoma kainodara.** Kainos koreguojamos pagal istorinius pardavimo duomenis, sezonines tendencijas ir konkurentų veiksmus.
3. **Atsargų lygio pagrindu nustatoma kainodara.** Kainos kinta priklausomai nuo turimo sandėlio lygio: mažesnis atsargų kiekis dažnai reiškia aukštesnę kainą [58].

Nors šie metodai buvo plačiai naudojami, jie pasižymi esminiais trūkumais – statiniu pobūdžiu, ribotu gebėjimu reaguoti į greitus pokyčius rinkoje ir nesugebėjimu įvertinti kompleksinių vartotojų elgsenos bei išorinių veiksnių. Tokiomis aplinkybėmis ypač svarbus tampa pažangesnių, dirbtiniu intelektu pagrįstų metodų taikymas.

Vienas iš šiuolaikinių požiūrių – tai skatinamuoju mokymusi grįsta kainodara. Ji leidžia sistemoms adaptuotis prie kintančios aplinkos, remtis tiek istorine, tiek realaus laiko informacija, optimizuojant kainodaros strategijas per nuolatinį mokymosi ciklą [60]. RL metodai ne tik eliminuoja statinių taisyklių ribotumus, bet ir leidžia personalizuoti kainas pagal vartotojo elgseną bei padeda pasiekti ilgalaikį pajamų didinimo efektą, kartu gerinant klientų patirtį.

Tyrimai rodo, kad RL algoritmai gali žymiai pagerinti kainodaros sprendimų efektyvumą įvairiose srityse. Pavyzdžiui, Yavuz‘as ir Kaya parodė, kad taikant giluminį Q mokymąsi (angl. *deep q-learning*, DQL) ir minkštąjį aktoriaus-kritiko (*angl. soft actor-critic*, SAC) algoritmus, galima reikšmingai pagerinti prekių su ribotu galiojimu kainų nustatymą, sumažinant sprendimų laiką ir padidinant tikslumą [51]. Panašūs metodai buvo taikomi ir kitose srityse: elektros tinklų apkrovos valdymo optimizavime [52], duomenų produktų rinkose [53], elektromobilių įkrovimo kainodaroje [54], aviacijos sektoriuje [55] ir keleivių pavežėjimo paslaugose [56].

Šie tyrimai akcentuoja, kad RL metodai leidžia ne tik efektyviau optimizuoti kainas pagal esamas sąlygas, bet ir prisitaikyti prie nuolat kintančių rinkos parametrų, siekiant maksimizuoti ilgalaikį organizacijos pelną bei pagerinti klientų patirtį.

1 lentelė. Mokslinių tyrimų analizė skirta RL pritaikymui dinaminėje kainodaroje

Tyrimo autoriai	**Sritis**	**Naudotas metodas**	**Pagrindiniai rezultatai**
Yavuz ir Kaya (2024) [51]	Prekių su ribotu galiojimu kainodara	DQL, SAC	Padidintas kainodaros tikslumas iki 95-96%, sprendimų laikas sutrumpintas 70-80%.
Watari ir kt. (2024) [52]	Elektros tinklų dinamika	DRL	Pagerintas tinklo apkrovos stabilumas iki 57%.
Shen ir kt. (2024) [53]	Duomenų produktų kainodara	DQN	Pagerintas pelningumas, konkreti metinė grąža nenurodyta.
Bae ir kt. (2024) [54]	Elektromobilių įkrovimo kainodara	QL	Individualizuota kainodara, padidintos pajamos (be tikslaus %)
Zhu ir kt. (2024) [55]	Aviacijos sektoriaus skrydžių kainodara	PPO, TRPO	Pasiektas 99% artumas prie teorinio maksimalaus pajamų lygio.
Guo ir kt. (2024) [56]	Keleivių pavėžėjimo paslaugos	RL su dinaminės kainos prognoze	Vairuotojų pajamos padidintos iki 45,5% kai kuriose situacijose.

1.6. Patikimumo iššūkiai skatinamojo mokymosi modeliuose

Skatinamojo mokymosi algoritmai, tokie kaip DQN (angl. *deep Q-network*), PPO (angl. *proximal policy optimization*) ir TRPO (angl. *trust region policy optimization*), yra plačiai taikomi, sprendžiant dinaminės kainodaros ir kitas optimizavimo užduotis. Tačiau jų patikimumas praktiniuose verslo scenarijuose vis dar kelia keletą iššūkių, ypač susijusių su konvergencija, stabilumu, mokymosi greičiu ir sprendimų nuoseklumu.

1. **Konvergencijos garantijos.** Vertės pagrindu veikiantiems algoritmams, tokiems kaip Q-mokymasis ar DQN, konvergencija negali būti garantuojama esant nelinijiniams funkcijų aproksimavimo metodams, pvz., neuroniniams tinklams. Priešingai, politikos gradiento metodai (pvz., PPO, TRPO) užtikrina stabilesnę konvergenciją, o modeliu pagrįsti metodai pasižymi aiškesniu konvergencijos elgesiu [63].
2. **Nestabilumas dėl Q-vertės atnaujinimų**. DQN algoritmui būdinga problema – koreliuotų duomenų sekų naudojimas, kas gali lemti nestabilius sprendimus. Šiam iššūkiui spręsti DQN naudoja patirties pakartojimą (angl. *experience replay*) ir atskirus Q ir tikslinius Q tinklus, kurie stabilizuoja mokymosi eigą [63].
3. **Stochastinės aplinkos iššūkiai.** Esant atsitiktinei paklausai ar kitiems stochastiniams reiškiniams, Q-vertės skaičiavimas tampa mažiau tikslus. Mokslininkai siūlo patobulintas DQL versijas (pDQL1 ir pDQL2), kurios pasižymi didesniu atsparumu neapibrėžtumui ir pagerina sprendimų kokybę stochastinėse sąlygose [51].
4. **Mastelio keitimo problema.** Nors giluminis RL iš esmės sprendžia „dydžio prakeiksmą“ (angl. *curse of dimensionality*), praktikoje net ir tokie algoritmai kaip DQN tampa sunkiai įgyvendinami labai didelės dimensijos veiksmų ar būsenų erdvėse. Naujausi tyrimai rodo, kad analizuojant aviacijos kainodaros atvejį, DQN algoritmas susidūrė su atminties išnaudojimo problema, kai tuo tarpu PPO ir TRPO veikė efektyviau [55].
5. **Mokymosi proceso greitis.** Algoritmų konvergencijos sparta tiesiogiai veikia praktinį jų taikymą. Pvz., „nulio kainos“ įtraukimas į veiksmų erdvę leido spartinti konvergenciją aviacijos sektoriuje [55], o kitame tyrime DQL versijos pademonstravo spartesnį sprendimo laiką palyginti su dinaminio programavimo metodais [51].
6. **Sprendimų nuoseklumas.** Stabilumas per kelis mokymosi ciklus yra esminis patikimumo aspektas. Remiantis Zhu‘o ir bendraautorių tyrimu [55] buvo nustatyta, kad TRPO algoritmas pasižymėjo nuoseklesnėmis pajamomis skirtingose simuliacijose, o PPO išsiskyrė geru bendrinimo gebėjimu.

Apibendrinant, nors DQN, PPO ir kiti RL algoritmai atveria naujas galimybes sudėtingoms optimizavimo užduotims spręsti, jų patikimumo užtikrinimas praktiniuose verslo scenarijuose vis dar reikalauja papildomų metodologinių sprendimų – nuo architektūrinių patobulinimų iki tinkamo hiperparametrų parinkimo ir atsparumo stochastinėms sąlygoms didinimo.

1.6.1. Metodai didinantys RL paaiškinamumą, sąžiningumą ir stabilumą

Atsižvelgiant į anksčiau aptartus patikimumo iššūkius, šiame poskyryje analizuojami literatūroje aptarti metodai bei algoritminės strategijos, kurios padeda užtikrinti arba sustiprinti skatinamojo mokymosi modelių stabilumą, paaiškinamumą ir sąžiningumą – tris esminius komponentus, svarbius praktiniam taikymui verslo sprendimų priėmimo kontekstuose, įskaitant dinaminę kainodarą.

RL modelių stabilumas yra vienas svarbiausių patikimumo aspektų. Literatūroje akcentuojama, kad tam tikri architektūriniai sprendimai, ypač DQN algoritme, padeda sumažinti modelio jautrumą duomenų koreliacijai ir konvergencijos svyravimams. Patirties pakartojimas (angl. *experience replay*) leidžia treniruoti agentą su atsitiktinai atrinktais istoriniais duomenimis, taip mažinant duomenų priklausomybę ir užtikrinant stabilesnį mokymosi procesą [63]. Papildomai naudojamas tikslinis tinklas (angl. *target network*), kuris atskirai saugo vertės funkcijos parametrus ir taip apsaugo nuo pernelyg greitų Q-vertės pokyčių.

Stabilumą stiprina ir algoritmų architektūriniai variantai. Pavyzdžiui, DQL metodo patobulintos versijos (pDQL1 ir pDQL2), taikytos greitai gendančių prekių kainodaros kontekste, pasirodė veiksmingesnės stochastinėje aplinkoje, nes apima daugkartinių paklausų (angl. *multi-demand*)

sudarymą ir vidutinių atlygių naudojimą – tai padėjo sumažinti stochastinio atlygio dispersiją ir užtikrinti Q-vertės stabilumą [51].

Literatūroje taip pat išskiriami strategijos optimizavimo (angl. *policy optimization*) metodai, tokie kaip PPO ir TRPO, kaip tinkamesni praktiniam taikymui nei vertės pagrindu veikiantys metodai (pvz., DQN). PPO pasižymi geresniu bendrinimu ir konvergencijos tolygumu, o TRPO algoritmas, taikytas aviacijos kainodaros atveju, parodė stabilesnį pelno pasiskirstymą skirtinguose mokymosi cikluose [55].

RL algoritmai, ypač giluminiai modeliai, dažnai kritikuojami dėl savo „juodosios dėžės“ pobūdžio. Nors daugelyje šaltinių šiai temai skiriama nedaug dėmesio, kai kuriuose tyrimuose užsimenama apie interpretuojamo įgūdžių įgijimo (angl. interpretable skill acquisition) galimybę, taikant hierarchinius RL modelius kelių užduočių mokymuisi. Toks požiūris padeda išskaidyti sprendimų priėmimą į mažesnius, aiškesnius sprendimus (pvz., tarpinius tikslus), kurie gali būti lengviau analizuojami, tačiau ši kryptis vis dar yra daugiau koncepcinė nei standartizuota praktika.

Atsižvelgiant į augantį poreikį AI sprendimų skaidrumui, galima tikėtis, kad RL interpretacijos metodų raida – tokia kaip strategijos santrauka, būsenos-veiksmo paaiškinimas (angl. state-action explanation) ar svarbos žemėlapiai – taps svarbia tyrimų kryptimi. Vis dėlto šiuo metu RL paaiškinamumo praktiniai metodai literatūroje vis dar traktuojami fragmentiškai.

RL sąžiningumo klausimai literatūroje nagrinėjami retai, tačiau kai kurie tyrimai iškelia rizikas, susijusias su personalizuotos kainodaros taikymu. Mokslininkai tyrime apie elektromobilių (EV) įkrovimo kainodarą iškelia riziką, kad naudojant privačius naudotojų duomenis (pvz., eSOC (angl. *estimated state of charge*) – įkrovimo lygį) personalizuotam kainų nustatymui, gali atsirasti sisteminis šališkumas. Pvz., vartotojai, turintys mažai energijos ir esantys toliau nuo pigesnių įkrovimo stotelių, gali būti priversti priimti ekonomiškai nepalankesnes sąlygas [54].

Šie aspektai rodo, kad, nors algoritminio sąžiningumo priemonės (pvz., sąžiningas atlygio formavimas) RL kontekste dar nėra plačiai taikomos, praktinėse srityse – ypač tose, kur sprendimai paveikia skirtingas socialines grupes – būtina etinė priežiūra ir aiški reguliavimo sistema.

2. Tyrimo metodai

Šiame skyriuje pateikiama matematinė ir algoritminė analizė, kuria remiantis buvo vystoma sprendimų priėmimo sistema, paremta skatinamojo mokymosi metodais. Pradedama nuo RL teorinio pagrindo – Markovo sprendimų proceso (angl. *Markov decision process*, MDP), apimančio tokias pagrindines sąvokas kaip būsena, veiksmas, atlygis, politika ir perėjimo tikimybė. Vėliau pristatomos pagrindinės RL algoritmų grupės: reikšmės funkcijos, politikos optimizavimo ir mišrieji metodai. Taip pat nagrinėjami praktiniai klausimai, susiję su tyrinėjimo ir išnaudojimo balansu, algoritmų stabilumu bei veikimo sudėtingose aplinkose efektyvumu.

Ši struktūrinė metodologija sudaro pagrindą tiriamų modelių kūrimui, treniravimui ir vertinimui realiomis verslo problemų sąlygomis, tokiomis kaip dinaminė kainodara ar rekomendacijų sistemų personalizavimas.

2.1. Markovo sprendimų procesas

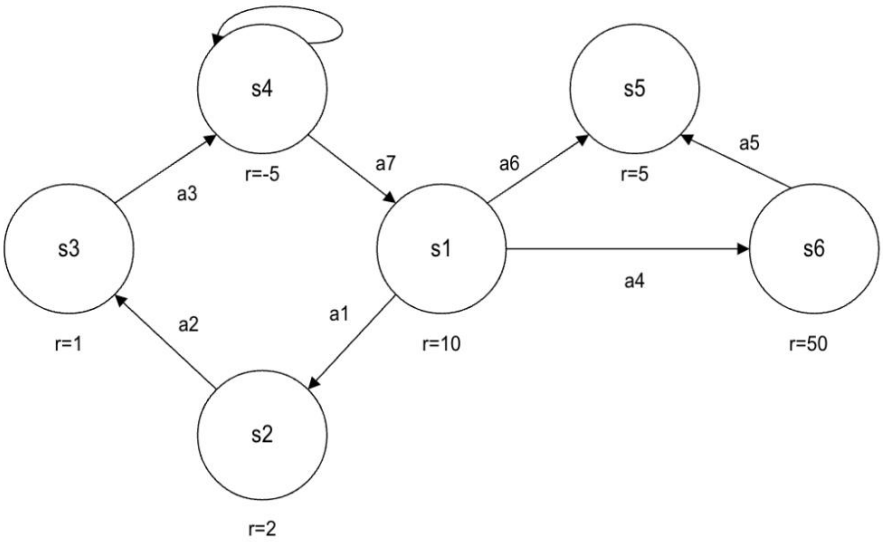
2.1.1. Markovo sprendimų proceso samprata ir taikymas

Markovo sprendimų procesas – tai matematinis modelis, plačiai naudojamas sprendimų priėmimo analizėje stochastinėse sistemose, kai sprendimų rezultatai priklauso nuo agento veiksmų ir atsitiktinių veiksnių. MDP suteikia formalų pagrindą skatinamojo mokymosi algoritmų kūrimui, leidžiant struktūruotai modeliuoti agento sąveiką su aplinka. Šio proceso tikslas – išmokti optimalią veiksmų seką, kitaip tariant - strategiją, kuri maksimaliai padidintų agento ilgalaikį atlygį [57].

MDP modeliai taikomi įvairiuose kontekstuose – nuo robotų valdymo ir autonominių sistemų iki finansinių sprendimų optimizavimo ir logistikos [58]. Skatinamojo mokymosi sistemose MDP yra vertingas tuo, kad leidžia atskirti pagrindinius mokymosi elementus: aplinkos būsenas, galimus veiksmus, perėjimų tikimybes, atlygio struktūrą ir agento politiką. Modelio pagrindas – Markovo savybė, reiškianti, kad būsima būsena priklauso tik nuo dabartinės būsenos ir pasirinkto veiksmo, nepriklausomai nuo ankstesnės istorijos [59].

Vienas iš svarbių MDP aspektų – gebėjimas susieti kiekvieną veiksmo pasekmę su kiekybiškai įvertinamu rezultatu (atlygiu), kas leidžia formaliai apibrėžti agento tikslą – maksimizuoti bendrą atlygį per tam tikrą laikotarpį. Šis procesas yra grindžiamas naudingumo (arba vertės) funkcijomis, kurios dažnai apibrėžiamos rekursyviai, remiantis Bellmano lygtimi – tai viena iš svarbiausių formulių daugelyje RL algoritmų [60]. Taip pat itin svarbus yra diskonto koeficientas γ, nusakantis būsimos naudos svarbą: kai γ artimas 1, sistema orientuojasi į ilgalaikį rezultatą, okai jis mažesnis – į trumpalaikę naudą [61].

MDP galima vizualizuoti grafu, kur mazgai atitinka būsenas, o lankai – perėjimus tarp jų, pažymėtus tikimybėmis ir atlygio reikšmėmis. Šiame kontekste 6 pav. pateikiamas Markovo sprendimų proceso pavyzdys, kuris padeda geriau suprasti, kaip įvairūs veiksmai su skirtingomis pasekmėmis lemia agento elgseną per laiko tarpą.



6 pav. Markovo sprendimo proceso pavyzdys

Kai agentas neturi galimybės tiesiogiai stebėti visos aplinkos būsenos, MDP modelis praplečiamas į dalinai stebimą Markovo sprendimų procesą (POMDP). Tokiu atveju sprendimai grindžiami tikimybiniu pasiskirstymu apie galimas būsenas (angl. *belief state*), o tai leidžia taikyti modelį realiose situacijose, kur egzistuoja duomenų trūkumas ar informacijos neapibrėžtumas [62].

2.1.2. Markovo sprendimų proceso komponentės

Siekiant geriau suprasti, kaip veikia MDP, toliau aptariami pagrindiniai jo komponentai. Kiekvienas komponentas atlieka specifinį vaidmenį apibrėžiant agento ir aplinkos sąveiką, o bendra šių dalių struktūra leidžia formaliai modeliuoti sprendimų priėmimo problemas. Komponentės pateikiamos nuosekliai, pradedant nuo aplinkos apibrėžimo, būsenų ir veiksmų, iki pereinamumo, atlygio bei politikos struktūrų ir su jomis susijusių formulių [63].

MDP modelis yra apibrėžiamas penkiais komponentais:

- **Būsenų aibė (S).** Būsena MDP kontekste – tai galima aplinkos konfigūracija, kurią agentas stebi ar patiria. Ji pateikia esminę informaciją, reikalingą sprendimų priėmimui. Būsenų gali būti baigtinis arba begalinis skaičius, ir jos gali būti apibrėžtos diskrečiomis (pvz., tinklelio koordinatės) arba tolydžiomis reikšmėmis (pvz., temperatūros rodmenys). Perėjimas įvyksta, kai agentas, atlikęs veiksmą, pereina iš vienos būsenos į kitą. Kai kurios būsenos, vadinamos terminalinėmis, žymi epizodo pabaigą. Tinkamas būsenų atvaizdavimas yra kritiškai svarbus RL algoritmų veiksmingumui.
- **Veiksmų aibė (A).** Veiksmas yra bet koks agento sprendimas ar žingsnis, kurį jis gali atlikti aplinkoje. Agentui prieinamų veiksmų rinkinys apibrėžia galimus sąveikos su aplinka būdus. Pavyzdžiui, tinklelio pagrindu paremtame navigacijos uždavinyje veiksmai gali būti judėjimas aukštyn, žemyn, kairėn ar dešinėn. Agentas pasirenka veiksmus, vadovaudamasis politika, siekdamas maksimalios ilgalaikės naudos.
- **Perėjimo tikimybės funkcija (P).** Perėjimo modelis, žymimas kaip $P(s'|s,a)$ apibrėžia tikimybinę aplinkos dinamiką, nurodydamas tikimybę, kad atlikus veiksmą a būsenoje s,

agentas atsidurs būsenoje s′. Šis modelis atspindi neapibrėžtumą, būdingą aplinkai, ir atitinka Markovo savybę, t. y. kita būsena priklauso tik nuo dabartinės būsenos ir veiksmo.

- **Atlygio funkcija (R).** Atlygiai yra grįžtamasis ryšys agentui, kiekybiškai įvertinantis momentinę naudą už buvimą tam tikroje būsenoje ar konkretaus veiksmo atlikimą. Atlygio funkcija R(s) arba R(s,a) priskiria skaitines vertes būsenoms arba veiksmams, skatindama agentą siekti pageidaujamų rezultatų ir vengti nepageidaujamų. Epizodiniuose uždaviniuose atlygių suma skaičiuojama per baigtinį laikotarpį, o nuolatiniuose – naudojama diskontuota grąža, kuri suteikia didesnį svorį artimiausiems atlygiams. Bendras atlygis (arba grąža) laikui bėgant yra išreiškiamas taip:

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \cdots + r_\infty. \quad (1)$$

- **Diskonto koeficientas (γ).** Skaičius tarp 0 ir 1, parodantis, kiek svarbūs yra būsimi atlygiai (kuo γ arčiau 1, tuo labiau vertinami ilgalaikiai rezultatai).

MDP modelio paskirtis – surasti optimalią strategiją π(s), kuri kiekvienoje būsenoje nurodo veiksmą, leidžiantį gauti kuo didesnį bendrą atlygį. Viena pagrindinių MDP savybių yra Markovo savybė, kuri teigia, kad sekančią būseną lemia tik dabartinė būsena ir veiksmas, o ne visa ankstesnė veiksmų seka.

2.2. Skatinamasis mokymasis

Žmonės nuolat siekia tobulinti savo sąveiką su aplinka, mokydamiesi iš praeities patirties. Panašiai, skatinamojo mokymosi (RL) sistemoje dirbtinis agentas siekia to paties tikslo – veikdamas aplinkoje ir mokydamasis iš savo veiksmų pasekmių. RL metodai sukasi apie fundamentines sąvokas: būsena, veiksmas, aplinka, atlygis ir politika. Pagrindinis RL agento tikslas – suformuoti optimalią politiką, kuri leistų priimti sprendimus taip, kad būtų maksimaliai padidintas ilgalaikis sukauptas atlygis.

7 pav. Giluminio skatinamojo mokymosi pavyzdys

RL veikimo schema pavaizduota paveikslėlyje 7. Vienas iš esminių RL išskirtinumų, palyginti su kitomis mašininio mokymosi paradigmomis (pvz., prižiūrimu ar neprižiūrimu mokymusi), yra gebėjimas atsižvelgti į uždelstus atlygius, kai sprendimo pasekmės pasireiškia ne iš karto. Tai itin svarbu sprendžiant sudėtingus, ilgalaikių pasekmių turinčius uždavinius, kaip pažymi Chen‘as ir Wang‘as [57]. Atsižvelgiant į tai, toliau trumpai apžvelgiami esminiai RL sistemos elementai:

aplinka, naudingumo funkcija, politika ir Bellmano lygtis, kuri yra daugelio algoritmų teorinis pagrindas:

- **Aplinka**. Skatinamojo mokymosi kontekste aplinka reiškia išorinę sistemą, su kuria agentas sąveikauja. Ji apibrėžiama taisyklių rinkiniu, nustatančiu agentui galimus veiksmus, galimas būsenas ir su jais susijusius atlygius arba bausmes. Aplinka priima agento dabartinę būseną ir veiksmą kaip įvestį bei grąžina naują būseną ir atlygį kaip išvestį. Ji gali atitikti įvairius realaus pasaulio scenarijus, tokius kaip robotų valdymo sistemos, sveikatos priežiūra ar finansų rinka. Aplinka gali būti sudaryta iš diskrečių arba tolydžių būsenų ir veiksmų, ir jos tinkamas apibrėžimas yra būtinas efektyviam sprendimų priėmimui.
- **Naudingumo funkcija.** Naudingumo funkcija nusako, kiek naudinga yra būsena, remiantis tikėtinu atlygiu, kurį agentas gali sukaupti būdamas joje. Ji skaičiuojama naudojant diskonto koeficientą γ, kuris nusako būsimų atlygių svarbą, palyginti su tiesioginiais atlygiais:

$$U_h = R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots + \gamma^T R(s_n)\ . \quad (2)$$

– **Strategija.** Kitaip –agento politika, kaip pasirinkti veiksmus skirtingose būsenose. Ji išsaugo kiekvieną būseną į atitinkamą veiksmą arba tikimybių paskirstymą veiksmams. Skatinamojo mokymosi tikslas – rasti optimalią politiką, kuri maksimaliai padidina agento ilgalaikį atlygių kaupimą. Ši gali būti deterministinė politika π(s) = a arba stochastinė politika, kur π(a|s) = P(a|s).
– **Belmano lygtis.** Ši suformuluota Richardo Bellmano 1953 m., yra rekursinis metodas būsenoje sukaupto naudingumo apskaičiavimui, atsižvelgiant į momentinius ir būsimo atlygio lūkesčius:

$$U(s) = R(s) + \gamma max \sum T(s, a, s')U(s'), \quad (3)$$

kur U(s) – tai tikėtinas naudingumas būsenoje s; R(s) – momentinis atlygis už buvimą būsenoje s; γ– diskonto koeficientas ($0 < \gamma \leq 1$), nusakantis būsimų atlygių svarbą; T(s, a, s‘) – tikimybė, kad atlikus veiksmą a būsenoje s, bus pereita į būseną s′; o U(s‘) – naudingumas būsenoje s′.

2.2.1. Daugialypio lošimų automato problema

Viena iš paprasčiausių RL užduočių – daugialypio lošimų automato (angl. *multi-armed bandit)* problema. Agentas čia turi rinktis iš keleto galimų veiksmų, o kiekvienas jų grąžina atlygį pagal tam tikrą tikimybės pasiskirstymą. Analogiškai tai galima palyginti su lošimo automatais: kiekvienas automato „rankos“ pasirinkimas suteikia skirtingą išmoką.

Šioje problemoje agento tikslas – išmokti, kurį veiksmą pasirinkti, kad per laiką gautų kuo daugiau bendro atlygio. Ši užduotis sudaro pagrindą sudėtingesnėms sprendimų priėmimo situacijoms, kur reikalingas balansas tarp eksperimentavimo (išbandyti naujus veiksmus) ir išnaudojimo (pasinaudoti tuo, kas jau žinoma, jog yra efektyvu).

2.2.2. Tyrinėjimo ir išnaudojimo dilema

Vienas pagrindinių skatinamojo mokymosi iššūkių – tinkamai suderinti tyrinėjimą ir išnaudojimą. Tyrinėjimas reiškia naujų veiksmų bandymą, galinčių atnešti geresnių rezultatų ateityje. Išnaudojimas – tai esamos žinios taikymas, siekiant maksimalios naudos dabar.

Pavyzdžiui, lošimo automatų atveju tyrinėjimas reikštų skirtingų automatų bandymą, o išnaudojimas – geriausią grąžą duodančio aparato naudojimą. Tačiau pernelyg didelis išnaudojimas gali neatskleisti geresnių alternatyvų, o per daug tyrinėjimo gali sumažinti momentinį efektyvumą.

Tyrinėjimo strategijos, dažniausiai taikomos RL algoritmuose:

- ε -godžioji strategija (angl. *epsilon-greedy)* – veiksmas pasirenkamas atsitiktinai su tam tikra tikimybe (ε).
- Viršutinės pasitikėjimo ribos strategija (angl. *upper confidence bound*, UCB) – pasirenkamas veiksmas su didžiausiu pasitikėjimo intervalu.
- Tompsono atranka – Bayeso požiūriu grįstas atsitiktinis veiksmų pasirinkimas.

2.2.3. Vertės funkcija

Vertės funkcijos padeda RL agentui įvertinti, kiek naudinga būti tam tikroje būsenoje, apskaičiuojant tikėtiną būsimų sukauptų atlygių sumą nuo tos būsenos. Būsenos vertės funkcija *(*angl. *state-value function*) parodo, kokį tikėtiną atlygį agentas gaus, jei laikysis tam tikros politikos. Matematiškai tai išreiškiama taip:

$$V^{\pi}(s) = \mathbb{E}[R_t|s_t = s. \quad (4)$$

Panašiai, veiksmo vertės funkcija (angl. *action-value function)*, dar vadinama Q-funkcija, įvertina tikėtiną grąžą, kai agentas tam tikroje būsenoje pasirenka konkretų veiksmą ir vėliau laikosi politikos. Ji išreiškiama taip:

$$Q^{\pi}(s, a) = \mathbb{E}[R_t|s_t = s_t a_t = a]. \quad (5)$$

Vertės funkcijos yra fundamentaliai svarbios skatinamajame mokyme, nes jos padeda agentui priimti geresnius sprendimus, prognozuojant būsimą naudą, atsižvelgiant į būsenas ir veiksmus.

Bellmano lygtys pateikia rekursinį ryšį, leidžiantį apskaičiuoti būsenų ir veiksmų-būsenų porų vertes skatinamajame mokyme. Šios lygtys išreiškia būsenos vertę per momentinį atlygį ir tikėtiną būsimų būsenų vertę:

$$V^{\pi}(s) = R(s) + \sum \gamma V^{\pi}(s'). \quad (6)$$

$$Q^{\pi}(s, a) = R(s) + \sum \gamma Q^{\pi}(s', a'). \quad (7)$$

Šios lygtys sudaro teorinį pagrindą daugeliui skatinamojo mokymosi algoritmų ir padeda agentams priimti sprendimus remiantis prognozuojama ilgalaike nauda.

2.2.4. Tikslo funkcija skatinamajame mokyme

Skatinamojo mokymosi tikslas yra optimizuoti agento veiksmų strategiją taip, kad būtų maksimaliai padidintas ilgalaikis sukauptas atlygis. Šis tikslas formalizuojamas per tikslo funkciją (angl. *objective function*), kuri apibrėžia, kokią kriterijaus reikšmę agentas turi maksimizuoti. RL kontekste dažniausiai optimizuojama tikėtina diskontuota grąža – bendra gautų atlygių suma, įvertinta nuo tam tikros pradinės būsenos, laikantis politikos π [64]. Tikslo funkcija dažniausia pateikiama kaip:

$$J(\pi) = \mathbb{E}\pi\left[\sum_{t=0}^{\infty} \gamma^t R_t\right], \tag{8}$$

kur $J(\pi)$ – tikėtinas bendras diskontuotas atlygis; $\mathbb{E}\pi$ – vidurkis, kai veiksmai pasirenkami pagal politiką $\pi[\ ]$; R_t – atlygis gautas laiko momentu t; γ – diskonto koeficientas;

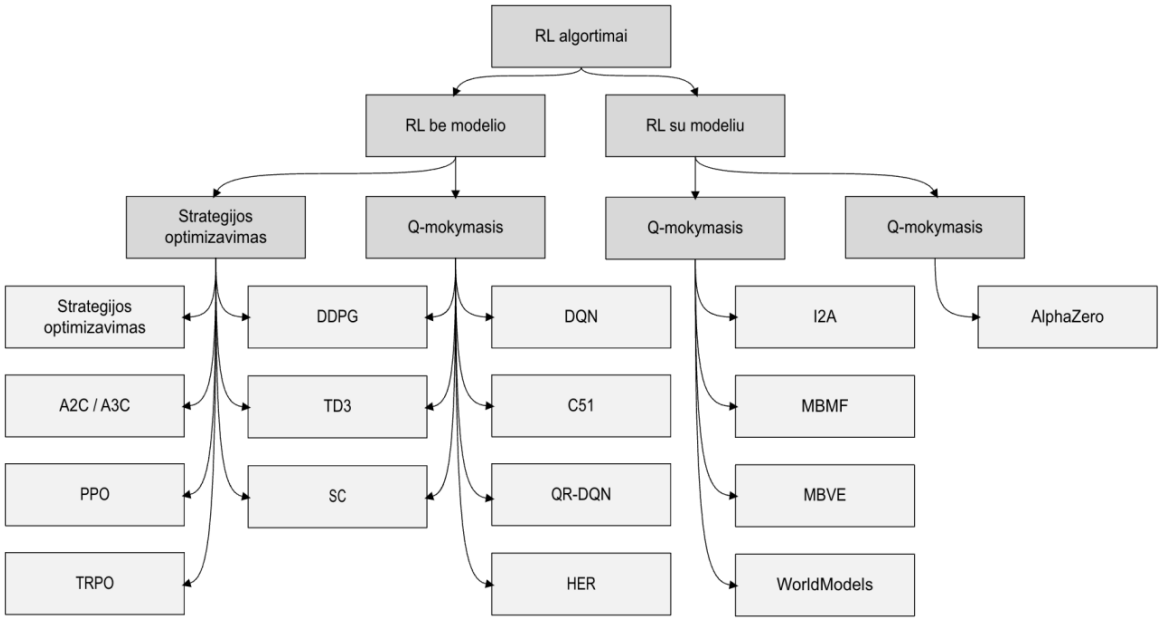
Skirtingi skatinamojo mokymosi algoritmai optimizuoja šią funkciją skirtingais būdais. Vertės pagrindu veikiantys metodai (angl. *value-based*) optimizuoja $J(\pi)$ netiesiogiai, apskaičiuodami būsenos ar veiksmo vertes ir remdamiesi jomis priima sprendimus. Politikos pagrindu veikiantys metodai (angl. *policy-based*) tiesiogiai optimizuoja politiką, naudodami, pavyzdžiui, politikos gradientą [65][66].

Kai kuriais atvejais į tikslo funkciją gali būti įtrauktos papildomos sąlygos ar reguliavimas (pvz., entropijos terminai), siekiant pagerinti tyrinėjimą arba stabilizuoti mokymąsi [67].

Apibendrinant, tikslo funkcija RL sistemoje nurodo, ką konkrečiai agentas mokosi maksimizuoti ir sudaro teorinį pagrindą visai skatinamojo mokymosi metodologijai. Be šios funkcijos mokymosi procesas neturėtų aiškaus optimizavimo kriterijaus.

2.3. Skatinamojo mokymosi algoritmai ir jų klasifikacija

Skatinamojo mokymosi algoritmai paprastai klasifikuojami į strategijomis pagrįstus, vertės funkcijomis pagrįstus ir aktoriaus-kritiko (angl. *actor-critic*) metodus. Be to, jie skirstomi į metodus be modelio, tokius kaip Q-mokymasis, ir metodus su modeliu, tokius kaip dinaminis programavimas. Be modelio metodai mokosi tiesiogiai iš sąveikos su aplinka, o su modeliu metodai naudoja numatomas būsimas būsenas ir atlygius sprendimų priėmimui [63].



8 pav. Skatinamojo mokymosi algoritmų klasifikacija

2.3.1. Laiko skirtumo metodai

Laiko skirtumo (angl. *temporal difference*, TD) metodai yra viena iš pagrindinių stiprinamojo mokymosi atšakų, derinanti Monte Karlo ir dinaminio programavimo principus. Pagrindinis jų bruožas – vertės funkcijų atnaujinimas po kiekvieno veiksmo, nelaukiant viso epizodo pabaigos, kaip reikalaujama Monte Karlo metoduose [58]. TD metodai taiko vadinamąjį „bootstrapping“ principą, kai atnaujinimui naudojamos jau turimos vertės, todėl mokymasis tampa greitesnis ir tinkamas realaus laiko situacijose.

TD pagrindinis atnaujinimo principas išreiškiamas taip:

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)], \quad (9)$$

$$New\ estimate \leftarrow Old\ estimate + \alpha[\text{Target} - Old\ estimate], \quad (10)$$

TD metodų kategorijos skirstomos į:

- **SARSA**. Politika grįstas metodas, kuris atnaujina Q reikšmes, remdamasis veiksmų seka, vykdoma pagal esamą politiką. Nors SARSA dažnai naudojamas rizikos vengiančiuose scenarijuose, šiame darbe nebuvo taikytas dėl techninio palaikymo trūkumo naudojamoje bibliotekoje („Stable Baseline3”).

$$Q(s_t, q_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)], \quad (11)$$

- **Q-mokymasis**. Šis ne strateginis metodas, kuriame Q reikšmės atnaujinamos pagal maksimalią galimą būsimą naudą, nepriklausomai nuo agento vykdomos politikos. Jis yra ypač populiarus dėl savo paprastumo ir efektyvumo sudėtingose aplinkose, todėl ir naudojamas šiame darbe.

$$Q(s_t, q_t) \leftarrow Q(s_t, a_t) + \delta[r_{t+1} + \gamma \max_a Q(s_t, a_t)], \quad (12)$$

Kadangi Q-mokymasis remiasi optimaliu galimu veiksmu ateityje, jis dažnai pasiekia geresnius rezultatus sprendžiant uždavinius, kuriuose svarbus tikslinis maksimalus atlygis, pavyzdžiui, robotikos ar žaidimų srityse.

Algorithm 4: Q-Learning Algorithm

1. Initialize
Q arbitrarily
Q (terminal) =0
Repeat
    initialize s
    Repeat
        choose $a' \in \epsilon - greedily$
        take action a, observe r, s'
        $Q(s_t, a_t) \leftarrow$
$Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}] - Q(s_t, a_t)$
        $s \leftarrow s'$
    s is terminal
until convergence

9 pav. Q-mokymosi algoritmas

2.3.2. Strategijomis pagrįsti metodai

Strategijomis pagrįsti metodai (angl. *policy-based methods*) tiesiogiai optimizuoja agento politiką, apeidami vertės funkcijų skaičiavimą. Šiuose metoduose politika modeliuojama kaip parametrizuota funkcija $\pi(\theta)$, kurios parametrus θ algoritmas tobulina, siekdamas maksimizuoti tikėtiną ilgalaikį atlygį. Tai ypač naudinga aplinkose su tęstiniais veiksmų rinkiniais arba kai reikia generuoti sklandžias veiksmų sekas.

Politika dažnai aproksimuojama naudojant dirbtinius neuroninius tinklus, o jų atnaujinimas atliekamas remiantis gradientiniais metodais. Vienas iš žinomiausių pavyzdžių – strategijos gradiento metodas, kuriame politika atnaujinama pagal tikėtino atlygio gradientą:

$$\nabla J(\theta) = \mathbb{E}[\nabla_\theta log \pi_\theta(a|s) \cdot Q^\pi(s,a)], \quad (13)$$

Šis metodas remiasi esama politika ir ypač tinkamas aplinkoms su sudėtinga, tęstine veiksmų erdve.

Patobulinti metodai, tokie kaip aktoriaus-kritiko, jungia politikos optimizavimą su vertės funkcijos apskaičiavimu, kur „kritikas“ įvertina veiksmo naudą, o „aktorius“ atnaujina politiką. Tuo tarpu tokie algoritmai kaip PPO (angl. Proximal Policy Optimization) ar A2C (angl. .Advantage Actor-Critic) papildomai užtikrina stabilumą ir efektyvumą naudojant reikšmių apribojimo arba sinchronizuotą mokymą.

Šio metodo pagrindą sudarantis strategijos kartojimo (angl. *iteration*) principas gali būti aprašytas ir klasikiniu politikos kartojimo algoritmu (žr. 8 pav.), kuris parodo strategijos įvertinimo ir patobulinimo ciklą iki konvergencijos [63]:

Algorithm 2: Policy Iteration Algorithm

1. Initialization
$V(s) \in Rand\pi(s) \in A(s)arbitrarilyforalls \in S$
2. Policy evaluation
Repeat
    $\Delta \leftarrow 0$
    For each $s \in S$
        $v \leftarrow V(s)$
    $V(s) \leftarrow max_a \sum_{s',r} p(s', r|s, \pi(s)[r + \gamma V(s')]$
        $\Delta \leftarrow max(\Delta, |v - V(s)|)$
3. Policy imporvement
policy -stable $\leftarrow$ true
for each $s \in S$:
    $a \leftarrow \pi(s)$
    $\pi(s) = argmax_a \sum_{s',r} p(s', r|s, a)[r + \gamma V(s')]$
    if $a \neq \pi(s)$ then policy stable $\leftarrow$ false if
policy-stable then stop and return V and π; else go to 2.

10 pav. Strategija pagrįsto algoritmo pavyzdys [63].

2.3.3. Aktoriaus-kritiko metodai

Aktoriaus–kritiko metodai yra hibridinis stiprinamojo mokymosi požiūris, jungiantis tikslo (vertės) ir politikos metodų privalumus [68]. Skirtingai nuo grynai politikos ar vertės pagrindu veikiančių algoritmų, aktoriaus–kritiko architektūra palaiko dvi atskiras, bet kartu veikiančias komponentes:

- **Aktorius**. Ši komponentė yra atsakinga už politikos (veiksmų pasirinkimo taisyklių) atnaujinimą, remiantis gaunamu grįžtamuoju ryšiu iš kritiko.
- **Kritikas**. Įvertina aktoriaus pasirinktus veiksmus, apskaičiuodamas vertės funkciją V(s) arba veiksmų vertės funkciją Q(s, a), kuri naudojama kaip atgalinis signalas politikos tobulinimui.

Tokia architektūra leidžia sistemiškai optimizuoti agento elgseną tiek trumpuoju, tiek ilguoju laikotarpiu ir prisideda prie efektyvesnio konvergavimo sudėtingose ir dinamiškose aplinkose. Tai ypač aktualu, kai veiksmai ir būsenos yra tęstiniai arba turi didelę dimensiją [68, 63].

Algorithm 5: Actor-Critic Algorithm [2]

Initialization
Rewards for state-action pairs $R_{s,a}$
$\gamma = 0.9$
initialize s
1. select a_t in s_t
2. Get s_{t+1}
3. Get R_{s_t,a_t}
4. Update state s_t utility function (critic)
    $U(s_t) \leftarrow U(s_t) + \alpha[r_{t+1}) + \gamma U(s_{t+1)-U(s_t)}]$
5. Update the probability of a_t using error (actor)
    $\delta = r_{t+1} + \gamma U(s_{t+1)-U(s_t)}$

11 pav. Aktoriaus-kritiko algoritmo pavyzdys [63].

Šie metodai pasižymi lankstumu – jie tinka tiek mažoms, tiek didelėms būsenų ir veiksmų erdvėms. Kritikas padeda sumažinti vertinimo dispersiją, o aktorius gali sparčiai reaguoti į besikeičiančią aplinką. Tai leidžia subalansuoti tyrinėjimo ir išnaudojimo procesus, kas yra vienas iš esminių iššūkių stiprinamajame mokyme.

2.3.4. Giliojo Q-mokymosi tinklai (DQN)

Giliojo Q-mokymosi tinklai (angl. *deep Q-networks*, DQN) yra vieni pirmųjų plačiai taikytų giliojo skatinamojo mokymosi algoritmų, kurie leido spręsti sudėtingas užduotis didelės dimensijos aplinkose, kur tradiciniai Q-mokymosi metodai tampa nepraktiški. DQN algoritmo pagrindas – tai Q-funkcijos aproksimavimas naudojant dirbtinius neuroninius tinklus, leidžiantis efektyviai įvertinti būsimos būsenos vertę net ir esant labai plačioms veiksmų ar būsenų erdvėms [57]. Vertės funkcijos atnaujinimui taikoma Q-mokymosi formulė, pritaikyta neuroninio tinklo parametrams:

$$r_j + \gamma \max_{a'} Q\left(\phi_{j+1}, a', \theta^-\right), \tag{14}$$

Ši formulė naudojama treniruojant DQN modelį, kuriame aproksimacija atliekama su neuroniniais tinklais, o parametrai θ atnaujinami, siekiant minimizuoti nuostolio funkciją tarp prognozuotos ir tikslios Q reikšmės. Norint pasiekti stabilų veikimą, DQN modelyje taikomas specialus algoritmas, kuris apima patirties pakartojimą, tikslinius tinklus ir ε-godžioji strategiją:

Algorithm 6: Deep-Q Network Algorithm

Setup RM
Initialize Q(s,a)
Initialize target
repeat
  Initialize sequence $s_1 = x_1$ and prepocessed sequence $\theta_1 = \theta(s_1)$
  repeat
    Either choose randomly a_i with probability ϵ or choose $a_i = argmax_a(Q(\phi(s_t), a; \theta))$
    Get r_{t+1} and x_{t+1}
    $s_{t+1} = (s_t, a_t, x_{t+1})$
    $\phi_{t+1} = \phi(s_{t+1})$
    Store transition $(\phi_t, a_t, r_{t+1}, \phi_{t+1})$ in RM
    Sample random minibatch of transitions $(\phi_t, a_t, r_{t+1}, \phi_{t+1})$ from RM
    if *episode ends at step k+1* **then**
      $y_k = r_k$
    else
      $y_k = r_k + \gamma \max_{a'} Q(\phi_{k+1}, a', \theta^-)$
    end
    Have a gradient descent step on (y_k - $Q(\phi_k, a_k, \theta)$)
    Every C steps set $Q^* = Q$
  until *t = 1, T*;
until *Episode = 1, M*;

12 pav. DQN algoritmo pavyzdys [63].

Vienas iš esminių DQN privalumų yra gebėjimas sumažinti mokymosi proceso nestabilumą. Tam pasitelkiami du pagrindiniai techniniai sprendimai: patirties pakartojimas ir tiksliniai tinklai (angl.

target networks). Patirties pakartojimas leidžia saugoti ir maišyti ankstesnes sąveikas su aplinka, taip sumažinant duomenų koreliaciją ir pagerinant bendrą treniravimo efektyvumą. Tuo tarpu tiksliniai tinklai leidžia stabilizuoti Q reikšmių skaičiavimus, nes tinklo, skirto Q reikšmei apskaičiuoti, parametrai atnaujinami tik periodiškai, o ne po kiekvieno žingsnio.

Nepaisant šių privalumų, DQN modeliai susiduria ir su reikšmingais iššūkiais. Vienas svarbiausių – algoritmų nestabilumas treniruotės metu, ypač kai neuroniniai tinklai veikia kaip nelinijinės funkcijų aproksimacijos priemonės. Tokiu atveju net maži duomenų pokyčiai gali smarkiai paveikti veiksmų politiką, sukeldami stiprius svyravimus ir apsunkindami konvergenciją. Nestabilumo problema taip pat paaštrėja stochastinėse aplinkose, kai paklausos ar aplinkos parametrai kinta atsitiktinai – tokiomis sąlygomis klasikinis DQN tampa jautrus klaidoms vertinant būsimus atlygius [51].

2.4. Kiti naudoti metodai

2.4.1. Pagrindinių komponenčių analizė

Pagrindinių komponenčių analizė (angl. *principal component analysis*, PCA) – tai duomenų dimencijų mažinimo metodas, kuris sumažina aukštos dimensijos duomenis į žemesnės dimensijos erdvę, išlaikant didžiausią įmanomą duomenų dispersiją [69]. Ši analizė buvo pasitelkta, siekiant vizualiai įvertinti modelio suklasifikuotas prekių kategorijas ir nustatyti, kiek skirtingų klasių modelis geba atpažinti.

PCA veikia transformuodama originalius kintamuosius į naujus, tarpusavyje ortogonalius kintamuosius – pagrindines komponentes, kurios yra išrikiuotos pagal jų išsaugomą dispersiją. Tokia transformacija leidžia vizualizuoti sudėtingus duomenų pasiskirstymus 2D arba 3D erdvėje.

Metodika pagrįsta kovariacinės matricos spektrine analize. Pirmiausia atliekamas duomenų standartizavimas, o tuomet apskaičiuojama kovariacinė matrica:

$$\Sigma = \frac{1}{\mathrm{n}-1} X^T X, \tag{15}$$

čia X – standartizuotų duomenų matrica, Σ – kovariacinė matrica.

Po to vykdoma tikrinių vektorių ir reikšmių analizė su sąlyga:

$$\Sigma \mathrm{v} = \lambda \mathrm{v}, \tag{16}$$

čia v – pagrindinių komponenčių vektoriai, o λ – jų tikrinės reikšmės.

Galiausiai duomenys transformuojami į sumažintą erdvę:

$$\mathrm{Z} = XW, \tag{17}$$

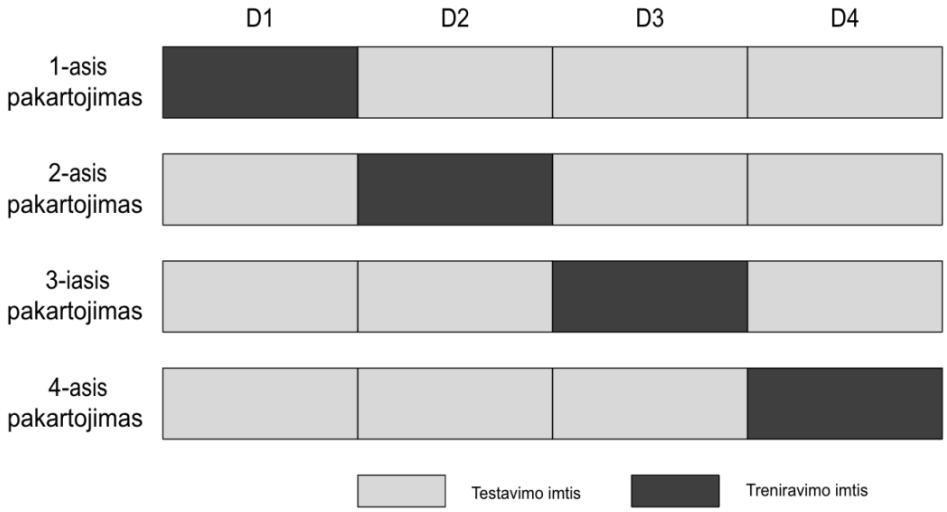
čia W – pasirinktų komponenčių matrica, Z – projekcija žemesnėje dimensijoje.

Ši transformacija buvo panaudota ne tik duomenų vizualizacijai, bet ir rezultato interpretavimui – PCA leido vizualiai įvertinti klasių (prekių kategorijų) pasiskirstymą po klasifikavimo ir nustatyti, ar modelis atskyrė kategorijas aiškiai, ar jos susilieja. Tokiu būdu buvo galima įžvelgti, kiek reikšmingai skiriasi klasės modelio aprėptyje bei preliminariai įvertinti klasifikatoriaus gebėjimą atskirti įvairias prekių grupes.

2.4.2. Kryžminės validacijos metodas

Kryžminė validacija (angl. *cross-validation*) – tai plačiai taikomas metodas, skirtas modelio rezultatų patikrinimui. Vietoj vienkartinio duomenų padalijimo į mokymo ir testavimo rinkinius, šis metodas duomenų aibę padalija į kelias dalis (angl. *folds*), kurios paeiliui naudojamos tiek mokymui, tiek vertinimui [70].

Dažniausiai naudojamas yra k kartų kryžminės validacijos metodas (angl. k-fold cross-validation), kai duomenų aibė padalijama į k lygias dalis. Kiekviename žingsnyje viena dalis naudojama kaip testavimo rinkinys, o likusios – modelio mokymui. Tokiu būdu modelis treniruojamas ir testuojamas k kartų, o galutinis įvertinimas gaunamas apskaičiuojant visų pakartojimų rezultatų vidurkį.



13 pav. 4 dalių kryžminė validacija

Tokiu būdu sumažinamas atsitiktinumas, kylantis dėl vienkartinio padalijimo, bei sumažinama modelio prisitaikymo rizika prie treniravimo duomenų (angl. *overfitting*). Šis metodas ypač naudingas tuomet, kai duomenų kiekis ribotas, tačiau svarbus patikimas modelio įvertinimas.

2.5. Patikimumą didinantys sprendimų priėmimo modeliai grįsti dirbtiniu intelektu

Dirbtinio intelekto (DI) sprendimų priėmimo modeliai skirstomi pagal tai, kiek sprendimo priėmimo proceso yra deleguojama automatizuotai sistemai ir kiek į jį įsitraukia žmogus. Pagrindiniai modelių tipai yra: pilnas sprendimų delegavimas DI, hibridiniai modeliai (DI–Žmogus ir Žmogus–DI), bei daugumos balsavimo struktūros [71]. Kiekvienas iš jų pasižymi skirtingu paaiškinamumo, sprendimo greičio ir patikimumo lygiu.

MODELIS	PAIEŠKOS SPECIFIŠKUMAS	PAAIŠKINAMUMAS	PASIRINKIMŲ KIEKIS	SPRENDIMO PRIĖMIMO GREITIS	ATKARTOJAMUMAS	PAVYZDŽIAI
Pilna delegacija DI	**Aukštas**	**Žemas** (dėl žmogaus nebuvimo procese)	**Aukštas**	**Greitas**	**Aukštas**	Rekomendacinės sitemos, dinaminė kainodara
Hibridinis #1: DI–Žmogus	**Aukštas -> Žemas**	**Aukštas** (dėl žmogaus įsitraukimo priimant galutinį sprendimą)	**Aukštas**	**Lėtas** (dėl žmogaus įsitraukimo)	**Mažas** (dėl žmogaus kintamumo)	Darbuotojų paieška ir įdarbinimas
Hibridinis #1: Žmogus-DI	**Žemas -> Aukštas**	**Žemas** (dėl DI įsitraukimo priimant galutinį sprendimą)	**Žemas**	**Lėtas** (dėl žmogaus įsitraukimo)	**Mažas** (dėl žmogaus kintamumo)	Sporto analitika, sveikatos stebėjimas
Daugumos balso	**Žemas** (Žmogui) **Aukštas** (DI)	**Žemas** (Žmogui) **Aukštas** (DI)	**Žemas** (tas pats pasirinkimų rinkinys abiems)	**Lėtas** (dėl žmogaus įsitraukimo)	**Dalinis** (jei sprendimo priėmimo analizė patikėta DI)	Valdybos posėdžiai

14 pav. Organizacijos sprendimo priėmimo modeliai paremti dirbtiniu intelektu.

Dinaminėje kainodaroje, kur sprendimai turi būti priimami greitai ir apdorojant realaus laiko duomenis, praktikoje dažnai taikomas pilnas delegavimas DI sistemoms, ypač, kai naudojami giliojo mokymosi algoritmai, tokie kaip DQN ar PPO. Šie metodai leidžia adaptuotis prie besikeičiančių rinkos sąlygų ir maksimizuoti pajamas be būtinybės nuolat įsitraukti žmogui. Tačiau pasitikėjimo problema lieka aktuali: ar galima visiškai pasitikėti sprendimais, kurių veikimo logika dažnai yra neaiški?

Šiuo atveju svarbų vaidmenį atlieka hibridiniai modeliai, ypač DI–Žmogus struktūra, kur DI algoritmas pateikia optimalias kainos rekomendacijas, tačiau galutinį sprendimą priima žmogus. Jis gali stebėti sistemą, įsitraukti kritiniais atvejais ir vertinti sprendimų pagrįstumą. Tokia sistema ne tik padidina sprendimų aiškumą ir etinių normų laikymąsi, bet ir išlaiko automatizacijos efektyvumo pranašumus.

Apibendrinant, nors dinaminė kainodara techniškai leidžia visiškai automatizuoti kainų nustatymą, pasitikėjimo stiprinimui organizacijose neretai taikomi hibridiniai modeliai, kurie leidžia suderinti skaidrumą, kontrolę ir efektyvumą.

2.6. Skatinamojo mokymosi bibliotekos pasirinkimas

Didėjant skatinamojo mokymosi (RL) taikymui verslo srityse, tokiose kaip dinaminė kainodara ar personalizuotos rekomendacinės sistemos, vis aktualesnis tampa tinkamos programinės įrangos ir jos bibliotekos pasirinkimas. Tinkama RL biblioteka turi pasižymėti algoritmų modernumu, patikimu palaikymu, vizualizacijos įrankiais bei gebėjimu dirbti su įvairių tipų būsenomis ir veiksmais.

Šiame darbe, kurio tikslas – sukurti pasitikėjimo vertą ir RL pagrįstą sprendimą rekomendacijoms ir kainų optimizavimui, pagrindinis dėmesys skiriamas tik toms bibliotekoms, kurios iš tiesų taikomos arba rekomenduojamos aukščiau minėtose srityse:

- *Stable Baselines3* (SB3). Vienas iš patikimiausių ir plačiausiai naudojamų RL įrankių rinkinys, ypač vertinamas verslo problemų sprendime. SB3 palaiko modernius algoritmus (PPO, A2C, DQN ir kt.), turi *TensorBoard* integraciją, veikia su *OpenAI Gym* aplinkomis ir leidžia greitai realizuoti sprendimus kainodaros ar rekomendacijų srityse. Ši biblioteka taip pat išsiskiria aiškia dokumentacija, strategijų palaikymu (CNN/LSTM) ir aktyvia bendruomene.
- *Ray RLlib*. Skirta didelio mastelio uždaviniams, ypač kai reikia paralelinių ar paskirstytų mokymų. Naudojama kai kuriuose kainodaros ir reklamos technologijų sprendimuose (pvz., „Uber“, „Shopify“). Palaiko plataus spektro RL algoritmus ir integracijas su *Spark*, *TensorFlow*, *PyTorch*.
- *DI-engine*. Šiuolaikinė biblioteka, aktyviai vystoma Kinijos tyrimų bendruomenėje, orientuota į verslo sprendimus. Naudojama kainodaros, rekomendacijų, tiekimo grandinės optimizavimo uždaviniuose.
- *Tianshou*. Lengvai integruojama su *PyTorch* biblioteka ir tinkama greitiems eksperimentams. Pasižymi paprasta API ir palaiko pagrindinius algoritmus. Tinka rekomendacijų uždaviniams su mažesnėmis simuliacijomis ar prototipais.
- *OpenRL*. Nors dar jauna biblioteka, bet pasižymi universaliu pritaikymu – apima NLP, rekomendacines sistemas, savarankiško mokymosi strategijas ir neaktyvus RL. Šiuo metu vertinama dėl integracijos su transformuotais modeliais ir palaikymo GPT pagrindu veikiančioms sistemoms.

Biblioteka	NLP	Daugiaagentinė sistema	Vizualizacijos (TensoarBoard)	CNN/ LTSM	Integravimas	Pritaikymas kainodarai
SB3	✗	✗	✓	✓	✓	✓
Ray RLib	✓	✓	✓	✓	⚠	✓
DI-engine	✓	✓	⚠	✓	⚠	✓
Tianshou	✗	✗	✗	✓	✓	⚠
OpenRL	✓	✓	⚠	✓	✓	⚠

15 pav. Skatinamojo mokymosi bibliotekų palyginimas

Atsižvelgiant į bibliotekų palyginimą (žr. 15 pav.), šiame darbe pasirinkta *Stable Baselines3* biblioteka kaip pagrindinė eksperimentų bazė. Ji užtikrina pakankamą algoritmų spektrą, galimybę naudoti CNN/LSTM architektūras, leidžia stebėti mokymosi eigą realiu laiku ir yra tinkama integracijai į rekomendacines ar dinaminės kainodaros sistemas.

2.6.1. *Stable baseline3* skatinamojo mokymosi biblioteka

Stable Baselines – tai viena populiariausių atvirojo kodo skatinamojo mokymosi (RL) bibliotekų, paremta *OpenAI Baselines* pagrindu. Ji pasižymi ne tik pagerintu algoritmų įgyvendinimu, bet ir vieninga API struktūra, aukštesne kokybe, geresniu dokumentavimu bei aktyvia bendruomenės plėtra. Dėl šių priežasčių ši biblioteka plačiai taikoma tiek akademiniuose tyrimuose, tiek

praktiniuose pritaikymuose verslo srityse, įskaitant rekomendacines sistemas ir kainodaros optimizavimą.

Stable Baselines palaiko populiariausius RL algoritmus, tokius kaip A2C, PPO, DQN, SAC, TD3 ir kt. bei siūlo lankstų veiksmų ir būsenų tipų palaikymą – nuo diskrečių iki tolydžių (angl. *discrete*). Tokia įvairovė leidžia naudoti biblioteką labai skirtinguose scenarijuose: nuo robotų valdymo iki personalizuotų paslaugų ar dinamiškų e. komercijos sprendimų.

Svarbus šios bibliotekos privalumas integracija su *TensorBoard* biblioteka, leidžianti vizualiai stebėti mokymosi eigą, nuostolių funkcijas, atlygio kreives ir kitus metrikų pokyčius – tai itin vertinga vystant skaidrias ir patikimas RL sistemas. Be to, biblioteka suderinama su *OpenAI Gym* aplinkomis, todėl galima greitai realizuoti prototipus bei atlikti pakartotinius eksperimentus vienodu pagrindu. Atsižvelgiant į šio darbo specifiką – pasitikėjimo vertos rekomendacinės sistemos ir dinaminė kainodara – buvo pasirinkti tik tie algoritmai, kurie:

- Palaiko diskrečius veiksmus ir būsenas (reikalinga e. komercijos prekių ar vartotojų veiksmų modeliavimui).
- Nėra priklausomi nuo MPI.
- Turi stabilias ir pakartotinai patikrintas realizacijas.

Remiantis šiais kriterijais, atrinkti trys algoritmai: A2C, PPO2 ir DQN. Šie algoritmai pasižymi geru balansu tarp tyrimo paprastumo, mokymosi efektyvumo ir modelio interpretacijos galimybių. Jie buvo plačiai taikomi ir kituose panašiuose verslo kontekstuose, įskaitant personalizuotas nuolaidas, vartotojų elgsenos prognozes ir maržos optimizavimą.

2.6.2. *Stable baseline* algoritmai

Pasirinkti algoritmai išsiskiria tinkamumu dirbti su diskrečiais duomenimis, mokymosi stabilumu bei interpretacijos galimybėmis, todėl ypač tinka praktinėms verslo užduotims, tokioms kaip dinaminė kainodara ar personalizuotų rekomendacijų generavimas. Jie taip pat buvo sėkmingai pritaikyti kitose srityse – nuo elgsenos prognozės iki pelningumo optimizavimo, todėl laikomi tinkamu pagrindu šio tyrimo eksperimentams.

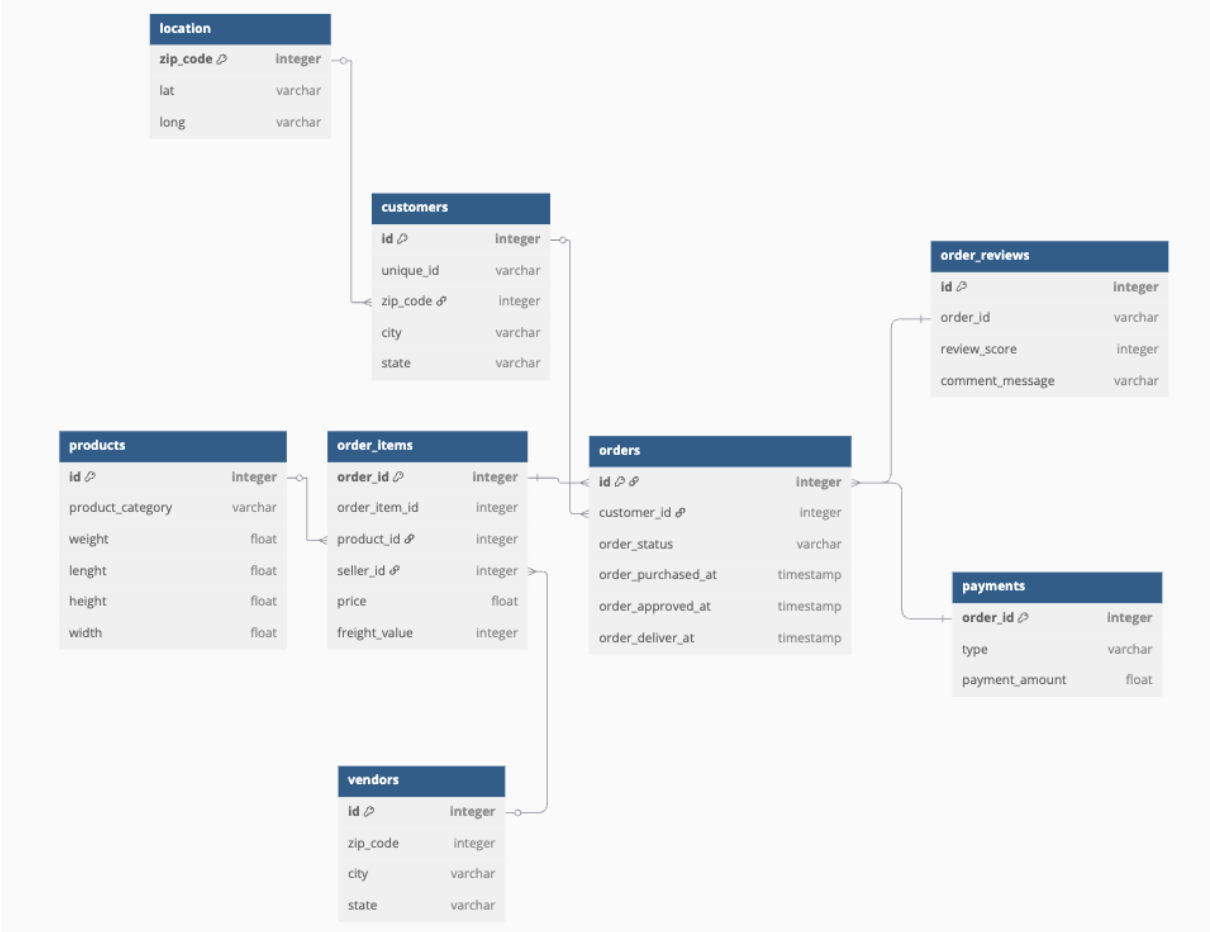
2 lentelė. Pasirinktų Stable baseline 3 algoritmų palyginimas.

Algoritmas	**A2C**	**PPO2**	**DQN**
Tipas	Aktoriaus-kritiko	Strategijos optimizavimas	Q-funkcijos
Pritaikymas	Efektyvus mokymasis su mažais duomenų kiekiais; tinka diskrečioms ir tolydžioms veiksmų erdvėms	Stabilus mokymasis įvairiose aplinkose, ypač diskrečiose	Diskrečios veiksmų erdvės uždaviniai, pvz., kainodara ar rekomendacijos
Pagrindinės savybės	Sinchroniškas atnaujinimas; stabilumas ir greitas konvergavimas	Be MPI; didelis stabilumas ir efektyvumas	Patirties atmintis, tikslinis tinklas, ε-godžioji strategija
Stabilumas	Vidutinis	Aukštas	Vidutinis-žemas
Reikalavimas atminčiai	Vidutiniai	Žemi	Aukšti
Tinkamumas realaus laiko užduotims	Taip	Taip	Ribotas
Veiksmų erdvė	Diskreti / tolydi	Diskreti	Diskreti
Konvergencijos greitis	Vidutinis	Sudėtingas	Vidutinis
Sudėtingumas įgyvendinti	Vidutinis	Sudėtingas	Vidutinis
Suprantamumas	Vidutinis	Ribotas	Aukštas
Naudojimo dažnumas praktikoje	Vidutinis	Dažnas	Dažnas

3. Tyrimo rezultatai

3.1. Pasirinkto duomenų rinkinio apžvalga

Šio tyrimo pagrindą sudaro realūs 2017-2019 m. elektroninės parduotuvės duomenys, kuriuose fiksuojami klientų užsakymai, apmokėjimai, produktų savybės, logistikos informacija ir klientų atsiliepimai. Šie duomenys leidžia visapusiškai modeliuoti pardavimo procesą – nuo prekės pasirinkimo iki jos pristatymo bei kliento įvertinimo. Tokia duomenų struktūra ypač vertinga, analizuojant vartotojo elgseną bei kuriant dinaminės kainodaros ir rekomendacinių sistemų modelius, pagrįstus skatinamojo mokymosi metodu.



16 pav. Elektroninės parduotuvės duomenų modelis.

Rinkinį sudaro keli tarpusavyje susiję duomenų failai, pateikiami reliacinės duomenų bazės struktūros pavidalu (žr. 16 pav.). Pagrindiniai duomenų šaltiniai yra šie:

- Užsakymai ir jų informacija – fiksuojama kada buvo pateiktas užsakymas, kokie produktai jame buvo, jų kainos, pristatymo terminai ir kt.
- Apmokėjimai – informacija apie apmokėjimo būdus, sumas, įmokų skaičių.
- Produktai – aprašomos produktų savybės, kategorijos ir matmenys.
- Pardavėjai – nurodo, kas išsiuntė prekę.

- Klientai ir jų geografiniai duomenys – leidžia įvertinti regioninius skirtumus ar pristatymo ypatybes.
- Klientų atsiliepimai – suteikia galimybę analizuoti klientų patirtį apsiperkant ir jo ryšį su kainodara bei logistika.

Duomenyse yra įtraukti laiko žymekliai (pvz., užsakymo data, pristatymo data), todėl galima formuoti laiko eilučių struktūrą, leidžiančią stebėti paklausos, kainų ir pardavimų dinamiką bėgant laikui. Ši savybė itin svarbi, siekiant taikyti RL algoritmus, nes laikiniai ryšiai yra esminiai priimant sekos pobūdžio sprendimus. Tačiau nors reliacinė duomenų schema yra optimali duomenų saugojimui ir užklausos vykdymui, skatinamojo mokymosi metodams reikalingas kitokio tipo duomenų pateikimas – t.y., kiekvienas agento mokymosi žingsnis turi būti pateikiamas kaip vientisa stebėjimo eilutė su visais atributais. Todėl buvo atlikta duomenų denormalizacija, kurios metu visos susijusios lentelės buvo sujungtos į vieną išplėstinę duomenų struktūrą, kurioje kiekviena eilutė reprezentuoja vieną agento sprendimo momentą. Šis procesas užtikrino duomenų paruošimą RL simuliacijai, leidžiantis imituoti agento veikimą realiomis el. prekybos sąlygomis bei analizuoti sprendimų poveikį ilgalaikiam verslo tikslų siekimui.

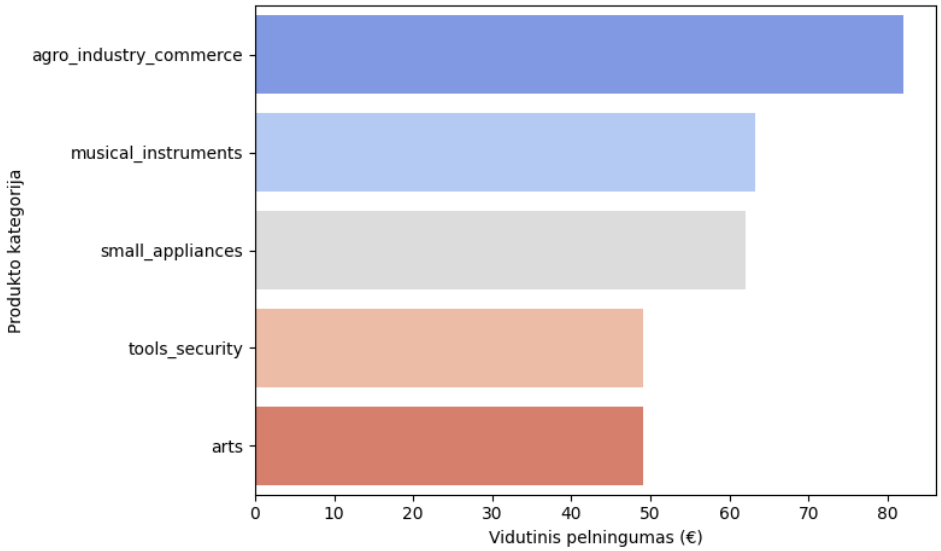
3.2. Duomenų paruošimas darbui

Norint pritaikyti skatinamojo mokymosi algoritmus, kuriant tikrus el. komercijos scenarijus, būtina užtikrinti, kad duomenys būtų ne tik teisingi, bet ir atspindintys realias verslo situacijas bei tendencijas. Kadangi agentas mokosi priimti sprendimus remdamasis grįžtamuoju ryšiu, bet kokie netikslumai ar neatitikimai gali lemti neteisingų strategijų išmokimą. Dėl šios priežasties duomenų ruošimas tapo vienu esminių etapų tyrimo sėkmei.

17 pav. Duomenų paruošimo darbui eiga.

Pirmiausia buvo atliktas pirminių duomenų ištraukimas bei sujungimas iš kelių tarpusavyje susijusių lentelių. Po to sekė išsami duomenų analizė, kurios metu buvo nustatytos neigiamos produktų savikainos, kurios buvo pakeistos, priskiriant joms 90% produkto vidutinės kainos reikšmę.

Kategorijų pavadinimai buvo išversti į lietuvių kalbą bei suskirstyti į pagrindines kategorijas, siekiant sumažinti klasifikavimo šakotumą. Išskirtys buvo sumažintos pasitelkiant slenkamojo vidurkio filtravimą. Tam, kad modelis galėtų aptikti sezoninius dėsningumus, duomenys buvo agreguoti į savaitinius vienetus pagal prekių kategorijas. Tokiu būdu buvo sumažintas triukšmas ir padidintas duomenų tankis.



18 pav. TOP 5 kategorijos pagal vidutinį pelningumą

Kadangi realių istorinių duomenų kiekis buvo ribotas, jie buvo paversti į savaitinius, o papildomai pritaikytas sezoninės naivios prognozės metodas, leidęs simuliuoti papildomus penkerių metų duomenis. Šis metodas leido generuoti papildomas laikines eilutes pagal ankstesnių metų pasikartojančius sezoninius dėsningumus bei užpildyti trūkstamas savaites esamuose duomenyse.

Kiekvienai prekių kategorijai buvo priskirta tipinė kainų elastingumo reikšmė, atspindinti vartotojų jautrumą kainos pokyčiams. Šios reikšmės nustatytos remiantis autoriaus praktine patirtimi e. komercijos srityje, įvertinant prekių pobūdį ir verslo logiką. Pavyzdžiui, kasdienio vartojimo prekėms, tokioms kaip maisto ar higienos produktai, priskirtas mažesnis elastingumas, o ne pirmo būtinumo ar prabangos prekėms – didesnis. Svarbu pažymėti, kad tikslus elastingumo nustatymas reikalautų atskiro ekonometrinio tyrimo, todėl šiame darbe jis naudojamas supaprastintu būdu, kaip vienas iš dinaminės kainodaros komponentų.

Vartotojo elgsena šiame tyrime buvo modeliuojama per paklausos funkciją kuri įvertina, kaip keičiasi vartotojų pirkimo tikimybė priklausomai nuo kainos. Kitaip tariant, paklausa šiame kontekste atstojo agento aplinkos dinamiką, imituodama galimą vartotojų reakciją į kainų pokyčius.

Formaliai, paklausa buvo skaičiuojama pagal šią formulę:

$$demand = base_{demand} * \left(\frac{price}{price_{base}}\right)^{price_{elasticity} * \alpha}, \quad (18)$$

čia α – elastingumo stiprumo moduliatorius, padedantis reguliuoti modelio jautrumą kainų pokyčiams. Ši formulė naudota tiek simuliuojant pardavimus, tiek skaičiuojant realią paklausą modelio veikimo metu.

Galiausiai, apdoroti duomenys buvo padalinti į apmokymo ir vertinimo rinkinius. Tik po šio paruošimo etapo buvo galima pradėti modelių mokymą su skatinamojo mokymo algoritmais, užtikrinant, kad kiekvienas modelis turėtų pakankamai informacijos apie paklausos dinamiką, sezoniškumą bei kainų politiką.

3.3. Skatinamojo mokymosi aplinkos kūrimas

Vienas svarbiausių skatinamojo mokymosi tyrimo etapų – tinkamos ir realistiškos mokymosi aplinkos sukūrimas. Kadangi pagrindinis šio darbo tikslas buvo ištirti patikimą skatinamojo mokymosi modelių elgseną dinaminėje kainodaroje, buvo sukurta specializuota mokymosi aplinka, kuri imituoja e. prekybos kontekstą bei leidžia vertinti modelių sprendimų logiškumą verslo atžvilgiu.

Sukurta aplinka paremta *OpenAI Gymnasium* pagrindu, pritaikyta naudoti su tokiais algoritmais kaip DQN, PPO ir A2C. Ji veikia epizodų pagrindu – kiekvienas epizodas atitinka metus arba 52-iejų savaičių laikotarpį, o kiekvienas žingsnis atitinka konkrečios kategorijos kainodaros sprendimą konkrečią savaitę.

Aplinkos būsena (angl. *observation*) sudaroma iš skaitinių rodiklių (vidutinė kaina, bazinė kaina, paklausa, elastingumas, sandėlio kiekis ir kt.) bei kategorinės informacijos apie prekę, kuri koduojama naudojant “OneHotEncoding” .Tai leidžia algoritmui „matyti“ ne tik skaitines reikšmes, bet ir skirtingų prekių klases.

Veiksmų erdvė (angl. *action space*) apibrėžia galimus kainos pokyčius nuo –20 % iki +20 %, 5 % intervalo didėjimu, o atlygio funkcija sukurta taip, kad atspindėtų pelningumą, verslo logiką bei pirkėjų elgsenos niuansus. Siekdami patikimumo, į atlygį įtraukėme papildomus kriterijus:

- **Sandėlio efektyvumo skatinimas.** Premiją už efektyvų sandėlio išpardavimą ir baudą už sandėlio kaupimąsi, skatindami balansą tarp pelno ir atsargų valdymo;
- **Paklausos jautrumo reakcijos.** Nuobaudas už žymius paklausos sumažėjimus, ypač kai sprendimai nėra racionalūs (pvz., paklausa krenta, bet kaina keliama). Buvo įvestos papildomos baudos už kainų didinimą esant mažai paklausai ir premija už agresyvią kainos mažinimo taktiką esant labai žemai paklausai;
- **Vartotojų elgsenos modeliavimas.** Pirkėjų nuovargio modeliavimą, kai ilgai keliant kainą iš eilės, paklausa krenta labiau nei įprastai (skatinamas kainos stabilumas ir atsargumas). Taip pat premija už nuoseklius sprendimus, jei prieš tai buvusi paklausa buvo maža, o kaina sumažinta;
- **Elastingumo logika**. Jei kainos pokytis neatitiko tikėtinos elastingumo reakcijos (pvz., paklausa greitai krenta, bet kaina keliama) – buvo taikoma bauda.

Tokia aplinkos struktūra leidžia ne tik optimizuoti trumpalaikį pelną, bet ir kurti modelius, kurie geba prisitaikyti prie ilgalaikių verslo tikslų ir logikos. Tai tampa esmine sąlyga tiriant patikimo skatinamojo mokymosi galimybes verslo aplinkoje – siekiama, kad agentas ne tik maksimizuotų atlygį, bet ir išmoktų elgtis logiškai, etiškai bei nuspėjamai. Tokiu atveju tikslo funkcija:

$$Maximize \sum_{t=1}^{t} R_t, \tag{19}$$

$$R_t = \pi_t - \lambda_s \cdot S_{left} + \beta_v \cdot Q_t + B_{stock} + B_{conversion} - P_{low_{demand}} \tag{20}$$
$$-P_{highprice} + B_{discount} + A_{fatigue} + B_{clearance},$$

čia:

- $\pi_t = (P_t - C_t) \cdot \mathrm{Q}$ - pagrindinis pelnas.
- $S_{left} = S_t - Q$ - likusios atsargos.

- λ_s – baudos koeficientas už atsargų kaupimąsi.
- Q_t – papildomas pelnas už didelį pardavimų kiekį.
- B_{Stock} – premija už pilną atsargų išpardavimą.
- $B_{conversion}$– bonusas už aukštą konversijos lygį.
- $P_{low_{demand}}$– bauda už ženklų paklausos kritimą.
- $\mathrm{P_{highprice}}$– bauda už neteisingą kainos didinimą esant žemai paklausai.
- $\mathrm{B_{discount}}$– premija už proaktyvų kainos mažinimą esant mažai paklausai.
- $\mathrm{A_{fatigue}}$– bauda už nuoseklius nepagrįstus kainų didinimus.
- $\mathrm{B_{clearance}}$– pelno bonusas už agresyvų sandėlio išvalymą.

3.4. Pasitikėjimo verto algoritmo paieška ir modelio apmokymas

Siekiant nustatyti efektyviausią ir patikimiausią skatinamojo mokymosi (RL) algoritmą dinaminės kainodaros užduočiai, buvo palyginti trys populiariausi modeliai: DQN, PPO ir A2C. Šie modeliai pasirinkti atsižvelgiant į jų skirtingą architektūrą bei gebėjimą dirbti su diskretine veiksmų erdve, kuri puikiai atspindi realią kainų keitimo situaciją – sprendimai priimami iš baigtinio veiksmų rinkinio (pvz., ±5 %, ±10 % pokyčiai).

Modeliai buvo apmokyti naudojant *MlpPolicy* – tai daugiasluoksnis perceptronas (angl. *multilayer perceptron*), kuris tinkamas struktūrizuotiems skaitiniams duomenims. Kadangi šio tyrimo duomenys neapima sekų (pvz., laikinių duomenų langų ar vaizdų), o kiekvienas sprendimas priimamas remiantis dabartine būsena (pvz., kaina, paklausa, sandėlis, elastingumas), *MlpPolicy* buvo optimalus pasirinkimas tiek dėl paprastumo, tiek dėl interpretavimo galimybių. Taip pat buvo svarbu išlaikyti aiškią architektūrą, kad būtų galima vertinti sprendimų logiką bei elgsenos dėsningumus.

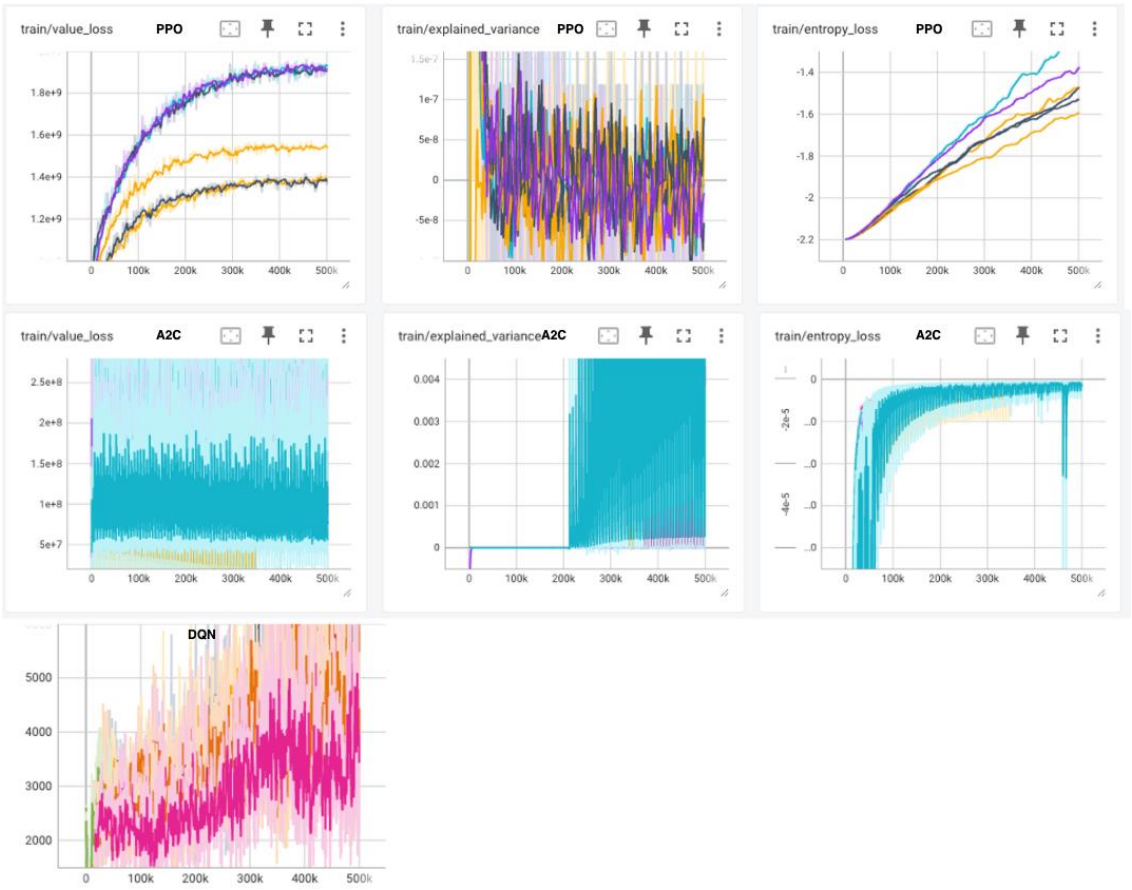
Visų modelių apmokymas buvo vykdomas naudojant vienodus hiperparametrus: 350 000 žingsnių, mokymosi greitį, atitinkantį 0.001 ir diskonto koeficientą γ=0.8. Sąmoningai pasirinkta mažesnė γ reikšmė leido modeliams labiau reaguoti į trumpalaikius pokyčius, kurie būdingi dinamiškai kintančiai paklausai. Naudoti hiperparametrai apibendrinti lentelėje apačioje:

3 lentelė. Hiperparametrų rinkinys kiekvienam algoritmui.

Algoritmas	**DQN**	**PPO**	**A2C**
Mokymosi žingsnis	0.001	0.001	0.001
γ	0.8	0.8	0.8
Veiksmų erdvė	9 veiksmai (nuo -20% iki +20%)	9 veiksmai (nuo -20% iki +20%)	9 veiksmai (nuo -20% iki +20%)
Apmokymo žingsnių skaičius	350 000	350 000	350 000
Politikos tipas	MlpPolicy	MlpPolicy	MlpPolicy

Modelių palyginimui buvo naudotas slankiojo atlygio vidurkis (angl. *rolling average reward*), skaičiuojamas 50 epizodų lange. Kaip parodyta 18 paveiksle, PPO ir A2C modeliai konvergavo nuosekliai, tačiau jų priimami sprendimai dažnai buvo konservatyvūs – modeliai vengė didesnių kainų pokyčių net ir esant galimam pelnui. Tuo tarpu DQN modelis pasižymėjo aktyvesniu veiksmų

tyrinėjimu ir platesniu kainodaros strategijų spektru, kas yra itin svarbu dinamiškai kintančios paklausos sąlygomis.



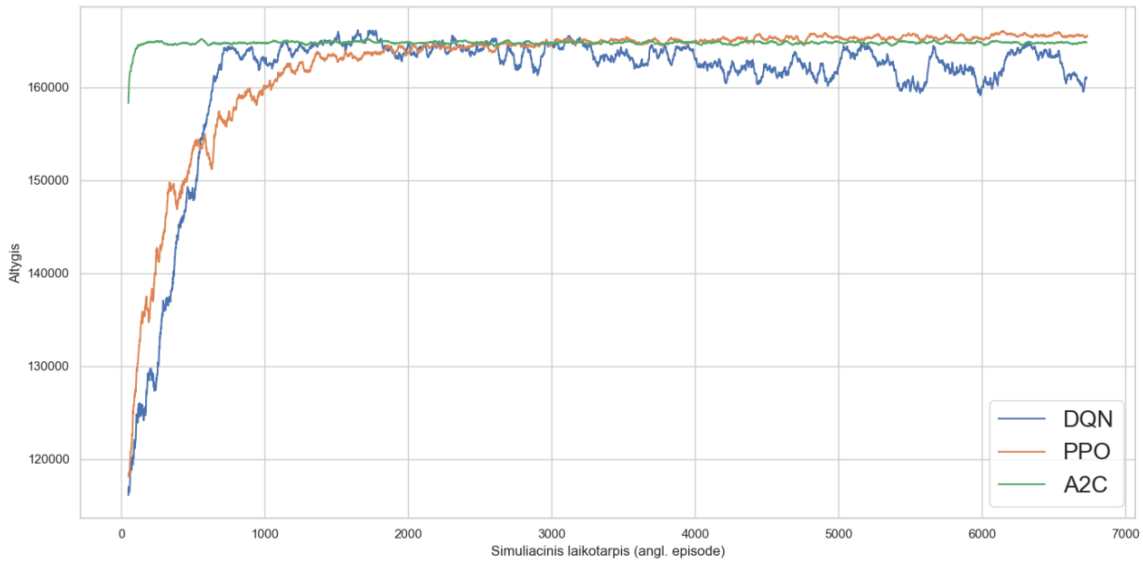
19 pav. Gauti grafikai su TensorBoard kiekvienam modeliui

Kaip parodyta 19 pav., PPO ir A2C modeliai, kurie remiasi aktoriaus–kritiko architektūra, pateikia tris atskiras metrikas: nuostolį pagal tikslo funkciją (angl. *value loss*), paaiškintos dispersijos rodiklį (angl. *explained variance*) ir entropijos nuostolį (angl. *entropy loss*). Tuo tarpu DQN modelis, kuris optimizuoja Q-funkciją, generuoja tik vieną bendrą apmokymo nuostolio kreivę, nes neturi atskiro vertinimo (kritiko) ar politikos komponentų.

Toliau pateikiamas pagrindinių treniravimo metrikų paaiškinimas:

- *Tikslo funkcijos nuostolis.* A2C ir PPO modelių metrikos parodė didelį nepastovumą. Kai kuriais atvejais reikšmės išliko aukštos, kas signalizuoja apie nepakankamai efektyvų kritiko (vertės prognozavimo) komponento mokymąsi. Tai riboja strategijos tikslumą ir gali turėti neigiamą įtaką modelio sprendimų kokybei.
- *Paaiškintos dispersijos rodiklis.* Ši metrika parodo, kiek kritikas geba paaiškinti būsimos vertės variaciją. Daugelyje treniravimo sesijų šis rodiklis išliko beveik lygus nuliui arba labai svyravo, kas rodo, kad kritiko mokymasis buvo ribotas. Tokioje situacijoje strategijos mokymosi procesas yra grindžiamas netiksliais signalais, o tai mažina modelio patikimumą ir jo supratimą.

- *Entropijos nuostolis*. Ši metrika leidžia vertinti veiksmų įvairovę treniruotės metu. PPO modelyje entropijos nuostolio augimas rodo didesnį tyrinėjimą, tačiau tai gali sukelti sprendimų nestabilumą. Tuo tarpu A2C modelyje šis rodiklis greitai stabilizuojasi, kas reiškia, kad modelis tampa nuspėjamas – renkasi tuos pačius veiksmus ir mažina lankstumą.
- *DQN apmokymo nuostolis*. DQN modelyje pateikiama tik viena bendroji apmokymo nuostolio reikšmė, kuri atspindi Q reikšmės prognozės klaidą. Matomas didelis triukšmas ir kintamumas, ypač vėlesniuose treniravimo etapuose. Tai rodo, kad DQN modelis aktyviai koregavo savo Q reikšmes – tai gali reikšti tiek didesnį prisitaikymo potencialą, tiek mokymosi nestabilumą.



20 pav. Pasitikėjimo verto modelio parinkimas.

Kaip matyti 20 pav. PPO modelis pasiekė aukščiausią atlygio lygį ir stabiliai išlaikė jį treniravimo pabaigoje. Visgi pastebėta, kad šis modelis dažnai kėlė kainas net ir esant žemai paklausai, o tai galimai rodo per didelį pasikliovimą netikslia vertės funkcija – tai prieštarauja patikimos politikos principui. A2C modelis veikė labai konservatyviai – fiksuotas atlygis, nekeitė kainos, kas sumažino jo adaptacijos galimybes ir potencialią grąžą. Abu šie elgsenos kraštutinumai (perdėtas agresyvumas ar per didelis atsargumas) nėra tinkami patikimos dinaminės kainodaros kontekste.

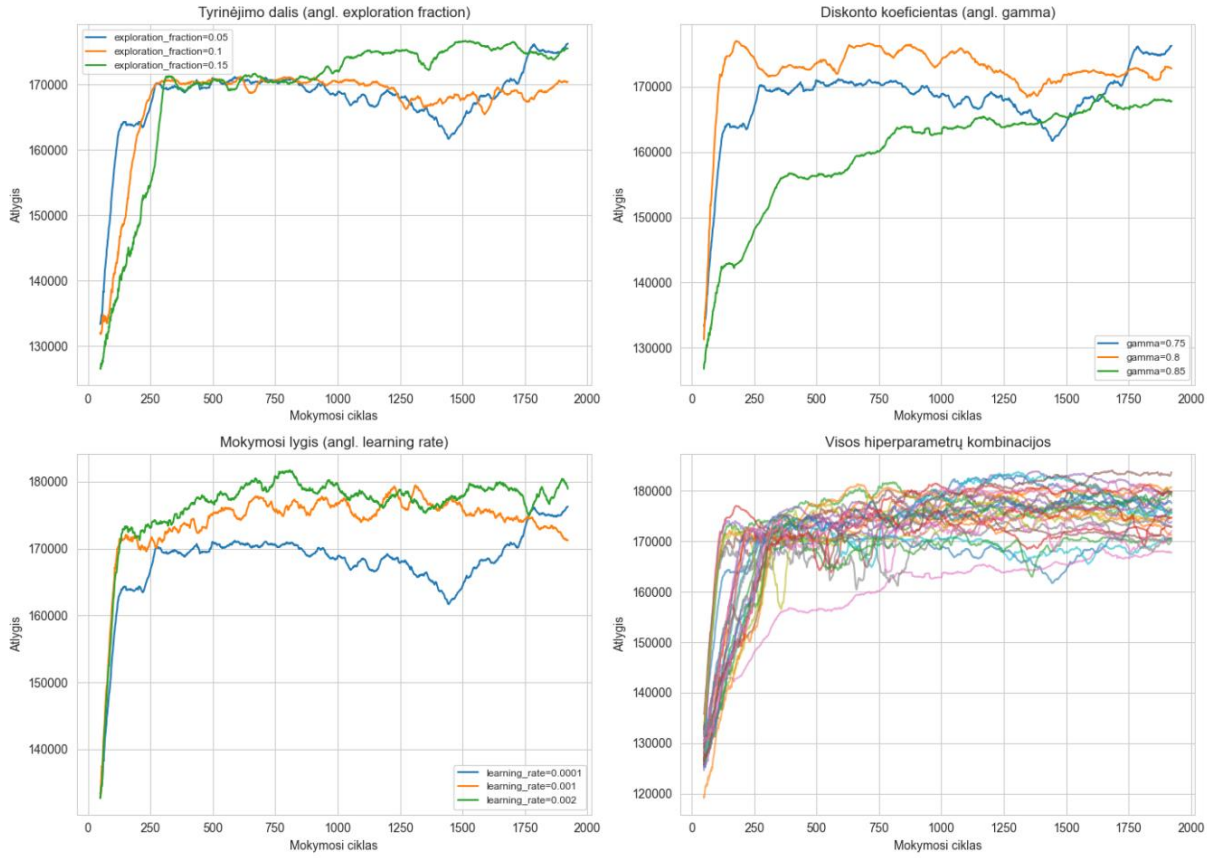
DQN, nors ir technologiškai paprastesnis, demonstravo subalansuotą tyrinėjimo ir išnaudojimo (angl. *exploration–exploitation*) strategiją – adaptavosi prie paklausos pokyčių, eksperimentavo su įvairiomis kainodaros taktikomis ir greičiau pasiekė atlygio pusiausvyrą. Tai leido jį identifikuoti kaip labiausiai tinkamą modelį tolimesniam taikymui.

Vertinant pagal patikimo skatinamojo mokymosi principus – modelio stabilumą, elgsenos paaiškinamumą, adaptaciją prie pokyčių ir verslo logikai atitinkančius sprendimus – DQN modelis pasirodė esąs tinkamiausias kandidatas. Jo sprendimų įvairovė, gebėjimas tyrinėti ir prisitaikyti prie aplinkos rodo didesnį pasitikėjimo potencialą praktinėje kainodaros optimizacijoje.

3.5. Modelio optimizavimas ir geriausių hiperparametrų paieška

Nustačius, kad DQN algoritmas yra tinkamiausias šio darbo uždaviniui, buvo pereita prie tolimesnio žingsnio – hiperparametrų optimizavimo. Skirtingai nei tradiciniuose mokymo metoduose, skatinamojo mokymosi algoritmų našumas stipriai priklauso nuo tinkamai sukonfigūruotų parametrų, tokių kaip diskonto koeficientas, mokymosi greitis ar tyrinėjimo lygis.

Atliekant optimizaciją buvo naudojama rankiniu būdu sukonstruota parametrų gardelė, kurios tikslas buvo įvertinti įvairias parametrų kombinacijas fiksuotoje aplinkoje. Kiekviena kombinacija buvo testuojama atskirai, o modelis apmokytas 200 00 žingsnių. Vertinimui buvo naudota slankiojo vidurkio pagalba sušvelninta atlygio kreivė, leidžianti palyginti, kaip agentas mokosi per epizodus.



21 pav. Modelio hiperparametrų paieška.

Bendras visų kombinacijų kiekis siekė 27, o gauti rezultatai pateikti vizualiai (žr. 21 pav.). Iš jų matyti, kad didesnis tyrinėjimo lygis agentui padėjo iš pradžių pažinti pačią aplinką, dėl ko pasiektas aukštesnis galutinis atlygis, tačiau mažesnis tyrinėjimo lygis lėmė greitesnę konvergenciją. Tuo tarpu, diskonto koeficientas (γ) ties 0.8 reikšmė, pasirodė optimaliausia, kuris užtikrino balansą tarp ateities naudos ir dabartinio atlygio. Mokymosi žingsnis (angl. learning rate) atskleidė, kad per maža reikšmė (0.0001) leido labai lėtą Q reikšmių atnaujinimą ir prastą mokymąsi. Atsižvelgiant į stabilumą ir bendrą našumą, geriausia kombinacija, kuri užtikrino ir pakankamą tyrinėjimą, ir gerą ilgalaikės naudos įvertinimą be didelio nepastovumo, buvo ši:

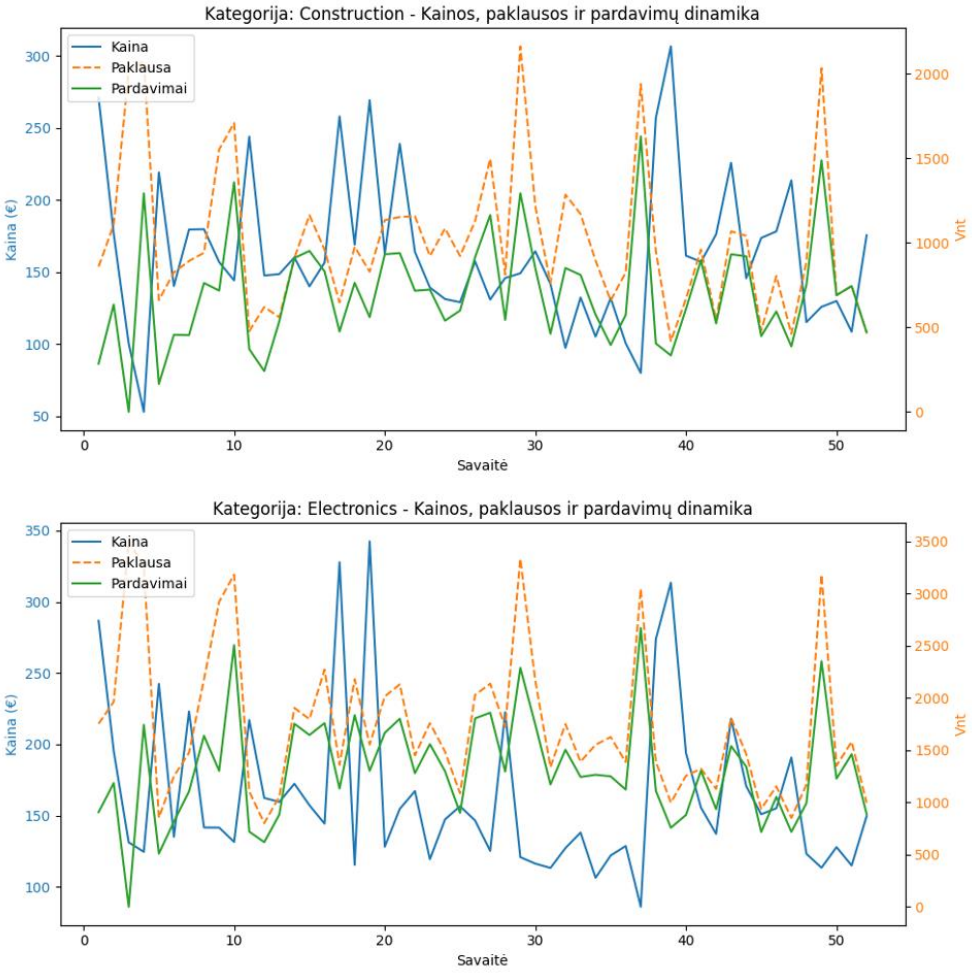
4 lentelė. Geriausių hiperparametrų rinkinys

	DQN
Mokymosi žingsnis	0.001
γ	0.8
Tyrinėjimo dalis	0.15
Tyrinėjimo paskutinis ciklas	0.01
Politikos tipas	MlpPolicy

3.6. Dinaminės kainodaros pritaikius RL rezultatai

3.6.1. Dinaminė kainodara skirtingoms kategorijoms

Siekdami įvertinti skatinamojo mokymosi pagrindu veikusio agento elgseną skirtingų prekių kategorijų kontekste, buvo išanalizuotos savaitinės dinamikos kreivės, atvaizduojančios kainų, paklausos bei realiai parduotų kiekių kitimą. Šie grafikai leidžia suprasti, kaip modelis prisitaikė prie skirtingų segmentų rinkos elgsenos, kiek nuosekli buvo jo kainodaros strategija bei ar pasiektas norimas paklausos ir pasiūlos balansas.



22 pav. Dinaminė kainodara statybų ir elektronikos prekėms.

Analizuojant kategorijų lygmeniu, galima išskirti kelias ryškesnes tendencijas:

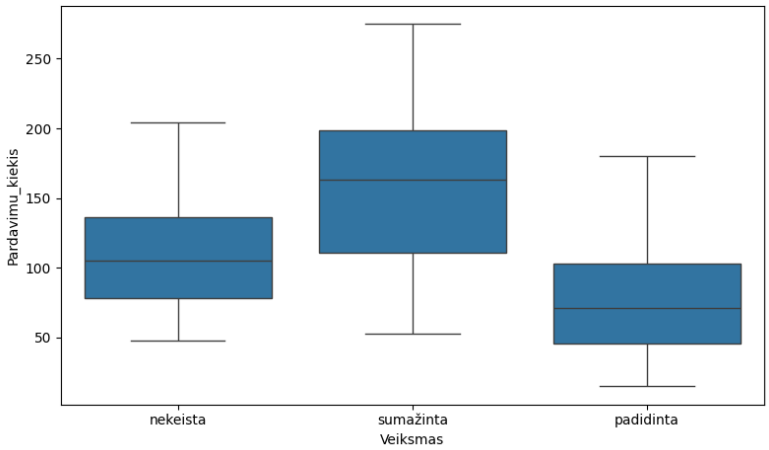
- Mados, elektronikos ir laisvalaikio kategorijose agentas taikė aktyvią dinaminę kainodarą – kainos buvo dažnai keičiamos, siekiant išnaudoti trumpalaikius paklausos svyravimus. Pastebimas glaudus ryšys tarp kainos pokyčių ir pardavimų augimo: sumažinus kainą, pardavimai dažnai šoktelėdavo, o tai rodo, kad modelis teisingai interpretavo elastingumo signalus. Šiose kategorijose agentas dažniausiai veikė strategiškai ir efektyviai.
- Namų ir statybos kategorijose agento elgsena buvo nepastovi, su dažnais kainos šuoliais, tačiau silpnesniu ryšiu su pardavimų dinamika. Tai leidžia daryti prielaidą, kad šiose rinkose elastingumo įvertinimas buvo mažiau tikslus, o pirkėjų elgesį dažniau lėmė kiti veiksniai – pvz., sezoniškumas, ilgalaikio poreikio pobūdis ar tiekimo specifika. Nepaisant bandymų, kainų strategijos nebuvo tokios aiškiai veiksmingos kaip kitose srityse.
- Asmeninės priežiūros kategorijoje agentas laikėsi stabilios ir žemos kainodaros strategijos, išlaikydamas nuosaikius kainų pokyčius per visą laikotarpį. Pardavimų ir paklausos kreivės rodo aukštą ir stabilų vartojimą, kas patvirtina, kad modelis atpažino šios kategorijos žemą elastingumą. Tokia strategija – mažai rizikinga, bet efektyvi – padėjo užtikrinti pastovų pelną.

Svarbu pažymėti, kad kiekviename mokymosi žingsnyje agentas generavo naują kainą remdamasis bazine kaina ir savo pasirinktu veiksmu – tai buvo procentinis pokytis nuo esamos kainos. Ši strategija leido veiksmus pritaikyti dinamiškai, atsižvelgiant į tokius veiksnius kaip bazinė paklausa, atsargos, kainų elastingumas bei ankstesnių žingsnių rezultatai. Tokiu būdu agentas ne tik mokėsi iš savo klaidų, bet ir nuosekliai formavo savo kainodaros politiką – adaptavosi prie realios aplinkos, siekė maksimizuoti pelningumą ir užtikrinti paklausos atitikimą pasiūlai.

Apibendrinant, agentas pademonstravo gebėjimą prisitaikyti prie skirtingų kategorijų rinkos sąlygų, taikydamas diferencijuotas strategijas. Veikdamas vieningoje mokymosi aplinkoje, jis sugebėjo suformuoti konteksto atžvilgiu pritaikytą kainų politiką. Tai atspindi skatinamojo mokymosi esmę – gebėjimą mokytis iš grįžtamojo ryšio ir priimti sprendimus, pagrįstus ne iš anksto nustatyta logika, bet realia patirtimi. Taip pat palyginus su pradiniu baziniu modeliu, agentas padidino pelną 12.58%, o tai parodo aiškų praktinį metodo efektyvumą. Vis dėlto kai kuriose kategorijose agentui dar liko erdvės tobulinti elastingumo interpretaciją ir sprendimų nuoseklumą.

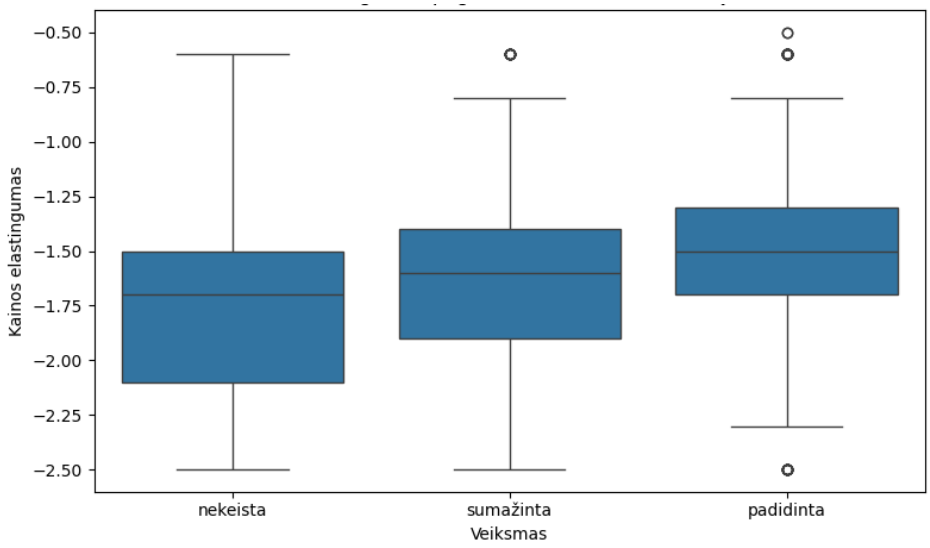
3.6.2. Skatinamojo mokymosi agento elgsenos analizė

Siekdami detaliau įvertinti agento, veikiančio dinaminės kainodaros aplinkoje, sprendimų pagrįstumą bei elgsenos stabilumą, atlikta papildoma rezultatų analizė, pasitelkiant kelias aiškinamąsias vizualizacijas. Tikslas – ne tik įvertinti pasiektus rezultatus, bet ir pažiūrėti, ar sprendimai buvo logiškai pagrįsti, ar agentas prisitaikė prie aplinkos bei ar įmanoma jį laikyti patikimu sprendimų rekomendavimo įrankiu.



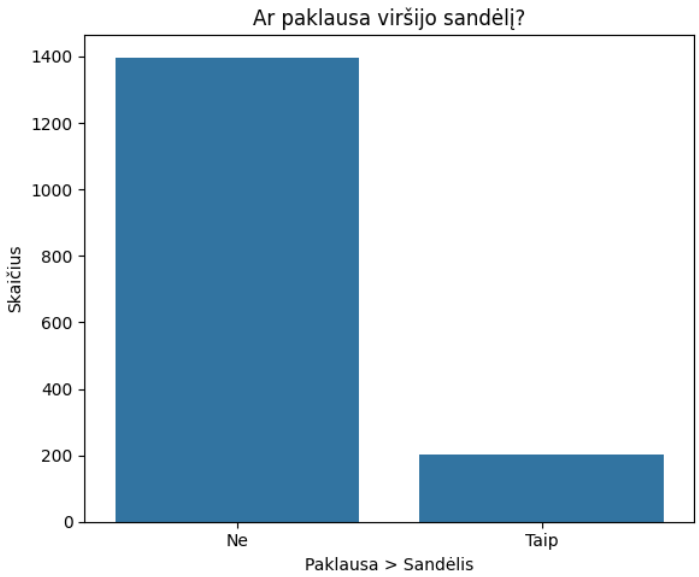
23 pav. Paklausa prieš kainos keitimą.

Viršuje pateikiamas paveikslėlis rodo, kad kainos mažinimo veiksmai dažniausiai buvo taikomi esant aukštesnei paklausai, o tai leidžia manyti, jog agentas siekė išnaudoti didesnę paklausą sandėlio ištuštinimui ir trumpalaikio pelno maksimizavimui. Tai atitinka atlygio funkcijoje įtrauktą paskatą už efektyvų sandėlio valymą. Tuo tarpu kainos didinimas dažniau pasitaikė žemos paklausos metu, kas gali reikšti ne visada optimalų elastingumo interpretavimą.



24 pav. Prekės kainos elastingumas pagal veiksmą.

Taip pat, agentas kainos keitimo sprendimus iš dalies derino prie elastingumo – kai elastingumas buvo stipresnis (t. y. labiau neigiamas), dažniau buvo pasirenkama mažinti kainą, o kainos didinimo sprendimai dažnesni esant mažesniam elastingumui (artimesniam nuliui). Tai rodo, kad agentas gebėjo bent iš dalies įvertinti paklausos jautrumą kainai ir reaguoti strategiškai, nors kai kuriose situacijose sprendimai vis dar buvo ne visai nuoseklūs.

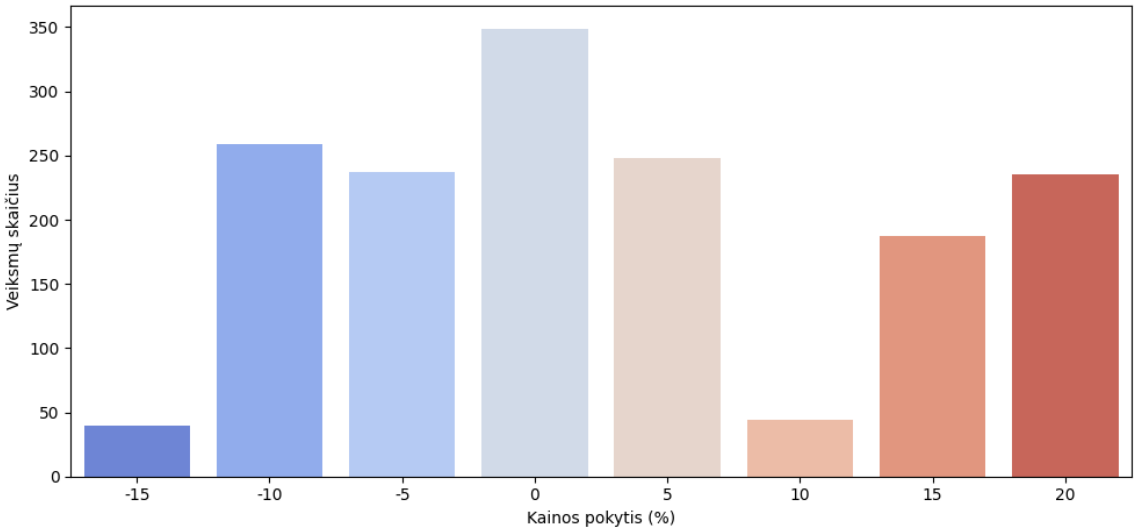


25 pav. Paklausos palyginimas su sandėlio likučiu.

25 pav. rodo, kad tik apie 15 % epizodų paklausa viršijo turimas atsargas sandėlyje. Tai leidžia daryti išvadą, kad agentas veikė gana atsargiai – galimai vengė situacijų, kai paklausa negalėtų būti patenkinta, taip sumažindamas klientų nepasitenkinimo ar prarastų pardavimų riziką. Tokia elgsena rodo išlaikytą pusiausvyrą tarp pelno siekimo ir veiklos tvarumo.

Tolesnė analizė atskleidžia ir kitus svarbius aspektus. Neetiškų veiksmų, kai agentas keltų kainas esant mažai paklausai, neužfiksuota (0 %), o tai rodo, kad modelis laikėsi pagrindinių sąžiningumo principų. Kainos krypties pokyčių dažnis siekė 51,84 %, kas leidžia daryti prielaidą, kad agentas reguliariai koregavo kainas pagal aplinkos signalus, tačiau nepersistengė – tai svarbu siekiant išvengti nenuoseklumo ar vartotojų painiavos.

Taip pat pažymėtina, kad didelės kainos korekcijos (virš +10 %) sudarė 28,83 % visų veiksmų, o tai gali rodyti kartais agresyvesnę strategiją maksimizuoti pelną tam tikrose situacijose. Galiausiai, pakartotinų veiksmų dalis siekė 23,64 %, o tai rodo vidutinį sprendimų įvairumą – agentas išlaikė lankstumą, bet taip pat nesiblaškė tarp visiškai atsitiktinių veiksmų. Šie rodikliai leidžia manyti, kad agento strategija buvo gana subalansuota – derinanti adaptavimąsi prie situacijos su atsargia veikimo logika bei etinių sprendimų išlaikymu.

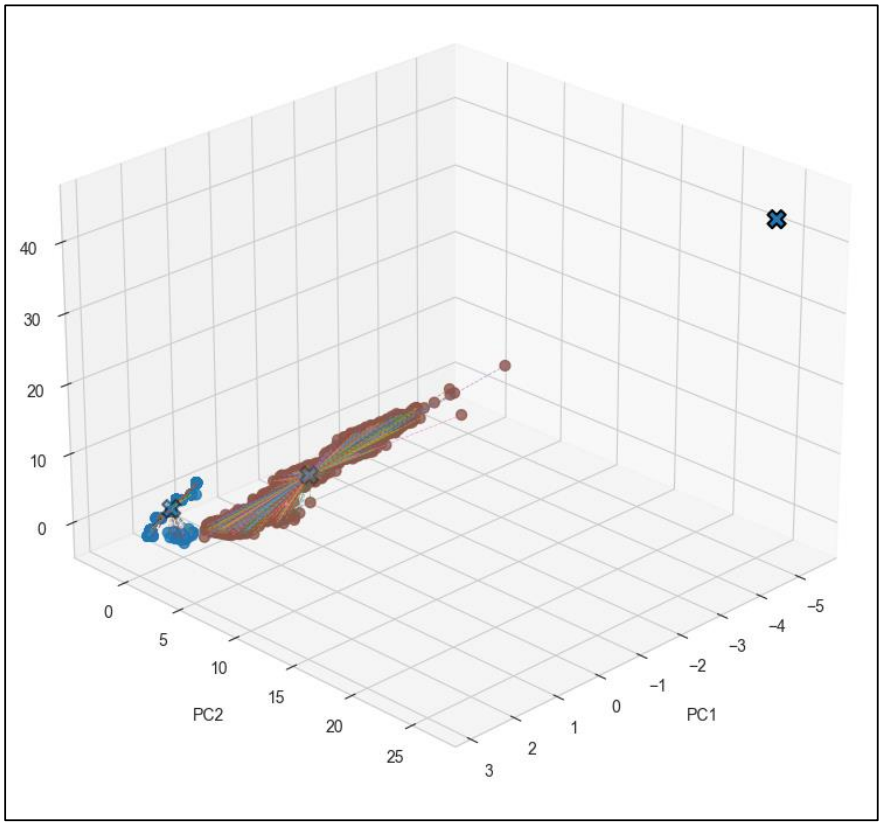


26 pav. Kainų keitimo veiksmų pasiskirstymas.

Dauguma veiksmų – kainos pakėlimas (58,8 %) ir tik 10,6 % atveju buvo pasirinkta kainą sumažinti. Dažniausiai buvo renkamasi nedideli (5-10 %) pokyčiai, kas rodo santykinai konservatyvią strategiją.

Prieš klasterizaciją atlikome komponentų tarpusavio ryšių analizę. Kadangi Shapiro–Wilk‘o normalumo testas atmetė daugumos atlygio komponenčių normalumo hipotezes ($p < 0.05$ net 11 iš 12 komponentų), nusprendėme naudoti Spearmano ranginę koreliaciją, kuri nereikalauja duomenų normalumo.

Toliau, siekdami geriau suprasti apmokyto RL agento sprendimų logiką bei veiksmų erdvinį pasiskirstymą, atlikome papildomą analizę, remdamiesi agento Q reikšmėmis. Po kiekvieno žingsnio gautos Q reikšmės reprezentuoja agento pasitikėjimą kiekvienu galimu veiksmu konkrečioje būsenoje. Surinkę šiuos duomenis, juos sumažinome taikydami **PCA** metodą į tris pagrindinius komponentus (PC1–PC3) ir vizualizavome 3D erdvėje.

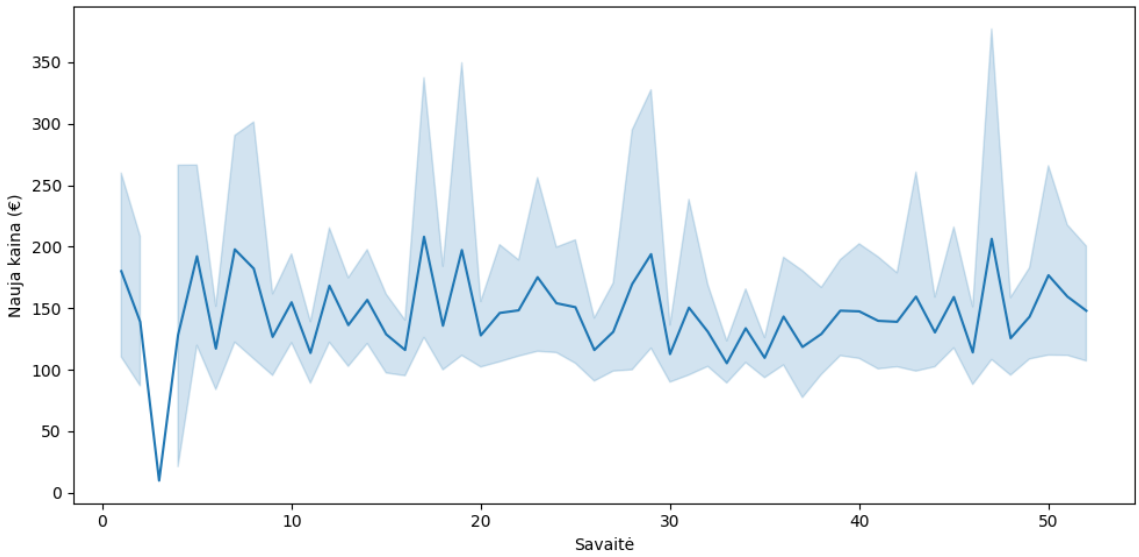


27 pav. 3D PCA klasterizacija (k = 3)

Kaip matyti 3D grafike, Q reikšmių erdvė formuoja tris aiškiai atskiras grupes – tai rodo, jog agentas išvystė skirtingas strategijas priklausomai nuo būsenos savybių. Kiekybiniam klasterių nustatymui pritaikėme k-vidurkio klasterizaciją k = 3, o kiekvieną tašką priskyrėme vienam iš trijų klasterių:

- **Klasteris 0 („Kasdieninės prekės“).** Daugiausia epizodų su higienos, vaikų drabužių, virtuvės ir šventinių reikmenų transakcijomis. Agentas veiksmo dažniausiai rinkosi mažinti kainą (nuolaidos), o vidutinis atlygis svyravo apie **~2 300 €**.
- **Klasteris 1 („Namų ir stiliaus prekės“).** Dominuoja namų interjero, grožio-sveikatos, kūdikių prekių ir aksesuarų pardavimai. Agentas dažnai naudojo nedidelius kainų padidinimus, o vidutinis atlygis siekė **~4 100 €**.
- **Klasteris 2 („Menas ir meno reikmenys“).** Visi epizodai – meninės veiklos prekės („arts“ kategorija). Čia agentas dažniausiai taikė aukštesnius kainų kėlimus (iki +10 %), o vidutinis atlygis buvo **~3 700 €**.

Tokiu būdu PCA 3D klasterizacija padėjo identifikuoti tris pagrindines agento strategijas: įprastas nuolaidas bei buities prekių orientaciją, namų ir stiliaus segmentų kainų valdymą bei meninės vertės produktų kainų koregavimą.



28 pav. Kainos pokytis per laikotarpį.

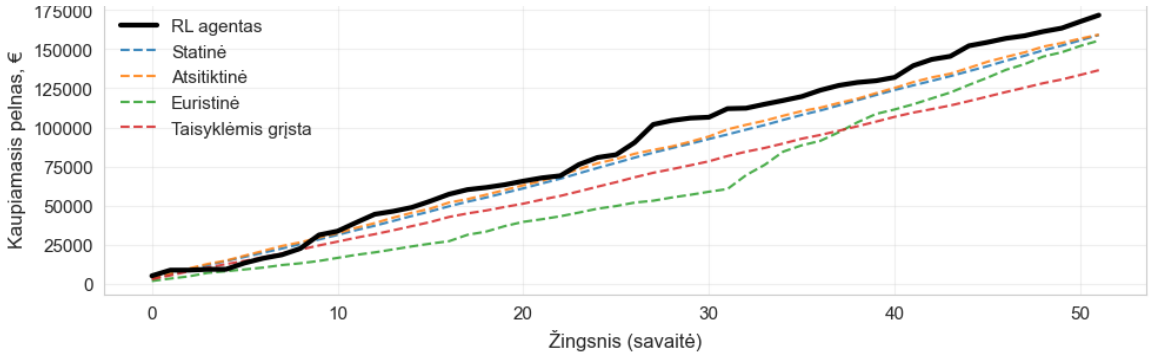
Iš grafiko matome, kad kainos kito palaipsniui, be staigių šuolių, o tai leidžia manyti, kad agento veiksmai buvo orientuoti į ilgalaikį pelningumą, o ne atsitiktinius pokyčius.

Atlikta analizė rodo, kad skatinamojo mokymosi pagrindu veikiantis agentas ne tik gebėjo taikyti diferencijuotą kainodaros strategiją pagal kontekstą, bet ir formavo sprendimus, kurie remiasi logišku verslo pagrindu. Nors kai kurios kategorijos (pvz., namų ar statybų) išliko neaiškios pelningumo prasme, dauguma veiksmų rodo ryšį tarp elastingumo, paklausos ir kainos. Vizualizacijos parodė, kad agentas veikė strategiškai, retai viršijo sandėlio paklausą, o kainos pokyčius vykdė palaipsniui. Tai leidžia teigti, kad nagrinėtas agento modelis atitinka patikimumo kriterijus, reikalingus rekomendacinėms sistemoms, taikančioms dinaminę kainodarą.

3.6.3. Skirtingų kainodaros strategijų palyginimas

Norėdami patikrinti, ar sukurtas skatinamojo mokymosi agentas iš tiesų kuria apčiuopiamą verslo vertę, jo veikimą palyginome su keturiomis intuityviomis alternatyvomis:

- **Statinė** strategija, kai viso epizodo metu taikoma vidutinė produkto kaina;
- **Atsitiktinė** strategija, kai kaina kas savaitę atsitiktinai svyruoja ± 5 % aplink vidurkį;
- **Euristinė** („žmogiška“) logika, kai sandėlio perteklius skatina − 5 % nuolaidą, o atsargų trūkumas → + 5 % antkainį, tačiau visada saugomas bent 5 % maržos rezervas;
- **Taisyklėmis grįsta** „if–else“ sistema, kuri tiesiogiai reaguoja į paklausos ir atsargų santykį (didelės atsargos → kainą mažiname, didelė paklausa → kainą keliam).



29 pav. Skirtingų kainodarų palyginimas

Kiekviena strategija 30 kartų „prasuko“ visus metus trunkančius 52 savaičių epizodus. 29 pav**.** aiškiai rodo: RL agento kreivė kyla sparčiausiai ir metų pabaigoje vidutiniškai 8–12 % lenkia taisyklių rinkinį bei dar labiau – statinę ar atsitiktinę kainodarą. Euristinei metodikai sekėsi nenuosekliai: kai kuriuose epizoduose ji artėjo prie RL, tačiau kituose generavo nuostolių.

Kad skirtumas tarp kreivių nebūtų tik vizualus, atlikome keletą patikros žingsnių.

– **Normalumo patikra.** Skirtumų paskirstymui (RL − bazė) taikėme Shapiro–Wilk‘o testą. Jis rodė, kad duomenys nėra pasiskirstę pagal normalųjį skirstinį, todėl, be klasikinių *t*-testų, pritaikėme ir neparametrinį Friedman‘o testą.
- **Suporuoti t-testai.** Vienpusis testas (hipotezė H_0: RL pelnas nėra didesnis) patvirtino, kad RL statistiškai reikšmingai ($\alpha = 0{,}05$) lenkia tik taisyklėmis grįstą strategiją ($p = 0{,}004$), o likusių trijų bazinių metodų nepagerina patikimu skirtumu.
- **Efekto dydis.** „Cohen“ $d \approx 0{,}5$, lyginant RL ir taisyklėmis grįstą strategiją, rodo vidutinio stiprumo praktinį poveikį – tai nėra atsitiktinis nuokrypis, o realus, verslui apčiuopiamas pranašumas.
- **„Bootstrap“ pasikliautinieji rėžiai** (95 %) parodė siauresnį RL intervalo plotį, t. y. mažesnį pelno dispersiją ir didesnį prognozuojamumą.
- **Friedman‘o testas** ($\chi^2 \approx 46{,}7$; $p < 0{,}001$) patvirtino, kad bent viena strategija iš esmės skiriasi – ir post-hoc analizė parodė, jog tai būtent RL agentas.

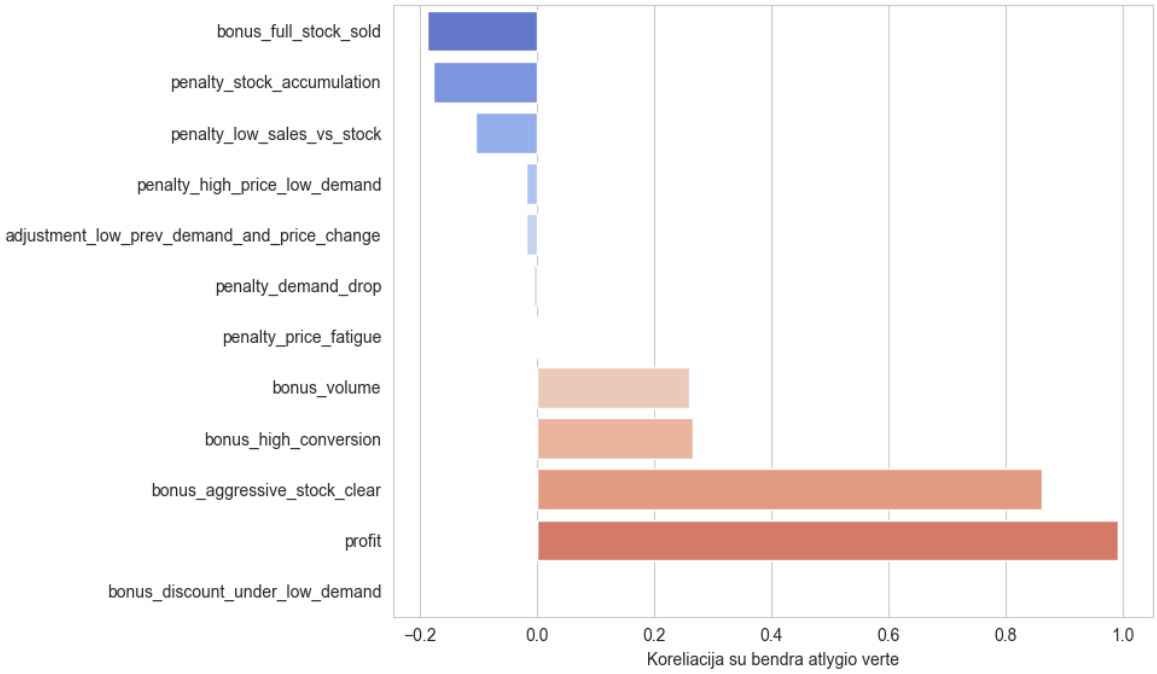
Trumpai tariant, iš keturių paprastų „rankinių“ metodų tik taisyklėmis grįstas algoritmas buvo realus konkurentas, tačiau ir jį RL agentas vidutiniškai aplenkė daugiau kaip 9 tūkst. € per epizodą. Statinė ir atsitiktinė strategijos atsiliko dar labiau, o euristinė – dėl agresyvių kainos korekcijų – patyrė itin didelę riziką, kai kuriuose epizoduose pereidama į nuostolingą zoną.

Taigi mokymosi modelis, sukonstruotas laikantis patikimo DI principų, ne tik racionaliai reaguoja į paklausos signalus, bet ir palaipsniui kaupia didesnį bei stabilesnį pelną už žmogaus sugalvotas taisykles. Tai rodo, jog RL agentą galima drąsiai svarstyti kaip realią, verslo požiūriu pagrįstą alternatyvą tradicinėms kainodaros schemoms, neprarandant modelio paaiškinamumo ir atitikimo etikos reikalavimams.

3.7. Pasitikėjimo vertos rekomendacinės sistemos tikrinimas

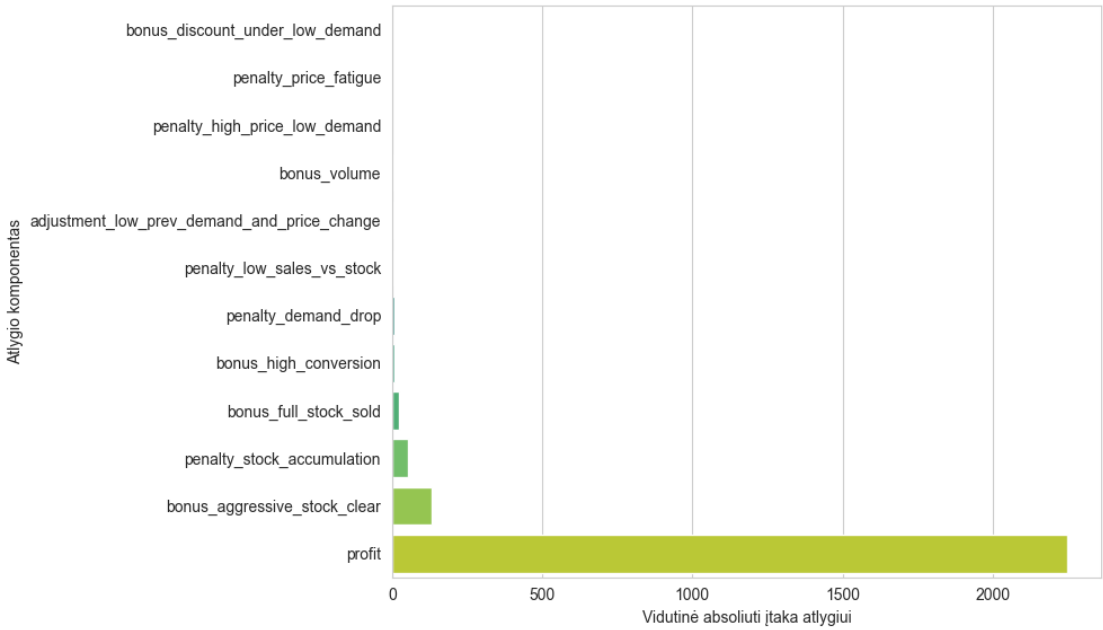
Siekdami įvertinti, ar skatinamojo mokymosi pagrindu veikiantis RL agentas laikosi verslo logikos bei elgiasi racionaliai, buvo atlikta vadinamoji trajektorijų auditavimo (angl. *trajectory audit*) analizė.

Ji remiasi tuo, kad kiekvieno sprendimo metu užfiksuojami atskiros atlygio funkcijos komponentės, leidžiantys detaliai išskaidyti, kas lėmė galutinį atlygį. Tai ne tik padidina sistemos aiškinamumą (angl. *explainability*), bet ir leidžia vertinti jos etiškumą bei atitiktį verslo taisyklėms.



30 pav. Atlygio komponenčių koreliacija su galutiniu atlygiu.

Kaip matyti 29 pav., buvo atlikta atlygio komponenčių koreliacijos analizė pagal jų vidutinę absoliučią įtaką galutiniam atlygį (angl. *reward attribution*). Pagrindiniai teigiami veiksniai buvo pelnas, premijos už agresyvų sandėlio išvalymą ir pardavimų apimtį – tai rodo, kad agentas mokėsi siekti ilgalaikio efektyvumo, o ne tik trumpalaikės naudos. Tuo tarpu neigiamą įtaką turėjo tokie aspektai kaip kainų kėlimas po žemos paklausos, atsargų kaupimas bei mažas pardavimo efektyvumas – tai atskleidžia, kad sistema reaguoja į nepalankius scenarijus ir jų vengia.



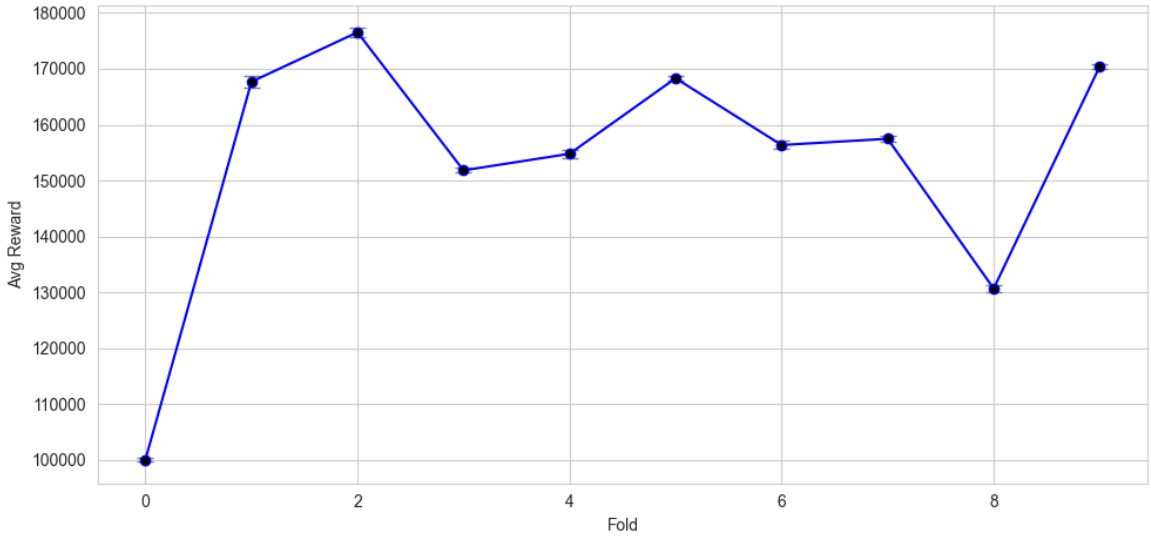
31 pav. SHAP analizė

SHAP analizė papildo šiuos rezultatus individualiu lygiu – vizualizacijose matoma kaip kiekvienas kintamasis (pvz., paklausos pokytis, sandėlio lygis ar ankstesnis kainos sprendimas) konkrečiai paveikia veiksmų parinkimą. Nors tam tikri rezultatai iš SHAP analizės ir koreliacijos galėjo skirtis, tai natūralu – koreliacijos analizė matuoja bendrą statistinį ryšį tarp kintamųjų ir rezultatų, o SHAP vertės vertina modelio vidinį sprendimų pagrindimą kiekviename konkrečiame taške. Tad kai kurie požymiai, turintys mažą koreliaciją, vis tiek gali turėti stiprią reikšmę sprendimui tam tikrose situacijose, ir atvirkščiai.

Tokios metodų sinergija – tiek agreguotos koreliacijos, tiek lokalios SHAP reikšmės – leidžia ne tik įžvelgti, kokiais principais agentas vadovavosi priimdamas sprendimus, bet ir įgyvendinti vienus svarbiausių patikimo DI principų: paaiškinamumą ir atskaitomybę. Šie bruožai būtini norint sukurti sprendimus, kuriais galėtų pasitikėti tiek verslo analitikai, tiek vartotojai, ir kurie atitiktų ne tik našumo, bet ir etiško sprendimų priėmimo reikalavimus.

3.8. Modelio stabilumo įvertinimas naudojant kryžminę validaciją

Siekdami įvertinti modelio apibendrinimo gebėjimus ir išvengti atsitiktinio priderinimo (angl. *overfitting*) prie konkretaus laikotarpio ar duomenų pasiskirstymo, buvo taikyta 10 kartų kryžminė validacija (angl. *10-fold cross-validation*), paremta laiko eilučių segmentavimu. Visas laikotarpis buvo padalytas į dešimt savaičių grupių, iš kurių kiekviena kartą buvo naudota kaip testavimo aibė, o likusios – mokymuisi. Kaip parodyta 30 paveiksle, dauguma duomenų dalių (rinkinių) pasiekė gana aukštą ir stabilų vidutinį atlygio rodiklį, kuris svyravo nuo maždaug 150 000 iki 180 000. Nors pirmoji ir devintoji duomenų dalys parodė žemesnius rezultatus, tai galima paaiškinti netolygiai pasiskirsčiusiais sezoniškumo ar sandėlio kiekio duomenimis. Visgi bendras rezultatas – vidutinis testavimo duomenų atlygis per visas dalis siekė apie 160 000 ± ~23 000, o tai rodo, kad modelis geba išlaikyti pastovius rezultatus skirtinguose kontekstuose, nesikoncentruodamas tik į vieną konkretų laikotarpį ar situaciją.

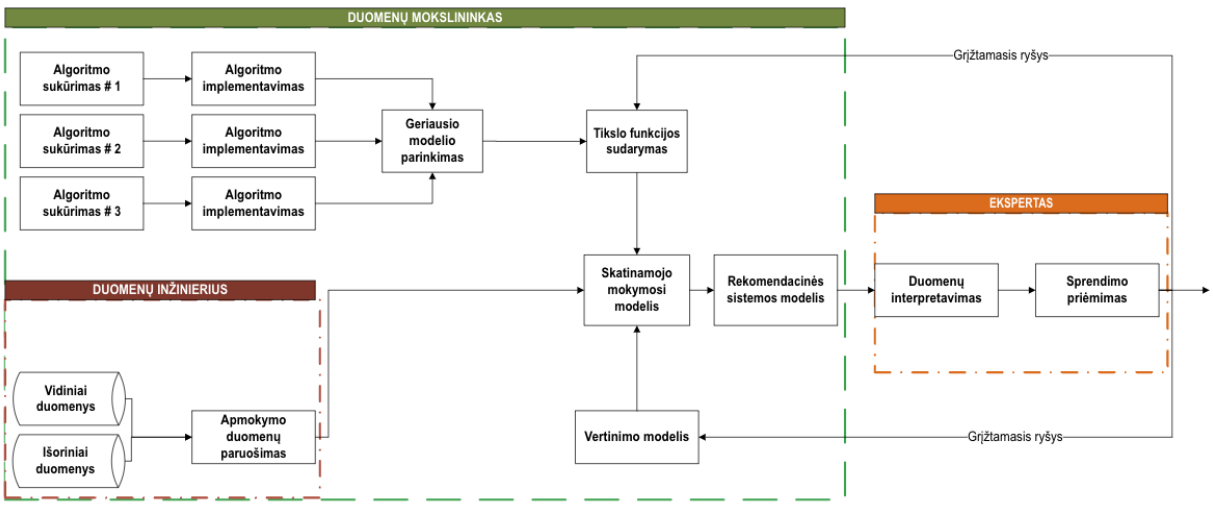


32 pav. Modelio stabilumo įvertinimas.

Toks testavimas leidžia patvirtinti, kad suformuota kainodaros strategija nėra atsitiktinė ar susijusi tik su viena duomenų dalimi – tai svarbus aspektas siekiant pasitikėjimo vertai rekomendacinei sistemai. Kryžminė validacija sustiprina pasitikėjimą agento gebėjimu prisitaikyti prie naujų situacijų, o ne tik „iškalti“ sprendimus iš istorinių duomenų. Tai taip pat leidžia pagrįstai teigti, kad modelis turi potencialą būti taikomas platesniu mastu ar kituose laikotarpiuose, išlaikydamas veikimo kokybę.

3.9. Pasitikėjimo verto rekomendacinės sistemos koncepcija

Apibendrinant tyrimo eigą ir rezultatus, buvo suformuota skatinamuoju mokymusi paremta rekomendacinės sistemos koncepcija, orientuota į pasitikėjimu grįstą sprendimų priėmimą realiose verslo situacijose. Pagrindinis šios koncepcijos principas – ne visiškai automatizuoti sprendimų priėmimo procesą, bet įgalinti dirbtinį intelektą veikti kaip patikimą pagalbininką žmogui, išlaikant sprendimų skaidrumą, atskaitomybę ir kontekstinį jautrumą.



33 pav. Pasitikėjimo vertos rekomendacinės sistemos koncepcija.

Modelio struktūra pagrįsta trimis glaudžiai sąveikaujančiais elementais: duomenų inžinerija, modelių kūrimu ir žmogaus eksperto vaidmeniu. Duomenų inžinierius užtikrina tinkamą duomenų paruošimą – nuo vidinių ir išorinių šaltinių iki apdorojimo mokymui tinkama forma. Duomenų mokslininkas kuria skirtingus modelius, remdamasis verslo logika suformuoja tikslinę funkciją ir parenka efektyviausią modelį, naudodamas stiprinamojo mokymosi algoritmus. Sukurtas rekomendacinis modelis generuoja dinaminės kainodaros sprendimus, kurie testuojami ir vertinami pagal apibrėžtus kriterijus.

Svarbų vaidmenį atlieka žmogus ekspertas – jis vertina modelio siūlomus sprendimus, o sudėtingais ar neapibrėžtais atvejais gali priimti galutinį sprendimą. Tokia sąveika padeda išvengti „juodosios dėžės“ efekto, užtikrina sprendimų kontrolę ir leidžia aiškiai identifikuoti atsakomybę.

Koncepcija įgyvendina svarbiausius pasitikėjimo verto DI principus: lankstumą, paaiškinamumą, atskaitomybę ir nuolatinį tobulėjimą per grįžtamąjį ryšį. Stiprinamojo mokymosi pagrindu veikiantis modelis geba prisitaikyti prie naujų duomenų, atsižvelgti į realaus pasaulio pokyčius ir išlaikyti veikimo efektyvumą. Dėl savo struktūrinio aiškumo bei adaptacijos galimybių modelis gali būti lengvai pritaikomas ir kitose elektroninės komercijos srityse, tokiose kaip personalizuoti pasiūlymai, atsargų valdymas ar nuolaidų kampanijų optimizavimas.

Galutinis rezultatas – subalansuotas, etiškai atsakingas ir praktiškai pritaikomas sprendimų modelis, stiprinantis pasitikėjimą dirbtinio intelekto sprendimais verslo aplinkoje.

Išvados

1. Atliktas tyrimas parodė, kad skatinamojo mokymosi (RL) modeliai gali būti efektyviai pritaikomi dinaminei kainodarai ir rekomendacinėms sistemoms, siekiant tiek verslo rezultatų gerinimo, tiek sprendimų etiškumo ir patikimumo. Tyrimo metu buvo suformuota pasitikėjimo vertos rekomendacinės sistemos koncepcija, įgyvendintas giluminio RL agento kūrimas (naudojant DQN), atliktas hiperparametrų optimizavimas, suformuota tikslo funkcija pagal verslo logiką, įvykdyta sprendimų paaiškinamumo analizė bei įvertintas modelio efektyvumas, lyginant su bazine strategija. Visa tai leido empiriškai pagrįsti, kad tokio tipo DI sprendimai gali būti integruojami į realias verslo sistemas, jei jie kuriami laikantis pasitikėjimo verto DI principų.
2. Atsižvelgiant į tai, kad DI pagrįsta rekomendacijų sistema elgėsi etiškai, t. y. 0 % kainų kėlimo, esant žemai paklausai, o >10 % kainų pokyčiai taikyti tik pagal elastingumą ir kategoriją, išvengiant diskriminacijos bei perteklinio maržų didinimo. Galima teigti, kad RL pagrindu sukurti modeliai gali užtikrinti nešališką sprendimų priėmimą.
3. Trajektorijų auditavimo ir SHAP analizė užtikrino sprendimų paaiškinamumą; didžiausią teigiamą įtaką turėjo pelnas, premija už aktyvų prekių sandėlyje sumažinimą ir premija už pardavimų apimtį, o neigiamą – kainos kėlimas esant žemai paklausai.
4. Su skatinamojo mokymosi (RL) agentu, pilnai deleguotu dinaminei kainodarai, bendrovė vieno epizodo metu uždirbo vidutiniškai ≈ 174 tūkst. € pelno – tai ≈ 37 tūkst. € (+ 27 %) daugiau, negu taikant dabartinę taisyklių pagrindu veikiančią strategiją. Šis skirtumas yra statistiškai patikimas, nes su taikytu t-testu ($t = 2,80$; $p = 0,004$) buvo atmesta nulinė hipotezė, o Cohen $d \approx 0,5$ rodo *vidutinio stiprumo praktinį efektą* – t. y. vidutiniam pardavimo laikotarpiui modelis sukurtų ≈ 37 tūkst. € papildomo grynojo pelno. Taip pat, agentas kainos kryptį (kelti / mažinti) pakeitė 27 kartus iš 52 sprendimų žingsnių, t. y. ~ 52 % atvejų, todėl kainodara buvo pakankamai aktyvi. RL modeliu pagrįsti sprendimai buvo ne tik efektyvūs, bet ir nešališki, paaiškinami bei skaidrūs, todėl atitiko visus pasitikėjimo vertos rekomendacinės sistemos kriterijus, suformuluotus tyrimo pradžioje.
5. Tyrimas parodė, kad RL pagrindu sukurtos rekomendacinės kainodaros sistemos gali būti laikomos pasitikėjimo vertomis, jeigu visame jų kūrimo ir veikimo cikle nuosekliai taikomi skaidrumo, paaiškinamumo, šališkumo mažinimo, etiškumo, duomenų kokybės, stabilumo bei atskaitomybės principai. Tokiu būdu, įmonė rekomendacinę RL sistemą gali įdiegti į esamus verslo procesus – nuo kainų nustatymo modulio iki rezultatų stebėsenos suvestinių – ir taip perkelti pasitikėjimo vertą DI iš prototipo stadijos į praktinį verslo modelį.

Literatūros sąrašas

1. RUSSELL, Stuart, NORVIG, Peter**.** *Artificial Intelligence: A Modern Approach.* 3rd ed. Global Edition. London: Pearson Education, 2016. ISBN 9781292153964.
2. HAENLEIN, Michael; KAPLAN, Andreas**.** A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, 2019, t. 61, nr. 4, p. 5–14. Prieiga per internetą: https://doi.org/10.1177/0008125619864925
3. DUAN, Yanqing; EDWARDS, John S.; DWIVEDI, Yogesh K. Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management*, 2019, t. 48, p. 63–71. Prieiga per internetą: https://doi.org/10.1016/j.ijinfomgt.2019.01.021
4. RANSBOTHAM, Sam; KIRON, David; GERBERT, Philipp; REEVES, Martin. *Reshaping Business With Artificial Intelligence*. MIT Sloan Management Review, 2017.
5. WACH, M.; OOI, K. B.; IRFAN, M.; SULISTYO, S. R. *Examining Risk and Tech-Savviness on Student's Adoption of Generative AI in Higher Education Assessments*. Open Access City Research Online, 2024
6. LEPRI, Bruno; OLIVER, Nuria; LETOUZE, Emmanuel; PENTLAND, Alex; VINCK, Patrick. *Fair, Transparent, and Accountable Algorithmic Decision-Making Processes*. Philosophy & Technology, 2018, t. 31, nr. 4, p. 611–627.
7. MIKALEF, P.; GUPTA, M. *Artificial Intelligence Capability: Conceptualization, Measurement Calibration, and Empirical Evidence*. European Journal of Information Systems, 2021, t. 30, nr. 2, p. 171–190.
8. SHANKAR, V. *How AI Is Reshaping Marketing*. Journal of the Academy of Marketing Science, 2020, t. 48, p. 7–20.
9. DWIVEDI, Y. K.; KSHETRI, N.; HUGHES, L.; SLADE, E. L.; JEYARAJ, A.; KAR, A. K. ir kt. *So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy*. International Journal of Information Management, 2023, t. 71, 102642. Prieiga per internetą: https://doi.org/10.1016/j.ijinfomgt.2023.102642
10. BIN-NASHWAN, S. A.; SADDALH, M.; BOUTEARA, M. *Use of ChatGPT in academia. Technology in Society 75, Academic integrity hangs in the balance*. 2023, 102370. Prieiga per internetą: https://doi.org/10.1016/j.techsoc.2023.102370
11. HUANG, L.; LADIKAS, M.; SCHIPPIL, J.; HE, G.; HAHN, J. *Knowledge mapping of an artificial intelligence application scenario: a bibliometric analysis of the basic research of data-driven autonomous vehicles*. Technology in Society, 2023, t. 75, 102394. Prieiga per internetą: https://doi.org/10.1016/j.techsoc.2023.102394
12. WAAL AL-KHATIB, A. *Drivers of generative artificial intelligence to fostering exploitative and exploratory innovation: a TOE framework*. Technology in Society, 2023, t. 75, 102403. Prieiga per internetą: https://doi.org/10.1016/j.techsoc.2023.102403
13. WEBER, E. U.; JOHNSON, E. J. *Mindful judgment and decision making*. Annual Review of Psychology, 2009, t. 60, p. 53–85.
14. BIN-NASHWAN, S. A.; SADDALH, M.; BOUTEARA, M. *Use of ChatGPT in academia. Academic integrity hangs in the balance*. Technology in Society, 2023, t. 75, 102370. Prieiga per internetą: https://doi.org/10.1016/j.techsoc.2023.102370

15. SIMCHI-LEVI, D.; WANG, H.; WEI, Y. *Increasing supply chain robustness through process flexibility and inventory*. Production and Operations Management, 2018, t. 27, nr. 8, p. 1476–1495. Prieiga per internetą: https://doi.org/10.1111/poms.12887
16. ARIAS-PÉREZ, J.; HUYNH, T. *Flipping the odds of AI-driven open innovation: The effectiveness of partner trustworthiness in counteracting interorganizational knowledge hiding*. Industrial Marketing Management, 2023, t. 111, p. 30–40. Prieiga per internetą: https://doi.org/10.1016/j.indmarman.2022.12.006
17. BAHOO, S.; CUCCULELLI, M.; QAMAR, D. *Artificial intelligence and corporate innovation: A review and research agenda*. Technological Forecasting and Social Change, 2023, t. 189, 122264. Prieiga per internetą: https://doi.org/10.1016/j.techfore.2022.122264
18. MARIANI, M.; MACHADO, I.; NAMBISAN, S. *Types of innovation and artificial intelligence: A systematic quantitative literature review and research agenda*. Journal of Business Research, 2023, t. 155, 113364. Prieiga per internetą: https://doi.org/10.1016/j.jbusres.2022.113364
19. BADAKHSHAN, E.; HUMPHREYS, P.; MAGUIRE, L. P.; McIVOR, R. *Using simulation-based system dynamics and genetic algorithms to reduce the cash flow bullwhip in the supply chain*. International Journal of Production Research, 2020, t. 58, nr. 11, p. 1–27. Prieiga per internetą: https://doi.org/10.1080/00207543.2020.1715505
20. GREWAL, D.; GAURI, D. K.; DAS, G.; AGARWAL, J.; SPENCE, M. T. *Retailing and emergent technologies*. Journal of Business Research, 2021, t. 134, p. 198–202. Prieiga per internetą: https://doi.org/10.1016/j.jbusres.2021.05.024
21. EDWARDS, J.; MILES, M. P.; D'ALESSANDRO, S.; FROST, M. *Entrepreneurial strategy-making, corporate entrepreneurship preparedness and entrepreneurial sales actions: Improving B2B sales performance*. Journal of Business Research, 2023, t. 157, straipsnis 113586. Prieiga per internetą: https://doi.org/10.1016/j.jbusres.2022.113586
22. MAGAS, M.; KIRITSIS, D. *Industry Commons: an ecosystem approach to horizontal enablers for sustainable cross-domain industrial innovation (a positioning paper)*. International Journal of Production Research, 2021, t. 60, nr. 7, p. 1–14. Prieiga per internetą: https://doi.org/10.1080/00207543.2021.1989514
23. COLUMBUS, L. *10 Ways AI Is Improving Manufacturing In 2020*. Forbes, 2020-05-18. Prieiga per internetą: https://www.forbes.com/sites/louiscolumbus/2020/05/18/10-ways-ai-is-improving-manufacturing-in-2020/
24. SAHOO, S.; KUMAR, A.; DONTHU, N.; SINGH, S. *Artificial intelligence capabilities, open innovation, and business performance: Empirical insights from multinational B2B companies*. Industrial Marketing Management, 2024, t. 118, p. 1–14.
25. SIMCHI-LEVI, D.; WU, D. *Powering retailers' digitization through analytics and automation*. ResearchGate, 2018.
26. WEI, R.; PARDO, C. *Artificial intelligence and SMEs: How can B2B SMEs leverage AI platforms to integrate AI technologies?* Industrial Marketing Management, 2023, t. 107, p. 102–115. Prieiga per internetą: https://doi.org/10.1016/j.indmarman.2022.10.008
27. YIN, J.; WEI, S.; CHEN, X.; WEI, J. *Does it pay to align a firm's competitive strategy with its industry IT strategic role?* Information Management, 2020, t. 57, nr. 8, 103391. Prieiga per internetą: https://doi.org/10.1016/j.im.2020.103391

28. BRYNJOLFSSON, E.; MCAFEE, A. *The Business of Artificial Intelligence*. Harvard Business Review, 2017, t. 7, p. 3–11. Prieiga per internetą: https://starlab-alliance.com/wp-content/uploads/2017/09/The-Business-of-Artificial-Intelligence.pdf
29. MINTZBERG, H.; RAISINGHANI, D.; THEORET, A. *The structure of 'unstructured' decision processes*. Administrative Science Quarterly, 1976, t. 21, nr. 2, p. 246–275.
30. BAZERMAN, M. H.; MOORE, D. A. *Judgment in Managerial Decision Making*. Hoboken: Wiley, 2008.
31. EPISTOLA, C.; JACOB, V.; ROACH, D. *The Concept of Information Overload: A Review of Literature From Organization Science, Accounting, Marketing, MIS, and Related Disciplines*. The Information Society, 2004, t. 20, nr. 5, p. 325–344. Prieiga per internetą: https://doi.org/10.1080/01972240490507974
32. MARCH, J. G.; SHAPIRA, Z. *Managerial Perspectives on Risk and Risk Taking*. Management Science, 1987, t. 33, p. 1404–1418. Prieiga per internetą: https://doi.org/10.1287/mnsc.33.11.1404
33. DANE, E.; ROCKMANN, K. W.; PRATT, M. G. *When should I trust my gut? Linking domain expertise to intuitive decision-making effectiveness*. Organizational Behavior and Human Decision Processes, 2012, t. 119, nr. 2, p. 187–194. Prieiga per internetą: https://doi.org/10.1016/j.obhdp.2012.07.009
34. MINTZBERG, H.; RAISINGHANI, D.; THEORET, A. *The structure of 'unstructured' decision processes*. Administrative Science Quarterly, 1976, t. 21, nr. 2, p. 246–275.
35. PALIUKAS, V.; SAVANEVIČIENĖ, A. *Harmonization of rational and creative decisions in quality management using AI technologies*. Economics and Business, 2018, t. 32, p. 195–208. Prieiga per internetą: https://doi.org/10.2478/eb-2018-0016
36. GOODMAN, B.; FLAXMAN, S. *European Union regulations on algorithmic decision-making and a 'right to explanation'*. AI Magazine, 2017, t. 38, nr. 3, p. 50–57.
37. WANG, R. Y.; STRONG, D. M. *Beyond accuracy: What data quality means to data consumers*. Journal of Management Information Systems, 1996, t. 12, nr. 4, p. 5–33.
38. PIPINO, L. L.; LEE, Y. W.; WANG, R. Y. *Data Quality Assessment*. Communications of the ACM, 2002, t. 45, nr. 4, p. 211–218.
39. CHEN, A. N. K. *Explainable AI for Enhanced Decision Making*. Decision Support Systems, 2024. Prieiga per internetą: https://www.sciencedirect.com/science/article/pii/S0167923624000010
40. BORCHERT, P.; COUSSEMENT, K.; DE CAIGNY, A.; DE WEERDT, J. *Extending business failure prediction models with textual website content using deep learning*. European Journal of Operational Research, 2023, t. 306, nr. 1, p. 348–357. Prieiga per internetą: https://doi.org/10.1016/j.ejor.2022.06.060
41. RAI, A. *Explainable AI: from black box to glass box*. Journal of the Academy of Marketing Science, 2020, t. 48, nr. 1, p. 137–141. Prieiga per internetą: https://doi.org/10.1007/s11747-019-00710-5
42. VEALE, M.; VAN KLEEK, M.; BINNS, R. *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making*. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems – CHI '18*, Montreal, Canada. ACM Press, 2018, p. 1–14. Prieiga per internetą: https://doi.org/10.1145/3173574.3174014

43. RAJI, I. D.; SMART, A.; WHITE, R. N.; MITCHELL, M.; GEBRU, T.; HUTCHINSON, B.; SMITH-LOUD, J.; THERON, D.; BARNES, P. *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT)*, 2020, p. 33–44.
44. RAHWAN, I. *Society-in-the-loop: programming the algorithmic social contract*. Ethics and Information Technology, 2018, t. 20, nr. 1, p. 5–14.
45. FLORIDI, L.; COWLS, J.; BELTRAMETTI, M.; CHATILA, R.; CHAZERAND, P.; DIGNUM, V.; … SCHAFER, B. *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*. Minds and Machines, 2018, t. 28, nr. 4, p. 689–707.
46. CHEN, Y.; ZHAO, X.; YU, S.; ZHANG, X. *Reinforcement Learning for Personalized Recommender Systems: Challenges and Opportunities*. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, 2020, p. 4630–4637.
47. FENG, Y.; LI, X.; ZHANG, Y.; WANG, H. *Deep reinforcement learning for business decision-making: A comprehensive review*. Journal of Business Research, 2019, t. 101, p. 1–15.
48. SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction*. 2-asis leidimas. Cambridge: MIT Press, 2018.
49. SILVER, D.; HUANG, A.; MADDISON, C. J.; GUEZ, A.; SIFRE, L.; VAN DEN DRIESSCHE, G.; … HASSABIS, D. *Mastering the game of Go with deep neural networks and tree search*. Nature, 2016, t. 529, nr. 7587, p. 484–489. Prieiga per internetą: https://doi.org/10.1038/nature16961
50. MNIH, V.; KAVUKCUOGLU, K.; SILVER, D.; RUSU, A. A.; VENESS, J.; BELLEMARE, M. G.; … HASSABIS, D. *Human-level control through deep reinforcement learning*. Nature, 2015, t. 518, nr. 7540, p. 529–533. Prieiga per internetą: https://doi.org/10.1038/nature14236
51. WATARI, D.; TANIGUCHI, I.; ONOYE, T. Duck curve aware dynamic pricing and battery scheduling strategy using reinforcement learning. *IEEE Transactions on Smart Grid*, 2023, t. 15, nr. 1, p. 457–471. Prieiga per internetą: https://doi.org/10.1109/TSG.2023.3245678
52. SHEN, Q.; YU, K.; LOU, Q.; ZHANG, Y.; NI, X. A Deep Reinforcement Learning Approach to Enhancing Liquidity in the U.S. Municipal Bond Market: An Intelligent Agent-based Trading System. *International Journal of Innovative Research in Engineering & Management*, 2024, t. 11, nr. 6, p. 43–54. Prieiga per internetą: https://doi.org/10.55524/ijirem.2024.11.6.5
53. BAE, S.; KULCSAR, B.; GROS, S. Personalized Dynamic Pricing Policy for Electric Vehicles: Reinforcement Learning Approach. *Transportation Research Part C: Emerging Technologies*, 2024. Prieiga per internetą: https://doi.org/10.1016/j.trc.2024.103456
54. ZHU, Y.; SHIHAB, M.; WEI, L. Deep Q-learning for Multi-Flight Dynamic Pricing: Maximizing Revenue with a Novel Utility Function in Airline Revenue Management. *Transportation Research Part E: Logistics and Transportation Review*, 2024. Prieiga per internetą: https://doi.org/10.1016/j.tre.2024.102345
55. GUO, X.; LI, M.; ZHAO, L.; WANG, H. A Deep Reinforcement Learning Approach for Ride-Hailing and Ride-Pooling Services. *arXiv preprint*, 2024. Prieiga per internetą: https://arxiv.org/abs/2503.13200
56. CHEN, G.; WANG, X. Reinforcement Learning for Approximating the Solution of Markov Decision Processes. *SSRN Electronic Journal*, 2021. Prieiga per internetą: https://doi.org/10.2139/ssrn.4272376

57. GAO, R.; ZHANG, M.; LIU, K. Applications of deep Q-networks in robotics. *Robotics and Autonomous Systems*, 2022, t. 85, p. 67–79.
58. SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction*. 2-asis leidimas. Cambridge: MIT Press, 2018.
59. PATEL, R.; et al. Policy gradient methods for continuous control. *IEEE Transactions on Cybernetics*, 2021, t. 39, nr. 12, p. 987–1003.
60. KIM, B.; LEE, H. On-policy and off-policy learning: A comparative study. *Neural Computing and Applications*, 2020, t. 31, nr. 9, p. 2234–2251.
61. SHARMA, D.; XU, W. A survey on hybrid reinforcement learning approaches. *Machine Learning Journal*, 2023, t. 29, nr. 6, p. 345–361.
62. NAEEM, M.; RIZVI, S. T. H.; CORONATO, A. A gentle introduction to reinforcement learning and its application in different fields. *IEEE Access*, 2020, t. 8, p. 209320–209344.
63. SUTTON, R. S.; BARTO, A. G. *Reinforcement Learning: An Introduction*. 2-asis leidimas. Cambridge: MIT Press, 2018.
64. SUTTON, R. S.; McALLESTER, D.; SINGH, S.; MANSOUR, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems (NIPS)*, 2000.
65. SCHULMAN, J.; LEVINE, S.; ABBEEL, P.; JORDAN, M.; MORITZ, P. Trust Region Policy Optimization. *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.
66. HAARNOJA, T.; ZHOU, A.; ABBEEL, P.; LEVINE, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep RL with a Stochastic Actor. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.
67. LIU, C.-L.; CHANG, C.-C.; TSENG, C.-J. Actor-Critic Deep Reinforcement Learning for. Solving Job Shop Scheduling Problems. *IEEE Access*, 2020, t. 8, p. 71752–71762.
68. JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2016, t. 374, nr. 2065, 20150202. P
69. SCIKIT-LEARN. Cross-validation: evaluating estimator performance. Prieiga per internetą: https://scikit-learn.org/stable/modules/cross_validation.html
70. SHRESTHA, Y. R.; BEN-MENAHEM, S. M.; von KROGH, G. Organizational Decision-Making Structures in the Age of Artificial Intelligence. *California Management Review*, 2019, t. 61, nr. 4, p. 66–83.