Vilniaus universitetas Filosofijos fakultetas

Filosofijos institutas

Gytis Pranskūnas

Filosofijos studijų programa

Kursinis darbas

**Loginis išvedimas kaip generatyvinių kalbos modelių problema**

Darbo vadovas: prof. Jonas Dagys

Vilnius 2024

Turinys

[Įvadas 3](#_Toc169812097)

[I dalis. Kriterijai 4](#_Toc169812098)

[Loginis išvedimas 4](#_Toc169812099)

[Išvedimo teorijos 4](#_Toc169812100)

[Taisyklių teorija 4](#_Toc169812101)

[Semantinė teorija 5](#_Toc169812102)

[Taisyklių-semantinė teorija 6](#_Toc169812103)

[II dalis. Generatyviniai kalbos modeliai 7](#_Toc169812104)

[Techninis kontekstas ir prielaidos 7](#_Toc169812105)

[Psicho-filosofinis pagrindas 8](#_Toc169812106)

[Protas kaip neuronų tinklas 8](#_Toc169812107)

[Proto teorijos 8](#_Toc169812108)

[Konekcionizmas 8](#_Toc169812109)

[Komputacinė 9](#_Toc169812110)

[Veikimo principai 10](#_Toc169812111)

[Neuroniniai tinklai 10](#_Toc169812112)

[Dėmesio mechanizmas arba Transformerio architektūra 11](#_Toc169812113)

[Transformerių veikimas 12](#_Toc169812114)

[III dalis. Vertinimas 14](#_Toc169812115)

[Loginis išvedimas kalbos modelių kontekste 14](#_Toc169812116)

[Kinų kambarys kaip ataka prieš sintaksinę reikšmės teoriją 16](#_Toc169812117)

[Tiuringo testas 16](#_Toc169812118)

[Kinų Kambario Argumentas 16](#_Toc169812119)

[Išvados 19](#_Toc169812120)

[Literatūra 21](#_Toc169812121)

## Įvadas

Dirbtinio Intelekto kaip mokslo šakos tikslas – sukurti protaujančias mašinas. Todėl nuo antrojo XXa. dešimtmečio pradėję vystytis generatyviniai kalbos modeliai kviečia svarstyti svarbias problemas susijusias su žmogaus protavimo procesais. 1950m. Tiuringas pasiūlė testą, kuriuo galime nustatyti ar kompiuteriui galime priskirti intelektą remiantis tuo, ar jis gali įtikinamai apsimesti žmogumi bendraudamas. Dirbtinio Intelekto mokslo tikslas yra nustatyti algoritmą, kuriuo žmogus protauja, o šio klausimo filosofinį svarstymą sąlygoja sudėtingos problemos suskaidymas į dalis ir vienos iš dalių svarstymas. Darbo objektas yra loginis išvedimas – sąlyginis mąstymas apie tai, kas išplaukia jeigu kažkas įvyksta (Holliday & Mandelkern 2024) – generatyvinių kalbos modelių kontekste. Šio darbo tikslai yra 1) apžvelgti teorijas aiškinančias žmogaus išvedimo procesą, 2) apsibrėžti generatyvinių kalbos modelių veikimo principus siekiant atsakyti į klausimą: koks yra kalbos modelių pagrindas loginiame išvedime – semantinis ar sintaksinis? Kitaip tariant, ar pagrįsta teigti, kad kalbos modeliai manipuliuoja formaliais simboliais, todėl neturi ryšio su semantika? Ginsiu tezę, jog generatyviniai kalbos modelių pagrindas, to pasekoje ir išvedimas, yra semantinio pobūdžio dėl jų architektūros ir veikimo principo. Generatyvinių kalbos modelių struktūros analizė susideda iš a) psicho-filosofinio pagrindo analizės – protavimo lokalizavimo smegenyse ir jo funkcijų kaip neuronų veikimo sumos; b) bazinio veikimo principo veikimo – dirbtinio neuroninio tinklo ir Dėmesio mechanizmo, būdingo daugeliui kalbos modelių, vadinamų transformeriais – aptarimas. Vertinamąją darbo dalį sudarys susijusių klausimų analizė. Pirma, kaip pritaikomos loginio išvedimo teorijos generatyviniams kalbos modeliams pagal jų veikimo principus? Atsakymas į klausimą susideda iš bandymo suprasti modelių veikimo principus teorijų kontekste ir atsakymo į semantikos priskyrimo modeliams priekaištus. Kitas klausimas susijęs su Tiuringo testu ir Kinų Kambario argumentu. Loginis išvedimas suponuoja suvokimą, o suvokimas – išvedimą. Vadinasi, turime pažiūrėti ar Kinų Kambario argumentas pritaikomas aptariamiems modeliams. Vertinamoji dalis generatyviniams kalbos modeliams padės suformuoti filosofinį pagrindą, o to pagrindu bus galima atlikti tolesnius tyrimus analizuojant generatyvinių kalbos modelių veikimo principus filosofiniame ir kituose kontekstuose.

## I dalis. Kriterijai

Loginis išvedimas

Logika yra vienas iš žmogaus pažinimo būdu protu, kuria tiriamas bet kokio turinio minčių jungimo būdas ir forma (Rimkutė 2007, p. 105). Kitaip tariant, logika tiria samprotavimo būdą ir išvedimo taisyklingumą. Samprotavimą sudaro prielaidos ir iš jų turinti sekti išvada – taip apibrėžiamas pagrįstas samprotavimas, o dažniausiai jį ir turime omenyje. Holliday ir Mandelkerno (2024) teigimu, sąlyginis mąstymas – mąstymas apie tai, kas išplaukia, jeigu kažkas įvyksta – yra „vienas iš būdingiausių žmogaus pažintinių gebėjimų“. Logika tiria minčių jungimobūdą, todėl ji neturi kokybinio santykio su faktais. Logikai faktai nerūpi. Turinio požiūriu samprotavimas gali būti klaidingas, tačiau pagrįstas (Rimkutė 2007, p. 105).

Samprotaudamas žmogus atsako į klausimus, planuoja ir priima sprendimus. Luo ir kiti (2023) pateikia loginio samprotavimo pavyzdį, kuris gerai iliustruoja gebos svarbą. Juodųjų Skylių (*angl.* Black Holes) egzistavimas pirmiausiai buvo išvestas iš kvantinės mechanikos ir reliatyvizmo principų. Iki pirmųjų empirinių Juodųjų Skylių stebėjimo įrodymų vien matematiškai buvo išvestas šių objektų egzistavimas – po 50 metų šie skaičiavimai buvo patvirtinti empiriškai[[1]](#footnote-1). Tikrovės pažinimas ir tiesos pasiekimas priklauso nuo taisyklingo samprotavimo. Vadinasi, išvedimas suponuoja suvokimą apskritai.

Išvedimo teorijos

# Taisyklių teorija

Žmogaus išvedimo procesą tyrėjai aiškina skirtingai. Rimkutė (2007, p. 116) išskiria tris. Viena iš jų yra *taisyklių* *teoriją*, pagal kurią samprotaudamas žmogus naudojasi formaliomis taisyklėmis, nuosekliai taikydamas tam tikra tvarka ir išvesdamas išvadas. Galime įsivaizduoti, kad turėdamas prielaidas „jei kažkas yra žmogus, tai jis yra mirtingas“ ir „Jonas yra žmogus“, asmuo naudodamasis Modus Ponnens (MP) taisykle – antecedentui esant teisingam, konsekventas taipogi teisingas – prieis teisingą išvadą: „Jonas yra mirtingas“. Kitaip tariant, šia teorija mąstymo dėsniai yra tiesiog logikos dėsnių naudojimas. Kita vertus, tyrimai nurodo, kad lengviau daryti teisingą išvedimą, kai samprotavimas susijęs su realiai gyvenimą liečiančias faktais žmonės geriau protauja aplinkybėse, kurios liečia gyvenimą arba kasdienybę (žr. Lurija 1982; Cole & Schribner 1977; Schribner 1977). Holliday ir Mandelkern (2024) teigimu, MP prastai atitinka „jei“ reikšmę natūralioje kalboje. Teiginių logikoje sąlyginis teiginys „jei *p*, tai *q*“ teisingas, kai *p* yra klaidingas arba *q* teisingas. Teiginys „studentai išlaikys egzaminus, jei stropiai mokysis“ implikuoja, kad „studentai stropiai mokysis“, tačiau tai nėra būtina. Pasak autorių, teiginio „jei *p*, tai *q*“ tikimybinė teisingumo reikšmė didėtų mažėjant *p* teisingumo reikšmei. Taip pat norėčiau paminėti materialiosios implikacijos problemą – teiginys teisingas visada, kai *p* yra klaidingas. Kitaip tariant, iš klaidingo teiginio galime išvesti bet ką – *ex falso quodlibet*.

Kadangi žmogus yra protaujanti būtybė, tai natūralu bandyti mąstymą apibrėžti kaip logikos dėsnių naudojimą, tačiau iš empirinių duomenų matome, kad žmonės geriau samprotauja, kai teiginiai yra susiję su praktine, kasdiene informacija. Kita vertus, logikos taisyklių – kaip dėsnių statusas – yra nevienareikšmiškas. Antecedento forma sąlyginiame teiginyje prastai atitinka natūralią kalbą, kuria žmogus ir reiškia savo mąstymą bei implikacijos teisingumas esant klaidingam antecedentui implikuoja svetimą santykį su žmogaus mąstymu. Kitaip tariant, vien taisyklėmis negalime apibūdinti žmogaus mąstymo proceso. Būtina semantika.

# Semantinė teorija

Kita išvedimo teorija - *semantinė teorija -* išvedimą konceptualizuoja kaip semantinį procesą. Išvedimas remiasi ne teiginių forma ir formaliomis taisyklėmis, o teiginių reikšme (Johnson-Laird 1999, p. 116). Mentaliniai modeliai tai „į schemą panaši, tik sudėtingesnė žinių struktūra“ reprezentuojanti informaciją ir jos apdorojimo būdą (Rimkutė 2007, p. 99). Pasak Johnson-Lairdo (1999, ibid.), kiekvieno mentaliniomodeliostruktūra ir turinys nurodo įvairius būdus, kuriais gali pasireikšti galimybė – mentalinismodelis*.* Kitaip tariant, esant teisingoms prielaidoms, samprotavimo išvados teisingumo tikimybė didėja proporcionaliai su randamais atitinkamais mentaliniaismodeliais. Išvada yra būtina, jeigu ji „galioja“ (*angl.* holds) visuose modeliuose. Pavyzdžiui, teiginio „yra trikampis ir apskritimas“ supratimas aiškinamas teiginio reikšmės perteikimu, iš kurios išplaukia su ja susijęs mentalinis situacijos modelis (*ibid*.):

O ∆.

Mentalinio modelio samprata miglota. Kiek mentalinių modelių reikia, kad būčiau užtikrintas išvedimo rezultatu? O jeigu turiu tik kelis mentalinius modelius, ar mano išvados bus pagrįstos, nors ir nerasiu nesutampančio mentalinio modelio? Kaip schema vystosi? Kaip susidarome modelius?

# Taisyklių-semantinė teorija

*Taisyklių* ir *semantinė* teorijos atstovavo kaip vienas kitą išskiriantys požiūriai į išvedimo procesą. Pirmoji mąstymą aiškina formalių taisyklių naudojimu, o pastaroji – semantiniais modeliais. Kaip du tarpusavyje susijusius pažinimo proceso polius susiejo Falmagne (1995). Semantinių modelių konstravimas priklauso nuo sintaksinių ir loginių teiginio savybių, kadangi teiginio reikšmė įtraukia ir jo loginę formą (p. 217). Dėl tos priežasties, jog logika tiria minčių – teiginių ir konceptų – jungimo būdą, nuo to priklauso ir sąmonėje formuojamos išvados teisingumo galimybės, t.y. mentaliniai modeliai. Kitaip tariant, vaizduotės būdu suformuotas teiginio modelis priklauso nuo teiginio loginės formos, sintaksės, nes adekvatus mentalinis modelis – jo struktūra ir reikšmė – priklauso nuo tinkamo teiginiais reiškiamų tikrovės vaizdo dalių santykio suvokimo.

Nėra vienareikšmiškos ir visuotinės išvedimo procesą paaiškinančios teorijos. Sintaksinis laiko logikos dėsnių naudojimą mąstymo pagrindu, tačiau tiek formalios taisyklės, tiek empiriniai duomenys kelia rimtas problemas, neleidžiančias pagrįstai aiškinti loginio išvedimo išskirtinai sintaksine prasme. Kita vertus, į intuiciją apeliuojanti psichinių modelių teorija atmeta sintaksinį, vien formos požiūriu traktuojamą mąstymą, tačiau modelių koncepcija yra miglota. Galiausiai, šie požiūriai yra suvienijami – adekvataus mentalinių modelių formavimas priklauso nuo teiginio suvokimo ir žodžių teisingo susiejimo – sintaksė santykiauja su semantika pateikdama mentalinių modelių formavimo principą.

## II dalis. Generatyviniai kalbos modeliai

Techninis kontekstas ir prielaidos

„Generatyvinis“ ir „kalbos“ reiškia, jog modelis generuoja tekstą. Diskusija apie Generatyvinius Kalbos Modelius (GKM) arba kitaip Didelius Kalbos Modelius (*angl.* large language models) turi pradėti nuo mokslo šakos aptarimo. Informatika (*angl*. computer science) turi daug skirtingų šakų, tačiau mums svarbiausia Taikomoji informatika (*angl.* applied computer science), kurioje lokalizuotas mūsų dėmesio objektas – Dirbtinis Intelektas (*angl.* artificial intelligence, DI). Pasak Bringsjord ir Govindarajulu (2024), DI yra mokslo sritis, kurios tikslas – kurti „dirbtinius gyvūnus (arba bent jau dirbtines būtybes, kurios tinkamuose kontekstuose *atrodo* kaip gyvūnai)“ arba „dirbtinius žmones (arba bent jau dirbtines būtybes, kurios tinkamuose kontekstuose *atrodo* kaip žmonės)“. Tikslas skyla į kelis DI apibrėžimus. Pirma, tai naujos pastangos priversti kompiuterius mąstyti (Haugeland 1985). Taip pat, ši mokslo sritis susijusi su protingo elgesio automatizavimu (Luger & Stubblefield 1993) ir kurianti racionaliai mąstančias sistemas (Winston 1992). Galiausiai, tai sistemos, kurios gali įtikinamai apsimesti žmogumi – išlaikyti Tiuringo testą (Turing 1950). Pastarąją poziciją aptarsiu trečioje darbo dalyje. Taigi, DI tikslas – imituoti žmogaus mąstymą, kuriant „skaičiavimo mašinas, kurios vykdytų pagrindines protines užduotis, tokias kaip samprotavimas, sprendimų priėmimas, problemų sprendimas ir pan.“ (Rescorla 2020).

DI srityje GKM yra lokalizuoti Mašininio Mokymosi (*angl*. Machine Learning, ML) srityje, kuri naudojasi algoritmais ir duomenimis siekiant „imituoti“ žmonių mokymosi būdą, palaipsniui didinant tikslumą (IBM 2021). Iš esmės, ML tikslai yra nuspėti (*angl.* predict) ir klasifikuoti duomenis: „remdamasis tam tikrais įvesties duomenimis, kurie gali būti pažymėti [*angl.* labeled] arba nepažymėti, [mašininio mokymosi] algoritmas parengia įvertį apie duomenų modelį“ (ibid.). Tarkime, kad turime duomenų rinkinį susijusį su širdies ligų simptomais (krūtinės skausmas, pasunkėjęs kvėpavimas, nuovargis ir t.t.) ir juos turinčiais žmonėmis bei atsakymais, ar jiems buvo diagnozuota liga. Remiantis duomenimis tinklas išgauna funkciją, į kurią įstačius reikalaujamų kintamųjų reikšmes bus galima nuspėti kurie žmonės turi širdies ligą. Mums svarbiausias algoritmas, skirtas išgauti funkciją – Dirbtinis Neuroninis Tinklas (*angl.* Artificial Neural Network, DNN), kuriuo remiasi GKM. Dar vienas svarbus komponentas sudarantis dabartinę kalbos modelių sėkmę – Dėmesio mechanizmas (*angl.* attention mechanism) (Turner 2024).

Psicho-filosofinis pagrindas

# Protas kaip neuronų tinklas

Protas yra lokalizuotas smegenyse ir mentaliniai procesai yra smegenų veiklos rezultatas - tai pagrindinė psicho-filosofinė DI prielaida. Smegenų procesų kaip neuronų veiklos rezultato koncepcija išaugo iš William Jameso (1891) ir Alexander Baino (1873). „Pagrindinė jų darbo idėja yra ta, kad tiek mintys, tiek kūno veikla yra neuronų procesų smegenyse rezultatas“ o „kiekvienai galimai veiklai reikia, kad įsijungtų tam tikras neuronų rinkinys“ (Bishop 2015, p. 27). Jameso modelyje prisiminimai ir veiksmai aiškinami tarp neuronų tekančia elektros srove. Vienu iš „asociacinių principų“ – kaip vieni smegenų procesai inspiruoja kitą, pvz. obuolys primena medį – teigiama, kad jeigu du smegenų procesai veikia vienas po kito arba kartu, tai kitą kartą, vienam iš jų pasikartojus, bus sužadintas kitas veikęs sąjungoje. Kitaip tariant, mąstymas ir žmogaus veiksmai aiškinami kaip smegenų procesų rezultatas, o šis yra mentalinių modelių turinio (prisiminimo, teiginio) pasiskirstymas per skaičių neuronų, dėl kurių įsižiebimo (*angl.* firing) ir tarpusavio jungčių turinys iškyla į sąmonės paviršių. Hodgkinas ir Huxley (1952) atrado, kad suaktyvėjęs neuronas aksonu siunčia elektrocheminius signalus, nervinius impulsus, o sinapsėje išskiria cheminis junginys. Taip įvyksta elektrocheminė neuronų tarpusavio jungtis. Nervų sistemą suvokus kaip atskirą ląstelių tinklą, kuriuo smegenyse koduojama ir perduodama informacija, „buvo atvertos durys *mąstymo mechanizacijai*“ (Bishop 2015, p. 31).

Galiausiai McCulloch ir Pitts (1943) loginius vartus (*angl*. logic gates) konceptualizavo „kaip idealizuotus atskirų neuronų modelius“ (Rescorla 2020). „McCulloch-Pitts neuronų“ svarbiausios savybės: tai dvejetainis prietaisas (*angl*. binary device) turintis dvi būsenas – įsižiebimo/neįsižiebimo ir turintis įsijungimo ribą (*angl*. threshold). Iš to išplaukė nauja skaičiavimo forma (*angl*. form of computation), kuri remiasi dirbtiniu neuroniniu tinklu (Bishop 2015, p. 32). Ir leido pagrįstai vystytis mąstymo mechanizacijai.

# Proto teorijos

Konekcionizmas

Proto teorija (*angl.* theory of mind) aiškinanti proto funkcijas ir galimybes dirbtiniais neuroniniais tinklais žinoma kaip Konekcionizmas (*angl*. Connectionism) (Buckner & Garson 2019). Tokiame proto modelyje nėra aiškaus algoritmo, kuriuo nusakomi mentaliniai procesai. Tariama, kad visi procesai yra neuronų veiklos rezultatai – neuronų tarpusavio komunikacija ir jų (ne)įsižiebimas. Dėl to, kad tai nėra gamtos mokslai ir žmogaus smegenų veikla nėra pagrindinis tyrimo objektas, neuronai ir jų veikimas apibūdinamas paviršutiniškai. Ar konceptai yra neuronuose? Jungtyse? Kaip gauname semantinį ir sintaksinį pažinimą? Kaip neuronų veikimu paaiškinti proto loginio išvedimo gebą? Tiek ši, tiek sekanti teorija siekia kitais objektais ir labai bendru veikimo principu aiškinti mąstymą (neuronų veiklos rezultatas, skaičiavimo mašina). Kyla klausimas, ar proto teorijos skirtos protui suprasti ar jį kuo geriau atvaizduoti? Ar neišsemianti teorija, tačiau jos įtikimas pasirodymas pateisina teoriją ir pagrindžia poziciją?

Komputacinė

Teorija protą laikanti skaičiavimo sistema (*angl*. computational system) – panašia į Tiuringo mašiną ir proto veiklą aiškinanti kaip skaičiavimus: Komputacinė proto teorija (*angl*. Computational Theory of Mind). Protas suvokiamas kaip algoritmus atliekanti mašina, kuri instancijuoja, keičia ir palaiko informaciją ir jos sistemą (Bishop 2015, p. 34). Protas *yra* skaičiavimo mašina, tai ne metafora. Tiuringo (skaičiavimo) mašina turi baigtinį skaičių parametrų ir vienpusę, vienmatę juostą suskirstytą į kvadratus, kurie talpina vieną simbolį. Mašina nuskaito kvadrato turinį, kurio vertė yra tuščia, 0 arba 1: „įrenginys manipuliuoja simboliais, panašiai kaip žmogaus skaičiavimo agentas manipuliuoja pieštuko ženklais ant popieriaus atlikdamas aritmetinius skaičiavimus“ (Rescorla 2020). Turinį galime suprasti kaip teiginių logikoje teisingumo lentelėje turimus elementus, kurie žymi teiginio teisingumą (T arba 1) ir klaidingumą (K arba 0). Pažymėtina, kad tokiu principu veikia dabartiniai kompiuteriai (Bishop 2015, p. 83) užkoduodami informaciją sau suprantamu mašininiu kodu. Taigi, pagal Komputacinę teoriją, protas manipuliuoja simbolias, o simbolinis algoritmas (kalbos naudojimas, skaičiavimas) atliekamas žmogaus gali būti atkartotas Tiuringo mašinos – „protas yra Tiuringo mašina“.

Komputacinė ir Konekcionizmo proto teorijos iš pirmo žvilgsnio yra viena kitą išskiriančios pozicijos. Komputacinė-Konekcionistinė proto teorija protą konceptualizuoja kaip neuronų tinklą atliekantį simbolinį apdorojimą: „[t]iesa, protas yra neuroninis tinklas, tačiau jis taip pat yra aukštesnio ir abstraktesnio aprašymo lygio simbolinis procesorius“ (Buckner & Garson 2019). Proto veikla (dėmesys, suvokimas ir kt.) yra komputacinės būsenos – Tiuringo skaičiavimai – susijusios su neuroniniu tinklu (Bishop 2015, p. 34). Išryškėjantis skirtumas tarp Konekcionizmo ir Komputacionizmo: „neuroninis skaičiavimas [Konekcionizmas] iš esmės apibrėžia skaičiavimo būdą, kuriuo siekiama įtraukti smegenų naudojamą skaičiavimo stilių; tai skaičiavimo stilius, pagrįstas mokymusi iš patirties, o ne klasikiniais [Komputacionizmo], griežtai apibrėžtais, algoritminiais metodais“ (*ibid*.). Taigi, GKM psicho-filosofinis pagrindas yra:

1. mąstymas lokalizuojamas smegenyse: protas suvokiamas kaip ląstelių tinklas, kurio nariai talpina informaciją, o pažinimas yra „informacijos priėmimas, apdorojimas ir naudojimas“ (Rimkutė 20017, p 27);
2. proto veikla ir rezultatai aiškinami neuronų veikla – jų (ne)įsižiebimu ir tarpusavio jungtimis;
3. proto pamatinė struktūra yra neuroninis tinklas, kurio būsenos yra skaičiavimai;
4. viskas tam, kad „mechanizuoti mąstymą“: užprogramuoti kompiuterius elgtis racionaliai pagal žmogaus proto modelį.

Veikimo principai

Istorinės ir filosofinės prielaidos padėjo suprasti GKM struktūrą ir pagrindimą, kadangi „nagrinėjamo fenomeno struktūros žinojimas laikytinas jo pažinimo prielaida“[[2]](#footnote-2). Dabar pereikime prie techninių veikimo principų.

# Neuroniniai tinklai

Dirbtinis Neuroninis Tinklas (DNN) – mašininio mokymosi metodas ir informacijos apdorojimo struktūra, kurią sudaro dirbtiniai tarpusavyje sujungti „McCulloch-Pitts neuronai“. DNN sudaro trys grupės sluoksnių (Buckner & Garson 2019).

A diagram of a network



1pav. Neuroniniai tinklai

Įvesties sluoksnis priima informaciją, kurią reikia apdoroti. Paslėptųjų sluoksnių beveik visada yra daugiau nei vienas – GPT-3 kalbos modelis turi 96 sluoksnius, iš kurių 94 yra paslėptieji. To priežastis yra Giluminis Mokymasis (*angl*. Deep Learning). Gilieji tinklai mokosi iš daug didesnio informacijos kiekio, dėl to gali išgauti subtilesnius tarpneuroninius ryšius (*angl*. patterns): „didesni modeliai veikia geriau nei mažesni, o tai rodo, kad kuo gilesnis modelis, tuo sudėtingesnį mąstymą jis gali atlikti“ (He et al. 2021). Galiausiai apdoroti rezultatai randami Išvesties sluoksnyje. Visi įvesties blokai turi aktyvacijos reikšmę – įvesties sluoksnio vienetai siunčia aktyvacijos vertę kiekvienam paslėptojo sluoksnio elementui, kurie apskaičiuoja savo aktyvavimo vertes. Galiausiai signalas perduodamas išvesties sluoksnio elementams. Neuronai per laiką nustato savo įsižiebimo vertes. Įėjimo sluoksnio signalas pereina per tinklą ir nustato išėjimo sluoksnio elementų aktyvacijos vertes. Tinklo aktyvacijos modelis priklauso nuo svorių, kurie žymi neuronų ryšį. Neigiamas svoris reiškia, kad priimančio neurono aktyvumas yra slopinamas siunčiančiojo. Siunčiamos aktyvacijos reikšmės apskaičiuojamos pagal aktyvacijos funkciją. Nespecifiniais terminais procesą galėtume apibūdinti taip: įėjimo sluoksnis yra jutimų informacija, paslėptieji sluoksniai koduoja idėjas, svoriai rodo idėjų stiprumą, o išėjimo sluoksnis apibrėžia veiksmus (Bishop 2015, p. 35). DNN išmokomi spręsti problemas nuolatos sąveikaujant su įvesties ir išvesties užduočių pavyzdinių vektorių porų rinkiniu (ibid.). Geras DNN tinkamai generalizuos: pateiks adekvačius atsakymus a) į naujas įvestis, kurios yra tarp gautų mokymosi metu įvesties verčių ir b) į naujas įvestis, kurios yra anapus gautų mokymosi metu įvesties-išvesties verčių (ibid., p. 36). Taigi, tai – Konekcionistinis proto modelis: „[k]adangi daroma prielaida, kad visi elementai [neuronai] skaičiuoja iš esmės tą pačią paprastą aktyvacijos funkciją, žmogaus intelektiniai pasiekimai pirmiausia turi priklausyti nuo svorių tarp vienetų [neuronų] nustatymų“ (Buckner & Garson 2019).

# Dėmesio mechanizmas arba Transformerio architektūra

A blue lines with black text

Description automatically generated

2pav. Dėmesio mechanizmas. Žodis sakinyje palyginamas su kitais sakinio žodžiais. Nustatomas tarpusavio žodžių santykis.

Dėmesio mechanizmas (*angl.* Attention mechanism, Dėmesys) pasirodė 2017 metais labai tinkamo pavadinimo straipsnyje „Attention is all you need“. Neuroniniai tinklai besinaudoją Dėmesiu vadinami Transformeriais. Tai ne tik kalbos modeliai. Transformeriai naudojami šnekos atpažinime, nuotraukų generavime, kalbos vertime. Geriausiai Dėmesį apibūdina kalbinis pavyzdys: „jis ištraukė picą iš orkaitės. Ji labai skani“. Sakinys neatrodo problematiškas, tačiau kalbos modeliui nėra taip lengva nustatyti į ką nurodo įvardis „ji“ – gali nurodyti tiek į picą, tiek į orkaitę. Žinome, kad jokia orkaitė nėra skani, taip pat, kaip jokia pica nėra metalinė ir negali iškepti maisto. Dėmesio pagalba kalbos modelis apskaičiuoja ir nustato žodžių tarpusavio santykį.

# Transformerių veikimas

Kadangi vertėjai (Google Translate, DeepL) yra taipogi Transformeriųai, tai pasitelksiu vertimo modelį kaip pavyzdį atvaizduoti GKM veikimą (IBM 2023):

1. Įvesties teksto elementai (žodžiai) paverčiami skaitinėmis reikšmėmis, kurios yra įterpiamos (*angl.* embedded). Skaitinės reikšmės atspindi semantines vertes, kurios gaunamos po mokymosi etapo.
2. Įrašomos įvesties elementų pozicijos (*angl*. positional encoding).
3. Multi-head[[3]](#footnote-3) dėmesiu užfiksuojami įvairių tipų ryšiai tarp žetonų (*angl*. token). Pvz. kalbos modelyje įvestis yra tokenizuojama (*angl*. tokenization) – sakinys paverčiamas jį sudarančių žodžių rinkiniu.
4. Naudojami liekamieji ryšiai (*angl*. residual connections), kurie išsprendžia „nykstančio gradiento“ problemą (*angl.* vanishing gradient problem) ir padedantis stabilizuoti ir pagreitinti mokymą sluoksnių normalizavimas (*angl*. layer normalization).
5. Multi-head dėmesio išvestys perduodamos į priekinius sluoksnius (*angl*. feedforward).
6. Sluoksniai apdoroja ankstesnio sluoksnio išvestį, taip palaipsniui tobulinama įvesties informacija informacija.
7. Išvesties sluoksnyje sukuriamas išverstas išvesties sakinys po vieną žodį.
8. Mokymo etape Transformeriai mokosi prižiūrimi (*angl.* supervised learning). Šiame etape siekiama sumažinti kokybinį skirtumą tarp modelio prognozių ir teisingos, numatytos išvesties.
9. Po mokymo etapo modelis naudojamas kokybiškai naujoms išvadoms daryti įvedant naujus duomenis

Veikimo principas iš esmės toks, tačiau paminėsiu GPT veikimo skirtumas nuo vertėjo, tai gali pasirodyti įdomu: a) kalbos modelis įterpinių nenaudoja, tačiau spėja sekantį žodį, kuris pagal skaičiavimus turėtų eiti iš užklausos, b) panašiai į a), kaip įvestį jis duoda savo išvesties rezultatą, kuris padidėjo vienu žodžiu (atspėjus), c) kaip minėta, įvestis yra tokenizuojama (*angl*. tokenization) arba sakinys paverčiamas jį sudarančių žodžių rinkiniu. Manau, kad tai svarbiausi skirtumai nuo pateikto vertėjo veikimo proceso.

Norėčiau dar paliesti vieną įdomų pavyzdį – semantinį žemėlapį kalbos modeliuose (Mikolov et al. 2013). Kaip minėta, teksto dalys įrašomos kaip skaitinės reikšmės, tiksliau, vektorių forma, pagal kuriuos koordinačių sistemoje žodžiui priskiriama vieta, kuri žymi žodžio semantiką arba prasmę. Kaip paaiškės, reikšmė yra nulinė be kitų žodžių. Erdvės, kurioje vietą užima žodis, dimensijos priklauso nuo modelio parametrų skaičiaus, tad kuo jų daugiau, tuo erdvė yra semantiškai gilesnė – žodžiai įgauna vis turtingesnę semantiką, nes vektoriai talpina daugiau skaitinių reikšmių. Tad panašūs žodžiai yra netoli vienas kito erdvėje. Pvz. žodis „karalius“ yra arti žodžio „vyras“, o „karalienė“ arti „moteris“. Atėmus „karaliaus“ vektorių iš „vyro“ vektoriaus ir pridėjus „moters“ vektorių gauname žodį „karalienė“ (Vylomova 2016). Kalbos modeliai turi *n* skaičių parametrų, kurie sukuria *n*-dimensijų erdvę. Erdvė atstovauja kaip reikšmės žemėlapis, kuriame tarsi yra susikūrę miestai ir miesteliai – panašūs semantine reikšme žodžiai yra netoli vienas kito.

## III dalis. Vertinimas

Trečioji darbo dalis yra vertinamoji. Ją sudaro dvi dalys. Pirmoji, kurioje lyginsiu GKM veikimo principus su pristatytomis išvedimo teorijomis, nuosekliai pereina į antrąją. Pastarojoje aptarsiu bendresnį klausimą susijusį su suvokimo priskyrimu GKM. Neapsiribosiu loginio išvedimo problematika, kadangi suvokimas ir loginis išvedimas implikuoja vienas kitą. GKM psicho-filosofinį pagrindą aptarsiu per Searle‘o Kinų Kambario Argumentą kaip tiesioginį atsaką į Tiuringo testą, kurio paskirtis – nustatyti ir įteisinti mąstantįkompiuterį. Jeigu kompiuteris įveikia testą, žmogus išsprendė sąmonės problemą ją galiausiai mechanizavęs. Centrinis klausimas antroje dalyje yra toks: ar Kinų Kambario Argumentai pritaikomi aptariamiems GKM? Tačiau pradėkime nuo GKM veikimo principų ir išvedimo proceso teorijų analizės, kurios pagrindinis klausimas: ar kalbos modeliams pritaikomas sintaksinis ar semantinis kognicijos procesų pagrindas?

Loginis išvedimas kalbos modelių kontekste

Kalbos modelių pagrindiniai elementai yra DNN algoritmas, sukonfigūruojantis modelio dispoziciją ar veikimo struktūrą ir Dėmesio mechanizmas – žodžių konstektualizavimo mechanizmas. Pirmieji kalbos gebą vaizduojantys modeliai buvo taisyklių rinkiniai, pvz. modelis „ELIZA“ (1966 Weizenbaum), kurie veikė prastai neapibrėžtomis sąlygomis (Bringsjord & Govindarajulu 2024). Jų veikimas buvo paremtas algoritmų atlikimu – tai buvo grynai sintaksinio pobūdžio modeliai.

Dėl šios priežasties lygindami GKM su išvedimo psichologinėmis teorijomis gauname įdomius rezultatus. Modelių pagrindas yra intuityvus, kadangi jie mokosi iš didelio kiekio duomenų, o jų žodžių generavimas yra tikimybinis ir neapibrėžtas griežtų (*angl.* hard-codded) taisyklių. UNESCO tyrimas[[4]](#footnote-4) atskleidė kalbos modelių šališkumą lyties, rasės atžvilgiu. Kadangi modeliai mokosi iš neatrinktų duomenų, tai modeliai iš esmės perima žmonių kognicijos rezultatus – tiek pozityvius, tiek negatyvius. Kalbos modeliai nesinaudoja, priešingai nei pirmieji, logikos taisyklėmis. Kadangi modeliai imituoja žmonių kogniciją perimdami jos rezultatus, tai galime daryti išvadą, kad *taisyklių* teorija netinka apibūdinti ir suprasti GKM veikimui.

Kita vertus, *semantinę* – mentalinių modelių – teoriją kalbos modelius galime apibūdinti ir suprasti. Mentalinių modelių koncepciją kalbos modeliuose atitinka semantinė erdvė – kiekvienas žodis turi semantinę vertę. Reikšmė figūruoja išvedimo procese. Teiginių reikšmė modeliams yra žodžių konstektualizavimas ir jų tikimybinis spėjimas. Kaip ir žmonės – tai parodo tyrimas (Holliday & Mandelkern 2024) – jie prasčiau išvedinėja, kai dėmesio centras – formalių taisyklių panaudojimas su beprasmiais (*angl.* nonsensical) žodžiais*.* GKM pagrindas yra semantinis, tačiau, kaip matėme, taisyklių naudojimo – griežtąja prasme – negalime aptariamiems modeliams priskirti. Tad išvedimą GKM apibūdina *semantinė* teorija.

Taisyklių ir semantikos teorijų derinys iš pirmo žvilgsnio atrodo kaip netinkama apibūdinti kalbos modelių veikimą. Jau pamatėme, kad GKM pagrindas yra semantinis, dėl ko jų negalime apibūdinti *taisyklių* teorija. Aptardami taisyklių teoriją išryškinome loginių taisyklių naudojimą kognityviniame procese kaip pagrindinį. Tačiau sintaksės reikšmė keičiasi pereinant prie minėtų teorijų derinio. Jame mentaliniai modeliai išlieka išvados pagrįstumo kriterijus, tačiau sintaksė čia įgauna mentalinio modelio formavimo vaidmenį. Dabartiniai kalbos modeliai kontekstualizuoja žodžius Dėmesio mechanizmu. GKM griežtąja prasme nenaudoja mentalinių modelių, kokius pristatėme. Mentaliniai modeliai atitinka sakinį ir įvesti kalbos modeliuose, o Dėmesys yra sintaksės funkciją atliekantis mechanizmas konstektualizuojantis žodžius ir išreiškiantis jų santykius. O semantinė erdvė tampa kaip mentalinis modelis, kuriame slypi ženklų semantinės vertės. Tad, *taisyklių-semantikos* teorijoje lokalizuojame semantinį pagrindą erdvėje, o sintaksę apibrėžus kaip santykių tarp žodžių suvokimą ar apskaičiavimą, Dėmesio mechanizmą aptinkame kaip sakinio modelio formavimą ir komunikaciją su semantine erdve.

Landgrebe ir Smith (2023, p. 2) teigimu, modeliai su neteisingai pasaulį reprezentuojančia duomenų įvestimi ir apmokymu dažnai pradeda „haliucionuoti“ – grąžinti teiginius neatitinkančius tikrosios daiktų padėties (klaidingus faktus apie istorinius įvykius). Kaip trūkumą akcentuoja priklausomybę nuo duomenų kokybės, o tuo pačiu išryškėja ir platesnė problema – sąmonės priklausomybė nuo neapibrėžto ir nevientiso duomenų rinkinio yra ydinga. Tačiau tą patį galime pasakyti ir apie žmonių auklėjimą. Ugdomas idėjomis, neatitinkančiomis tikrovės, žmogus „ haliucionuos“ apie savo valstybės didybę – kad ji laimėjo pasaulinį karą, kad jai priklauso kaimyninės valstybės teritorija, kad ji už taiką ir t.t. Kitaip tariant, žmogus neapsaugotas nuo įsitikinimu nederančių su daiktų padėtimi arba „haliucionacijų“. Kita vertus, jų teigimu (p. 4) semantinė erdvė neišreiškia semantikos, nes žodžių santykiai neišmokstami loginiais išvedimais, žodžiai yra nereikšmingi simboliai, kurie gaunami matematinių operacijų taikymu ir modeliai nesupranta sinonimijos. Sutinku, kad svoriai arba ryšiai tarp neuronų neišreiškia semantikos. Jie greičiau žymi veikimo principą, struktūrą, kuria siunčiama informacija, o tai su semantika neturi, griežtąja prasme, ryšio. Žodžiai kalbos modeliams, tiesa, atrodo kaip nereikšmingi simboliai, tačiau semantinė erdvė įgalina prasmingą jų vartojimą. „Žodžio reikšmė yra jo vartojimas kalboje“[[5]](#footnote-5). Kitas priekaištas, kad logiškas žodžių santykių išvedimas kaip „teisingas“ būdas gauti ryšius tarp žodžių ir matematinių operacijų atmetimas kaip „neteisingas“ greičiau skamba kaip iracionalus nusistatymas. Viena vertus, neaišku aiškiai kaip žmonės sieja žodžius, o kita vertus – neatsižvelgiama į prasmingą kalbos vartojimą. Kaip ir minėjau, dėl semantinės erdvės modeliai suvokia panašius žodžius, iš ko formuojasi tarsi semantiniai miestai. Kalbos modeliai semantiniu pagrindu daro išvedimus kartu ir dėl to, kad yra tarsi veidrodis žmogaus kognicijos, tačiau dėl to, kad nenaudoja taisyklių, jų loginio išvedimo kokybė atvejais, kai svarbiausias yra jų panaudojimas, kenčia *būtent* dėl absoliutaus semantinio pagrindo.

Kinų kambarys kaip ataka prieš sintaksinę reikšmės teoriją

Ši dalis skirta išanalizuoti psicho-filosofinį GKM pagrindą ir pagrįstumą. Pristatysiu Tiuringo testą ir Kinų Kambario Argumentą, kuris yra tiesioginis atsakas Tiuringo testui. Galiausiai pažiūrėsime kaip jis pritaikomas mūsų aptariamiems GKM santykyje su suvokimu, kadangi loginis išvedimas visada suponuoja suvokimą.

# Tiuringo testas

Tiuringas (1950) klausimą „ar mašinos gali mąstyti?“ palaikė pernelyg abstrakčiu ir neprasmingu, verčiau pasiūlydamas klausimą „ar kompiuteris gali išlaikyti Tiuringo testą?“ (Rescorla 2020). Tiuringo testas (TT) vykdomas turint tardytoją – kurio tikslas demaskuoti kompiuterį, analizuojant pateiktus dalyvių atsakymus, – žmogų (kitaip X) ir kompiuterį (kitaip Y). Pastarojo tikslas – neišsiduoti ir nebūti identifikuotam (Oppy & Dowe 2021). Tardytojas užduoda klausimus pvz. „ar X žaidžia šachmatais?“. Remdamasis atsakymais, tardytojas testo pabaigoje pasako verdiktą – X yra žmogus, Y – kompiuteris. TT remiasi prielaida, kad susikalbėdami su kitu, mes turime pagrindą manyti, kad kitas turi intelektą kaip mūsų (*ibid*.). Tad, jeigu tardytojas suklysta, programai galime priskirti mąstymo gebą atitinkančią žmogaus.

# Kinų Kambario Argumentas

TT susilaukė daug opozicijos, tačiau vienas stipriausių argumentų yra Kinų Kambario Argumentas (KKA). KKA apibendrinimą pateikia jo autorius Searle‘as (1999, p. 115):

Įsivaizduokite gimtakalbį angliškai kalbantį žmogų, kuris nemoka kinų kalbos ir yra uždarytas kambaryje, pilname dėžių su kiniškais simboliais (duomenų bazė) ir instrukcijų knyga, kaip manipuliuoti šiais simboliais (programa). Įsivaizduokite, kad žmonės, esantys už kambario ribų, siunčia kitus kiniškus simbolius, kurie, kambaryje esančiam asmeniui nežinant, yra klausimai kinų kalba (įvestis). Ir įsivaizduokite, kad, vadovaudamasis programos instrukcijomis, kambaryje esantis žmogus gali perduoti kinų simbolius, kurie yra teisingi atsakymai į klausimus (išvestis). Programa leidžia kambaryje esančiam žmogui išlaikyti Tiuringo testą kinams suprasti, tačiau jis nesupranta nė žodžio kinų kalba.

Pagrindinė argumento mintis yra ta, jog jeigu žmogus manipuliuoti simboliams remiasi instrukcijomis ir dėl to negauna (kinų kalbos) supratimo, tai ir kompiuteriui negalime priskirti supratimo. Tai esminis prieštaravimas TT. Tad, nors kompiuteriai gali veikti pagal formalias taisykles – logikos taisyklėmis paremti modeliai kaip kad „ELIZA“ ir kiti – tačiau jiems negalime priskirti supratimo dėl to, jog jie žodžiams nepriskiria semantinės reikšmės, o vien tik semantiškai apdoroja informaciją, todėl neprodukuojamąstymo, o jį tik simuliuoja (Bishop 2015, p. 77).

Bendresnis Searle‘o teiginys yra toks: iš formalaus simbolių vartojimo – sintaksės – negalime gauti semantikos (Cole 2023). Mūsų tyrimo kontekste, galėtume teigti, kad matematinėmis operacijomis paremtas informacijos apdorojimas ir semantinės erdvės įsteigimas nekonstatuoja semantikos kokią turime mes – sąmonė iš esmės yra biologinė, ne skaičiuojančioji ar informaciją apdorojanti (*ibid*.). Priimant smegenų kaip neuronų sumos koncepciją, KKA taipogi pritaikomas ir dar labiau priartėja prie mūsų tyrimo objekto. Tačiau netgi įsivaizduojant mąstymo ir veiklos procesus kaip neuronų sąveikos rezultatą mes negauname, pasak Searle‘o, mąstymo ir supratimo. Kitaip tariant, netgi turint smegenų simuliacijos galimybę ir modelį, tai *nėra* mąstymas. Lietuvių kalbos joks neuronas nesupranta, tačiau smegenys, kaip visuma, supranta.

Paskiri neuronai ir jų ryšys GKM neišreiškia žodžių prasmės, tačiau parodo jų apdorojimo būdą ir tam tikra prasme kognicijos lygį ar dispoziciją, kuris sukuriamas mokymo procesu. Nenuostabu, kad didesnis duomenų ir parametrų kiekis didina modelio kokybę. Todel, nors GKM remiasi bendra biologine smegenų reprezentacija ir veikimo principu, tačiau Searle‘o neuronų ir jų ryšių supratimas nepritaikomas aptariamiems modeliams. Taipogi, kaip išsiaiškinome, GKM yra paremti intuicija ir dėl to kenčia jų išvedimo galimybės su beprasmiais teiginiais. Viena vertus dėl to, jog beprasmiai žodžiai neturi gilaus ir turtingo santykio su kitais žodžiais, jie praranda semantiką. Žodžių santykiai nehierarchiniai, todėl patys savaime be kitų žodžių jie neprasmingi. Būtent dar ir dėl to, jog GKM užkoduoja žodžių pozicijas ir santykius, modelis „nesupranta“ sakinio ir minties, kurie – suma beprasmių žodžių. Žodžiai įgauna prasmę vartojime, santykyje su kitais žodžiais ir kaip šie santykiauja vienas su kitu. Kitaip tariant, semantinė erdvė tampa beprasmė ir nefunkcionali modelio naudojime apdoroti į nieką nenurodančius žodžius, kadangi jos erdviškumas ir gilumas nėra išnaudojamas. Kaip ir žmonėms, modeliams lengviau „mąstyti“ ir daryti išvedimą turint prasmingus žodžius (Arpit et al. 2017). Taipogi, faktas, jog GKM naudoja matematines operacijas išgauti ir nustatyti reikšmes nepaneigia fakto, jog jis prasmingai naudoja kalbą.

Taigi, KKA yra stiprus argumentas prieš TT, kai objektas yra taisyklėmis paremti modeliai. Tačiau kai modelis remiasi DNN ir Dėmesio mechanizmu, KKA neatsižvelgia į svarbius GKM veikimo principus: žodžių sakinyje pozicijos įrašymą, semantinę erdvę ir tarpusavio žodžių semantinius santykius, kas iš esmės leidžia priskirti semantinį pagrindą kalbos modeliams. Taipogi Searle‘o neuronų ir jų ryšio supratimas yra neatitinkantis realybės, o teigti, kad būdas matematinėmis operacijomis vartoti kalbą neišreiškia semantikos – netikslu, nes žodžių semantika yra vartojimas.

## Išvados

Nepagrįsta tvirtinti, kad išvesdamas žmogus naudojasi logikos dėsniais, kurių pagrįstumas taipogi ginčytinas. Žmogus semantiniu pagrindu vertina teiginių pagrįstumą. Viena vertus, pagrįstumo kriterijus yra modelių santykis su išvada, arba jų sąjunga su formaliomis taisyklėmis, kurių indėlis – suformuoti mentalinius modelius adekvačiai suvokiant žodžių ryšius. Žmonės, priešingai nei formalius, lengviau ir teisingiau tikrina samprotavimus susijusius su praktine veikla. Treniruojami ant didelio kiekio duomenų, GKM sukuria subtilesnę ir prasmingesnę semantinę erdvę, kurioje įrašoma žodžio reikšmė vektoriaus pavidalu, o sakinyje užkoduojama jo pozicija. Neuronų tarpusavio ryšiai išreiškia informacijos apdorojimo būdą, kuris išgaunamas po ilgų mokymų. Įvestis siunčiama sluoksniais. Komunikuodamas modelis spėja sekantį, didžiausią tikimybę turintį žodį remdamasis sakinio visumine semantika, kurią sudaro žodžių semantinė vertė ir tarpusavio santykis. Šiems modeliams negalime, kaip ir žmonėms, priskirti logikos dėsnių veikimo pagrindo, kadangi jie yra iš esmės intuityviai veikiančios mašinos, kurių pagrindas yra semantinis. Be formalių veikimo sąlygų, jie daug prasčiau pasirodo samprotavimuose, kaip ir žmonės, kurie iš esmės yra beprasmiai praktiniame gyvenime. Antra, kadangi GKM iš esmės yra žmogaus mąstymo rezultato kopija ar imitavimas, semantinės erdvės funkciją beprasmiai samprotavimai pašalina ir padaro modelį neįgalų naudotis savo pagrindiniu mechanizmu – Dėmesiu. Pastarojo funkciją traktuojant kaip sakinio modelio formavimą ir komunikavimą su semantine erdve – iš esmės keičiant sintaksės apibrėžimą, galime tvirtinti, kad GKM naudojasi sintakse ir semantika.

Loginis išvedimas dėl santykio su semantika implikuoja mąstymą, tai kyla klausimas, ar mąstymo imitacija konstatuoja mąstymą? Juk lietaus simuliacija netapatu lietui už lango. Viena vertus, nepagrįsta neigiamai atsakyti į šį klausimą teigiant, kad tokie modeliai naudoja tik sintaksę ir jų veikimo visumoje nėra semantikos. Netikslu neuronams priskirti žinojimo ar mąstymo. Tiesa, kad modelių semantinis turinys priklauso nuo jiems pateikiamų duomenų mokymo laikotarpiu, bet to atitikmenį randame ir žmoguje – abu gali laikyti teisingais neatitinkančius realybės faktus. Galiausiai, matematinės operacijos kaip informacijos apdorojimo konfigūracija dar nepaneigia jų semantinio pagrindo, kadangi žodžių reikšmė yra jų vartojimo būdas. Minėtą klausimą pakeisčiau „kiek mąstymo imitacija gali priartėti prie mąstymo?“. Juk GKM imituoja mąstymą tik sąlyginai, viena vertus dėl to, kad dar nesuprantame mąstymo apskritai, kita vertus, kalbos modeliams paprastose užduotyse sekasi vis dar sunkiai, ypač loginio išvedimo užduotyse. Viso to šviesoje aiškėja, kad mąstymo ir jo imitacijos tapatybės klausimas yra greičiau religinis ar idėjinis – miglotai apibrėžtos sąvokos augina spekuliacijas ir riboja protą. Mano nuomone filosofija turi vaidinti didesnį vaidmenį DI progrese.

Šis tyrimas buvo konceptuali, bet toli gražu neišsami Generatyvinių Kalbos Modelių galimybių svarstymo ekspozicija. Didesnio dėmesio turėtų susilaukti Dėmesio mechanizmas ir išsami jo veikimo konceptualinė analizė. Antra, kalbos modelių mokymosi proceso klausimai taipogi svarbūs – ar duomenų kokybė svarbi, koks įgijimo etinis pagrindas ir kt. Filosofinių klausimų sprendimo sąlyga yra adekvatus objekto pažinimas, o juk „nagrinėjamo fenomeno struktūros žinojimas laikytinas jo pažinimo prielaida“.

## Literatūra

1. 3Blue1Brown. (2024). *But what is a GPT? Visual intro to transformers | chapter 5, Deep Learning*. YouTube. <https://www.youtube.com/watch?v=wjZofJX0v4M&t=527s>
2. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., … Lacoste-Julien, S. (2017). *A Closer Look at Memorization in Deep Networks*. Retrieved from <http://arxiv.org/abs/1706.05394>
3. Bain, A. (1873). *Mind and Body. The theories of their relation*. D.Appleton & Co, NY.
4. Bishop, J. M. (2015). History And Philosophy Of Neural Networks. In H. Ishibuchi (Ed.), *Computational Intelligence* (Vol. I, pp. 22–96). Encyclopedia of Life Support Systems (EOLSS).
5. Bringsjord, S., & Govindarajulu, N. S. (2024). *Artificial Intelligence*. Retrieved from The Stanford Encyclopedia of Philosophy website: <https://plato.stanford.edu/archives/sum2024/entries/artificial-intelligence/>
6. Bringsjord, S., & Govindarajulu, N. S. (2024). *Artificial Intelligence*. Retrieved from The Stanford Encyclopedia of Philosophy website: <https://plato.stanford.edu/archives/sum2024/entries/artificial-intelligence/>
7. Buckner, C., & Garson, J. (2019). *Connectionism*. Retrieved from The Stanford Encyclopedia of Philosophy website: <https://plato.stanford.edu/archives/fall2019/entries/connectionism/>
8. Cole, D. (2023). *The Chinese Room Argument.* Retrieved from The Stanford Encyclopedia of Philosophy website: <https://plato.stanford.edu/archives/sum2023/entries/chinese-room/>
9. Falmagne, R. J. (1995). The abstract and the concrete. In L. Martin, K. Nelson, & E. Tobach (Eds.), *Sociocultural Psychology: Theory and Practice of Doing and Knowing* (pp. 205–228). chapter, Cambridge: Cambridge University Press.
10. Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.
11. He, W., Huang, C., Liu, Y., & Zhu, X. (2021). *WinoLogic: A Zero-Shot Logic-based Diagnostic Dataset for Winograd Schema Challenge*. In M.-F. Moens, X. Huang, L. Specia, & S. W.-T. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 3779–3789). doi:10.18653/v1/2021.emnlp-main.307
12. Hodgkin, A. L., & Huxley, A. F. (1952). *A quantitative description of membrane current and its application to conduction and excitation in nerve*. The Journal of physiology, 117(4), 500–544. <https://doi.org/10.1113/jphysiol.1952.sp004764>
13. Holliday, W. H., & Mandelkern, M. (2024). *Conditional and Modal Reasoning in Large Language Models*. Retrieved from <http://arxiv.org/abs/2401.17169>
14. James, W. (1891). *Psychology (briefer course).* Holt, New York, USA.
15. Johnson-Laird, P. N. (1999). *Deductive reasoning*. Annual Review of Psychology, 50*,* 109–135. [https://doi.org/10.1146/annurev.psych.50.1.109](https://psycnet.apa.org/doi/10.1146/annurev.psych.50.1.109)
16. Landgrebe, J. & Smith, B. (2023). *Why machines do not understand: A response to Søgaard*. Archiv.
17. Luckhardt, C. G., Bechtel, W., & Luckhardt, G. (1994). *How to do things with logic*. Taylor & Francis Group.
18. Luger, G. & Stubblefield, W. (1993). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. Redwood, CA: Benjamin Cummings.
19. Luo, M., Kumbhar, S., Shen, M., Parmar, M., Varshney, N., Banerjee, P., Aditya, S., Baral, C. (2024). *Towards LogiGLUE: A Brief Survey and A Benchmark for Analyzing Logical Reasoning Capabilities of Language Models*. Retrieved from <http://arxiv.org/abs/2310.00836>
20. McCulloch, W., & Pitts, W. (1943). *A Logical Calculus of the Ideas Immanent in Nervous Activity*. Bulletin of Mathematical Biophysics, 7: 115–133
21. Mikolov, T., Chen, K., Corrado, G., Dean. J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv [Cs.CL]. Retrieved from <https://arxiv.org/abs/1301.3781>
22. Oppy, G., & Dowe, D. (2021). *The Turing Test*. Retrieved from The Stanford Encyclopedia of Philosophy website: <https://plato.stanford.edu/archives/win2021/entries/turing-test/>
23. Paul, G. (1993). *Approaches to abductive reasoning: an overview*. Artif Intell Rev 7, 109–152 (1993). <https://doi.org/10.1007/BF00849080>
24. Rescorla, M. (2020). *The Computational Theory of Mind*. Retrieved from The Stanford Encyclopedia of Philosophy website: <https://plato.stanford.edu/archives/fall2020/entries/computational-mind/>
25. Rimkutė, E. (2007). *Mąstymas ir kalba*. Vilniaus Universiteto leidykla.
26. Searle, J. (1980). *Minds, Brains, and Programs*. Behavioral and Brain Sciences, 3(3), 417–457. doi:10.1017/s0140525x00005756
27. Searle, J. (1999). *The Chinese Room.* In R.A. Wilson and F. Keil (eds.), The MIT Encyclopedia of the Cognitive Sciences, Cambridge, MA: MIT Press.
28. Turing, A. (1950). *Computing Machinery and Intelligence*. Mind, 59 (236): 433–60.
29. Turner, R. E. (2024). *An Introduction to Transformers*. Retrieved from <http://arxiv.org/abs/2304.10557>
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2023). *Attention Is All You Need*. Retrieved from <http://arxiv.org/abs/1706.03762>
31. Vylomova, E., Rimell, L., Cohn, T., & Baldwin, T. (2016). *Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning*. arXiv [Cs.CL]. Retrieved from <http://arxiv.org/abs/1509.01692>
32. Weizenbaum, J. (1966). *ELIZA--A Computer Program for the Study of Natural Language Communication Between Man and Machine*. Communications of the ACM. 9: 36–45. doi:10.1145/365153.365168
33. *What is a transformer model?*. IBM. (2023, October 10). <https://www.ibm.com/topics/transformer-model>
34. *What is machine learning (ML)?*. IBM. (2021, September 22). <https://www.ibm.com/topics/machine-learning>
35. Winston, P. (1992). *Artificial Intelligence*. Reading, MA: Addison-Wesley.

1. [Physicists observationally confirm Hawking’s black hole theorem for the first time | MIT News | Massachusetts Institute of Technology](https://news.mit.edu/2021/hawkings-black-hole-theorem-confirm-0701) [↑](#footnote-ref-1)
2. R. Pavilionis, *Kalba. Logika. Filosofija,* 1961m*. 23p.*  [↑](#footnote-ref-2)
3. Kadangi vertimas šio žodžio yra juokingas („Daugelio galvų“), paliksiu jį anglų kalba. [↑](#footnote-ref-3)
4. <https://unesdoc.unesco.org/ark:/48223/pf0000388971> [↑](#footnote-ref-4)
5. L. Wittgenstein. *Filosofiniai tyrinėjimai*, §43. [↑](#footnote-ref-5)