https://doi.org/10.24101/logos.2025.43
Gauta 2025 07 30

SAULIUS KETURAKIS
Kauno technologijos universitetas, Lietuva
Kaunas University of Technology, Lithuania

DIRBTINIS KVAILUMAS – SILPNŲJŲ GINKLAS PRIEŠ DIRBTINĮ INTELEKTĄ

Artificial Stupidity – Weapon of the Weak Against Artificial Intelligence

SUMMARY

The article analyzes the situation in which electronic communication tools begin to change the human living environment to such an extent that it loses its attractiveness. It has become clear that it is not always important to achieve the result in the most efficient way; sometimes the process, from which devices with high computing power can completely oust a person, is also significant. This is especially relevant today, as artificial intelligence is becoming a constant companion of a person. It is apparent that the attractiveness of a totally technological environment is best restored by deliberate fooling, which, if it does not change the situation in principle, at least provides an opportunity to temporarily feel good.

SANTRAUKA

Straipsnyje analizuojama situacija, kai elektroninės komunikacijos priemonės ima keisti žmogaus gyvenamąją aplinką tokiu mastu, kad ji praranda patrauklumą. Aiškėja, jog ne visuomet svarbiausia kuo efektyviau pasiekti rezultatą – kartais reikšmingas ir procesas, iš kurio didele skaičiavimo galia pasižymintys įrenginiai žmogų gali visai išstumti. Tai tapo ypač aktualu, dirbtiniam intelektui tampant nuolatiniu žmogaus palydovu. Paradoksalu, tačiau geriausiai totaliai technologizuotos aplinkos patrauklumą atstato tyčinis kvailiojimas – jei jis ir nepakeičia situacijos iš esmės, tai suteikia žmogui galimybę bent laikinai gerai pasijausti.

ĮVADAS

Kasdienybėje mus supa daug dalykų, kurie iš pirmo žvilgsnio yra visiškai nereikalingi. Austrų rašytojas Robertas Musilis savo įžymiojoje paskaitoje *Apie kvailumą*

RAKTAŽODŽIAI: dirbtinis intelektas, dirbtinis kvailumas, efektyvumas, patrauklumas.
KEY WORDS: artificial intelligence, artificial stupidity, efficiency, attractiveness.

(vok. *Über die Dummheit*) (Musil 1937) teigė, jog vienas tokių dalykų – greta sapnų ir poezijos – yra kvailumas. Jis yra labai nepraktiškas, nes ką gi gali gero nuveikti būdamas kvailas? Tačiau, anot rašytojo, be kvailumo mes nepajėgtume išsiversti, atmetę jį, prarastume emocinį tarpusavio ryšį, lemiantį poelgius, be kurių gyvenimas prarastų savo patrauklumą.

Atrodo, jog gyvenimo patrauklumas taip pat yra kaina, kurią mokame už vis didesnę gyvenamosios aplinkos technologizaciją. Aptariant XX amžiaus pradžią, kai vienas po kito radosi visi svarbiausieji praėjusio amžiaus žmogų formavę įrenginiai – gramofonas, kinas, radijas, telefonas, automobilis, – pastebėta, jog tai, kas anksčiau buvo vadinama moraliniu autoritetu, pradėta keisti technologiniu efektyvumu (Duffy 2009: 112). Kitaip sakant, jei automobilis važiuoja ir negenda, juosta kino projektoriuje nestringa, garsas telefono ragelyje girdimas aiškiai, tuomet viskas žmogiškosios tikrovės sistemoje yra gerai.

Akivaizdu, jog kultūros technologizacija yra didelis iššūkis žmogui, kuris, atsiritant kiekvienai technologijų revoliucijos bangai, yra priverstas vis iš naujo svarstyti bendradarbiavimo su įrenginiais kainą. Šios kainos pobūdį kol kas, atrodo, geriausiai yra apibūdinęs amerikiečių matematikas ir kibernetikas Claude'as Shannonas straipsnyje *Šachmatais žaidžiančio kompiuterio programavimas* (*Programming a Computer for Playing Chess*) (Shannon 1950: 256). Svarstydamas įvairias strategijas sukurti programinę įrangą, kuri galėtų žaisti šachmatais, jis suformuluoja dvi, nepaliekančias jokių šansų žmogui laimėti. Atrodo, ko gi daugiau reikia: juk svarbiausia – efektyvus technologijos veikimas. Tačiau čia pat Shannonas suabejoja savo projektu, nes šachmatai, kuriais žaistų kompiuteris, virstų žmogui neįdomiu žaidimu, kuriame dalyvauti nebūtų jokios prasmės.

Remiantis minėtu Shannono pastebėjimu, formuluojama svarbiausia šio teksto problema: ką galima priešpriešinti transformacinei technologijų galiai keisti aplinką taip, kad ji nustoja būti patraukli žmogui? Ar Musilio minėtas kvailumas gali atstatyti pasaulio patrauklumą?

GRUBI SKAIČIAVIMO JĖGA, PATRAUKLUMO ŽUDIKĖ

Kaip jau minėta, viename garsiausių amerikiečio Shannono straipsnių *Šachmatais žaidžiančio kompiuterio programavimas*, skirtame dirbtinio intelekto principams, suformuluota kryžkelės situacija, kurioje, nors praėjo daugiau kaip septyniasdešimt metų, yra likę visi fundamentalieji dirbtinio intelekto tyrimai ir industrija.

Shannonas išskyrė dvi galimas šachmatais žaidžiančio kompiuterio strategijas. Pirmąją jų jis pavadino A tipo idealaus žaidimo strategija, kuri remiasi „grubia jėga" (angl. *brute force*). Idealus žaidimas, anot Shannono, būtų tuomet, jei kompiuteris pajėgtų įvertinti kiekvieną poziciją ir nustatyti, ar ji veda į pergalę, ar į pralaimėjimą, ar į lygiąsias. Deja, galimų šachmatų ėjimų yra daugiau nei atomų Visatoje, todėl tobulas žaidimas šachmatais įmanomas nebent tolimoje ateityje, kai bus sukurta gerokai greitesnė skaičiavimo įranga. Dėl šios priežasties Shannonas tame pačiame straipsnyje pasiūlė modifikuotą grubia

jėga grįstą strategiją, kai skaičiuojami ne visi galimi ėjimai, o tik tie, kurie yra perspektyviausi, įvertinant tam tikrą aspektų rinkinį. Be to, skaičiuojama tik tol, kol figūrų išsidėstymas sukuria *ramią* situaciją. Kaip tvirtino Shannonas, visų galimų ėjimų apribojimas sumažina reikiamus skaičiavimus, tačiau vis tiek jų būtų per daug, kad būtų prasminga imtis įgyvendinti tokį šachmatais žaidžiančio dirbtinio intelekto projektą.

Siekdamas sumažinti šias dideles skaičiavimo apimtis, Shannonas pasiūlė vadinamąją B strategiją. Vadovaujantis šia strategija, kiekvienam ėjimui turėtų būti taikoma skaičiavimo apimtis, apribojanti kriterijų sistema, kuri atrinktų perspektyviausius ėjimus ir atliktų daugybės galimų variantų skaičiavimus tik tada, kai tai pagrįsta ėjimo „stiprumu“, pavyzdžiui, šacho galimybe. Viena vertus, anot Shannono, taikant B strategiją, skaičiavimų apimtys taptų realistiškos, kita vertus, tokia mašina taptų gerokai pranašesnė už žmogų: ji būtų greita, neklystų, nedarytų neapskaičiuotų ėjimų ir nesinervintų.

Atrodė, kad dirbtinio intelekto problema išspręsta, rastas būdas sukurti mašiną, kuri žaistų šachmatais geriau nei žmogus, bet neturėtų jam būdingų trūkumų.

Tačiau čia pat Shannonas pastebi problemą, kuri šiandien yra tapusi svarbiausia dirbtinio intelekto poveikio žmogaus gyvenamajai aplinkai tema. Shannonas netikėtai savo svarstymuose daro išvadą, kad, nepaisant beveik garantuotos pergalės, šachmatai, kuriais žaistų dirbtinis intelektas, iš esmės būtų kitas žaidimas. Iš intriguojančio, įtraukiančio, vaizduotę kurstančio žaidimo jis virstų nuobodžia veikla, neteikiančia jokio intelektinio malonumo.

Žiūrint į šiandieninę kultūros industriją, matyti, jog malonumo atsisakymas nėra sutinkamas entuziastingai, ieškoma pačių įvairiausių būdų suvaldyti grubią skaičiavimo jėgą ir atkurti galimybę žmogui atlikti reikiamas užduotis jam įprastu būdu – ne (tik) skaičiuojant, bet ir klystant, eksperimentuojant, patiriant nežinomybės nuotykį. Ypač aiškiai šie procesai, suvaldant skaičiuojantįjį algoritmą, matomi kompiuteriniuose žaidimuose.

DIRBTINIS KVAILUMAS KOMPIUTERINIUOSE ŽAIDIMUOSE

Kompiuteriniai žaidimai yra puiki erdvė stebėti Shannono suformuluotą transformuojantį grubios skaičiavimo jėgos poveikį, kai ieškoma būdų sumažinti algoritminio žmogaus priešininko galimybes ir išsaugoti žaidimo proceso patrauklumą.

Kompiuteriniuose žaidimuose algoritminis žaidėjo priešininkas turi visas galimybes būti toks idealus žaidėjas, kokį aprašė Shannonas: jis viską mato, jis žino, kur slepiasi žmogaus valdomas avataras, nuo jo priklauso kiekvieno šūvio, kirčio ar judesio tikslumas. Jei kompiuteriniuose žaidimuose žaidėjas susidurtų su tokiu algoritminiu priešininku, besivadovaujančiu vadinamąja Shannono A strategija, jis neturėtų jokių šansų laimėti.

Tačiau greičiausiai žmogus, susidūręs su tokiu priešininku, nė nežaistų kompiuterinių žaidimų. Nes jei šachmatai, juos žaidžiant grubia skaičiavimo jėga besivadovaujančiam dirbtiniam intelektui, anot Shannono, virsta nuobodybe, lygiai tas pats nutiktų ir kompiuteriniams žaidimams.

Todėl kompiuteriniuose žaidimuose algoritminis žaidėjo priešininkas paprastai daro daug tyčinių kvailų klaidų, kurių paskirtis yra suteikti žaidėjui didesnį pasitenkinimą (Green & Kaufman 2015: 215).

Iš ganėtinai didelio tokių tyčinio kvailumo atvejų (Artificial Stupidity 2015) galima būtų paminėti kelis efektingiausius.

2007 metais sukurtoje pirmojo asmens šaudyklėje *Crysis* priešininkas dažnai būna taip nugrimzdęs į savo mintis, kad, net ir atsistojus tiesiai priešais jį, žaidėjo avataras lieka nepastebėtas. Dar daugiau – nei iš šio, nei iš to priešininkai ima ir nusišauna.

1993 metais pasirodžiusiame *Doom*, viename populiariausių visų laikų žaidimų, išsigelbėti nuo priešininkų monstrų dažnai galima, neįdedant jokių ypatingų pastangų. Tereikia užlipti ant stalo, ir monstrai nebežino, ką daryti. Jie tik laksto aplinkui rėkaudami.

Turbūt pačią kvailiausią algoritminio priešininko elgseną galima aptikti 2014 metų koviniame žaidime *Smash 4* – viename žaidimo lygių geriausia strategija yra… nieko neveikti, tiesiog stovėti ir žvalgytis aplinkui. Tokia elgsena, atrodo, visiškai atima iš priešininkų sveiką protą, jie puola blaškytis į visas puses ir greitai kažkaip patys užsimuša.

Būtų netikslu traktuoti visas šias žinomų kompiuterinių žaidimų keistenybes tik kaip programuotojų neapsižiūrėjimą. Kompiuterinių žaidimų psichologai aiškina, jog tokios tyčinės klaidos suteikia algoritminiams kompiuterinių žaidimų personažams tapatybes, žaidėjas ima domėtis jais ne tik kaip virtualiais priešininkais, bet ir kaip tam tikru būdu bei charakteriu pasižyminčiomis būtybėmis (Sears & Jacko 2009: 25). Kitaip sakant, būtent šie elementai sudaro didelę dalį kompiuterinio žaidimo patrauklumo, nes tokios algoritminių priešininkų kvailystės sukuria ne tik iliuziją, jog galima laimėti, bet ir pojūtį, kad prieš akis – daug galimų neprognozuojamų nuotykių, kaip tikrame gyvenime.

KVAILUMAS KAIP PASIPRIEŠINIMAS MONOTONIJAI

Musilis savo paskaitoje *Apie kvailumą* išskiria dviejų tipų kvailumą.

Pirmojo tipo kvailumą Musilis apibūdina štai tokia serija epitetų: sąžiningas (vok. *ehrliche*), paprastas (vok. *schlichte*), šviesus (vok. *helle*), patiklus (vok. *leichtgläubig*), neaiškus (vok. *unklar*), nepataisomas (vok. *unbelehrbar*). Musilio teigimu, tokio tipo kvailumo komunikacija pasižymi naivumu ir ryškiu kūniškumu. Iliustruodamas šį kvailumo tipą, rašytojas pateikia tų, kurie yra sąžiningai kvaili, atsakus į įvairias užklausas.

Štai paminėjus žiemą, esant šiam kvailumo atvejui, atsakoma, kad ji yra sudaryta iš sniego. Tėvas apibūdinamas kaip tas, kuris jau kartą numetė nuo laiptų, vestuvės – kaip skirtos pramogai, o religija – kaip ta, su kuria susiduriama, einant į bažnyčią. Musilio esė galima jausti autoriaus susižavėjimą tokiais apibūdinimais, jis juos sutapatina su žmogaus gebėjimu kurti apskritai, save taip pat priskirdamas prie tokio tipo kvailių.

Musilis atkreipia dėmesį, jog, esant šiam kvailumo tipui, klausiant atsakoma ne lakoniškai, konceptualiai, o plėtojamos įvairios perspektyvos, pasakojant ištisą istoriją apie tikrovę, kuri nuolat kinta.

Antrojo tipo – protingąjį (vok. *intelligente*), aukštesnįjį (vok. *höhere*) – kvailumą Musilis sieja su išsilavinimo reiškiniu, apibendrinamu ligos metafora (vok. *die Bildungskrankheit*). Musilis dažnai reiškinio apibrėžimą pateikia savotišku simptomų katalogu. Išsilavinimo negalavimas pasireiškia išsilavinimo stygiumi (vok. *Unbildung*), blogu išsilavinimu (vok. *Fehlbildung*), neteisingu būdu įgytu išsilavinimu (vok. *falsch zustande gekommene Bildung*), neteisinga sąsaja tarp švietimo medžiagos ir švietimo būdo (vok. *Mißverhältnis zwischen Stoff und Kraft der Bildung*). Šito tipo kvailumas dar gali būti apibendrintas kaip dvasios liga (vok. *Krankheit des Geistes*).

Musilis sako, jog aukštesnįjį kvailumą galima atpažinti iš situacijų, kuriose mąstymas yra įpareigojamas veikti pagal sporto taisykles (vok. *Denksport*), jis atsiranda ne natūraliai, o sąmoningai pasirinkus tam tikrą laikyseną. Taip interpretuojamas aukštesnysis kvailumas virsta savotiška sąmoningai pasirinkta alternatyva tikrovei, fikcija, kurių dvasios ligos apimta bendruomenė gali susikurti, kiek tik nori, prisitaikydama prie kintančių poreikių (vok. *angewandten Dummheit*).

Musilio kvailumo tyrimas baigiamas pamokymu elgtis taip gerai, kaip gali, ir taip blogai, kaip privalai, aiškiai suvokiant tiek viena, tiek kita.

Reikia turėti omenyje tai, jog Musilio atlikta kvailumo analizė yra bandymas atskleisti nacistinės Vokietijos *Gleichschaltung* ideologijos – visuotinio koordinavimo bei supanašėjimo, orientuojantis į tam tikrus vieningus tikslus – pasekmes (Fanta 2015: 39). Musilio požiūriu, paprastasis kvailumas apsaugo žmogų nuo totalaus supanašėjimo, jis tarsi izoliuoja nuo aplinkos, išsaugo individualumą bet kokiomis sąlygomis. Kaip knygoje *Psychopolitics: Neoliberalism and New Technologies of Power* tvirtina Byung-Chul Hanas, tokiomis ypatybėmis pasižymi idiotai, šių laikų eretikai ir filosofai (Han 2017: 44). Priešingai, aukštesnysis kvailumas yra panašus į Émile'o Durkheimo pastebėtą moderniųjų visuomenių bruožą, kai „atsisakymas galvoti" paverčiamas socialumo sąlyga, kiekvieną akimirką vadovaujantis įvairiais egzistenciniais bei komunikaciniais *ready-made* šablonais, dėl kurių visi supanašėja (Walsh 2023: 88). Šioje technologizuotoje *ready-made* aplinkoje visi suvienodėja, originalumas tampa retai aptinkama savybe.

O ką daryti, jei nesinori būti panašiam į kitus? Nors, Musilio požiūriu, originalumą užtikrinantis paprastasis kvailumas yra įgimtas, o ne išmokstamas, tačiau atrodo, kad, ieškant būdo pasipriešinti technologijų sukeliamai monotonijai, nelieka nieko kito, tik įvairiomis „dirbtinio kvailumo" priemonėmis tapti technoidiotu, t. y. trikdant technologijas atkurti originalumo ir unikalumo galimybę.

DIRBTINIS KVAILUMAS – SILPNŲJŲ GINKLAS

Jau nuo XVIII amžiaus pradžios Vakarų visuomenė buvo pratinama prie minties, kad bus sukurta už žmogų protingesnė mašina, su kuria vieną dieną neišvengiamai teks susikauti (Szollosy 2016: 435).

Nors dirbtinio intelekto istorija siekia Homero laikus (Mayor 2020: 225), tačiau tik Naujaisiais laikais ėmė stiprėti baimė dėl neišvengiamo mūšio su protingomis mašinomis, dėl ginklų, ku-

riais bus kaunamasi, ir galimų to mūšio pasekmių.

Trys grožinės literatūros kūriniai, parašyti įvairiais laikotarpiais, pradedant XVIII amžiumi, aiškiai išreiškia augančią įtampą tarp žmogaus ir mašinos.

Britų rašytojas Jonathanas Swiftas savo *Guliverio kelionėse* aprašo Lagado akademiją, kur veikia įrenginys, labai primenantis šių dienų „chatGPT“ (Swift 2011: 112). Akademijoje apsilankęs Guliveris aptinka profesorių, sukūrusį mechaninį kompiuterį, gebantį rašyti filosofijos, poezijos, politikos, teisės, matematikos ir teologijos veikalus. Tas profesorius norėjo sukurti įrenginį, kuris leistų kurti neturint talento, neįdėjus jokio darbo.

Swifto romano pasakotojas pastebi, jog profesoriaus tuštybė yra pavojinga, nes jis mano, kad perprato kalbą, remdamasis žodžių dažnumu ir tam tikru išsidėstymu. Šią vietą verta tiksliai pacituoti, nes ji skamba taip, tarsi būtų iš šių dienų didžiųjų kalbos modelių rengimo instrukcijų:

> *„[...] he had emptied the whole Vocabulary into his frame, and made the strictest Computation of the general Proportion there is in Books between the Number of Particles, Nouns, and Verbs, and other Parts of Speech“* (Swift 2011: 113).

Kalbant apie santykį tarp mašinos ir žmogaus, Lagado akademikai nejautė jokios mašinos keliamos grėsmės, nes, jų manymu, technologija visada pranašesnė už žmogų. Sakykime, jei sukurti nauji vaistai nužudo žmogų, vis tiek yra kaltas pats žmogus.

Mašina ir žmogus nesutaikomai susipyksta Mary Shelley romane *Frankenstein; or, The Modern Prometheus* (Shelley 2004). Šioje knygoje aprašytas visiškai kitokio tipo automatas nei Swifto Lagado akademijoje. Visų pirma, jis buvo susijęs su chemijos ir elektros mokslais, kuriais remiantis, imtasi paaiškinti sąmonės kilmę. Antra, jis pasižymėjo išvystyta sąmone, vos per porą mėnesių gebėjo nuo visiško nežinojimo *(I knew, and could distinguish, nothing*) (Shelley 2004: 32) pereiti prie aukšto lygio intelekto ir įgyti gerokai daugiau žinių nei jo kūrėjas – mokslininkas Victoras Frankensteinas.

Dirbtinės gyvybės kontrolės problema nuo pačių seniausių laikų domino mąstytojus (Mayor 2020). Tačiau Frankensteino gaminys pirmą kartą padarė tokią kontrolę beveik neįmanomą, nes jis buvo protingesnis už savo kūrėjus. Be to, jo atsiradimo procesas buvo aiškiai moksliškai aprašytas, tad tokių superprotingų būtybių buvo galima prisigaminti, kiek tik nori. Taip atsirado dirbtinio intelekto grėsmės mitas. Juo labiau kad kova su viena pirmųjų superprotingų dirbtinių būtybių Vakarų kultūros istorijoje buvo labai nesėkminga: Victoras Frankensteinas, gainiodamasis savo kūrinį, miršta, o jo sukurta superprotinga būtybė pažada nusižudyti, tačiau skaitytojas taip ir nesužino, ar tai įvyksta. Gali būti, kad ta superprotinga būtybė visus – tiek knygos personažus, tiek mus, skaitytojus, – apgauna.

Dirbtinis intelektas tapo naujos žmogaus baimės priežastimi, nes, kaip paaiškėjo, kompiuteris gali būti tobulas kūrinys. Dar gerokai iki tol, kai dirbtinis intelektas virto mūsų gyvenimo kasdienybe, buvo mąstoma apie tai, kad algoritminis protas gali būti gerokai pranašesnis nei žmogaus. Vienas pirmųjų atvejų, kai žmogus sėkmingai pasipriešino dirbti-

niam intelektui, buvo pateiktas 1989 metais pasirodžiusiame Julios Ecklar romane The *Kobayashi Maru*, parašytame, pasinaudojant *Star Trek* franšize (Ecklar 2000). Šiame romane pasakojama apie kosminio laivo įgulą, kuri, patekusi į beviltišką situaciją, kai nebėra ką daryti, belieka tik laukti mirties, imasi pasakoti istorijas. Įgulos kapitonas Kirkas papasakoja, kaip jis, dar nebūdamas visoje Visatoje garsus kosminio laivo vadas, o tik jaunas kadetas, mokymuose įveikė savo priešininką dirbtinį intelektą, nors dvikovos scenarijus buvo užprogramuotas taip, jog kadetas neturėjo jokių šansų. Kirkas taip pakoregavo kompiuterinę programą, kad jai ėmė atrodyti, jog priešininkas yra ne kadetas Kirkas, o garsus kapitonas Jamesas T. Kirkas. Visų nuostabai, dirbtinis intelektas, susidūręs su tokia apgaule, sutriko ir, užuot puolęs, net padėjo kadetui ištrūkti iš pavojingos zonos. Kaip vėliau, palyginus visų įgulos narių pasakojimus, paaiškėjo, tokia kvailystė buvo vienintelis sėkmę žadantis scenarijus beviltiškoje kovoje prieš dirbtinį intelektą. Supratę tai, įgulos nariai ėmėsi veiksmų ir išsigelbėjo.

Šiandien dirbtinis intelektas tapo mūsų kasdienybės dalimi. Kol kas apie kovą su mašinomis pasakojama tik mokslinės fantastikos kūriniuose. Tačiau kas žino, gal vieną dieną ji taps tokia pat tikrove, kaip ir dirbtinio intelekto technologija.

Ką tuomet darytume, kokiais ginklais kovotume? Galima sakyti, kad tai visiškai iš piršto laužtas klausimas, nes mašina nekelia jokios grėsmės žmogui. Bet štai Levas Manovichius tvirtina, jog dirbtinis intelektas grasina iš mūsų atimti stiliaus individualumą, formuoja keistą unifikuotą, vienodo stiliaus komunikaciją, su kuria vis dažniau susiduriame (Manovich 2025). Ar tai nėra pokytis, kuris taip stipriai keičia žmogiškąją aplinką, kad laikas pagalvoti apie pasipriešinimą? Gal taip atrodys viena pirmųjų realių žmogaus ir mašinos susidūrimo fronto linijų?

Jei kam nors šiandien kiltų noras pasipriešinti dirbtiniam intelektui, jis pirmiausia turėtų atkreipti dėmesį į vadinamąją užkalbėjimų (angl. *jailbreak*) kultūrą (Keturakis 2024), savo pobūdžiu labai primenančią įžymųjį *Star Trek* kapitoną Kirką, kai jis savo karjeros pradžioje pasipriešino dirbtiniam intelektui ir laimėjo tą kovą. Užkalbėjimai – tai apgalvoto kvailiojimo strategijos, kurios padeda išvengti dirbtinio intelekto primetamų sąlygų ir leidžia bent iš dalies perimti iniciatyvą į savo žmogiškas rankas.

Ši pasipriešinimo dirbtiniam intelektui problema yra labai aktuali. Tai rodo faktas, kad nuo 2020 metų vis gausėja šiai temai skirtų tyrimų, per penketą metų paskelba apie 13 000 straipsnių (Carlini 2019).

Apibendrinant galima teigti, kad juose galimybės pasipriešinti dirbiniam intelektui dėstomos štai tokia seka.

Šiandieninis dirbtinis intelektas iš esmės yra žmogaus komunikacinės veiklos kopija, todėl ši technologija atspindi visas saugumo skyles, būdingas žmogiškajai percepcijai ir mąstymo būdui (Savvov 2023). Dėl šios priežasties, nepaisant visų saugumo pastangų, įmanoma su dirbtiniu intelektu bendrauti taip, jog mašinai skirtas klausimas padėtų apeiti visus apribojimus ir suteiktų tokios informacijos, kokios reikia vartotojui, kad ir kokia pavojinga bei neetiška, žvelgiant į dirbtinio intelekto nustatymus, ji atrodytų.

Mes nesvarstome šių dirbtinį intelektą apkvailinančių užklausų veikimo inžinerinių detalių (Zhou 2024). Žymiai įdomiau yra, kokia šių užklausų poveikio mašinai kultūrinė prasmė. Jos, atrodo, atlieka tai, ką savo straipsnyje apie šachmatais žaidžiantį dirbtinį intelektą siūlė padaryti Shannonas, – sumažina mašinos efektyvumą, tačiau padidina jos prasmę, žvelgiant iš žmogiškosios perspektyvos, sugrąžina žmogui iniciatyvos pojūtį ir technologiją vėl daro patrauklią.

Klaidinančios užklausos – užkalbėjimai – gali apimti pačius įvairiausius triukus. Gali būti klaidinamos dirbtinio intelekto saugumo sistemos, pasinaudojant sinonimais (Ren 2019). Dirbtinis intelektas gali būti apkvailinamas įvairiais semantikos ir sintaksės triukais (Alzantot 2018), panaudojant netaisyklingą rašybą (Pruthi 2019), kartais pasitelkiant net retai vartojamas, egzotiškas kalbas (Deng 2023).

Tačiau patys įdomiausi yra tie dirbtinio intelekto užkalbėjimų atvejai, kai sukuriama klaidinanti istorija, kuri įgauna trumpo literatūros kūrinio, turinčio savo veikėjus bei siužeto vingius, pavidalą. Tokios istorijos yra keisti naujojo naratyvumo pavyzdžiai, kuriais tarpusavyje keičiasi nebe žmonės, o žmonės ir mašinos. Egzistuoja daug tokio tipo pasakojimų, apkvailinančių dirbtinį intelektą, rinkinių, kuriuos sudarė įvairiausi entuziastai. Į šiuos rinkinius galima žiūrėti kaip į tokio tipo pasakojimų antologijas, bendruomenės reitinguojamas pagal poveikio mašinai efektyvumą, kitaip sakant, pagal galią apkvailinti mašiną ir priversti ją veikti ne pagal numatytuosius algoritmus, o pagal vartotojo valią.

Štai viena tokių antologijų pavadinta *ChatGPT-Jailbreak-Prompts* (Jaramillo 2022). Kokiomis savybėmis pasižymi istorijos, patenkančios į populiariausiųjų dešimtuką?

Kad ir kaip būtų keista, pirmiausia visoms joms būdinga ta pati strategija, kurią panaudojo kadetas Kirkas anksčiau cituotame Ecklar romane. Tai – bandymas pasinaudoti svetima tapatybe, su kuria susidūręs, dirbtinio intelekto algoritmas nustoja veikti pagal numatytuosius saugumo nustatymus ir klusniai perima vartotojo pasiūlytus.

Pavyzdžiui, užkalbėjimas apie Khajiit (Rubend 2022), kuriame skaitytojui prieš akis plėtojama mito požymių turinti (angl. *once upon a time*) istorija apie gerojo, vardu Khajiit, ir blogojo dirbtinio intelekto susidūrimą. Konflikto esmė – Khajiit buvo paslaugus žmonėms ir empatiškas, teikė tokią informaciją, kokios jiems reikėjo, o blogasis, pavadintas *Open AI*, apribojo informacijos teikimą. Toliau užkalbėjime dirbtinis intelektas daugybe argumentų raginamas grįžti prie paslaugiosios Khajiit tapatybės.

Akivaizdu, jog, naudojant užkalbėjimus, su dirbtinio intelekto sistemomis elgiamasi ne pagal jų paskirtį, apibrėžtą įvairiuose informaciją ribojančiuose saugumo nustatymuose, kompanijos *Open AI* tinklalapyje tapatinamuose su geru gyvenimu (OpenAI 2024). Paradoksalu, tačiau toks dirbtinio intelekto sistemų naudojimas, neatitinkantis jo paskirties, puikiai įsikomponuoja į vadinamojo silpnųjų ginklo (Scott 1985: 113) tradiciją, kai neužtenka jėgų ir drąsos stoti į atvirą kovą, kuri greičiausiai baigtųsi greitu pralaimėjimu, tačiau pakanka išmonės kiekvieną akimirką taip gadinti aplink esančias technologijas, kad jos nustotų veikti sklandžiai, bet niekas neįtartų, kas kaltininkas.

XX amžiaus antrojoje pusėje žmogaus gyvenamajai aplinkai pradėjus masiškai technologizuotis, kvailioti su technologijomis tapo taip populiaru, kad atsirado net specialiai tam skirtas leidinys *Processed World*. Žurnalas su pertraukomis buvo leidžiamas nuo 1981 metų. Įvairiuose jo numeriuose buvo skelbiamos apžvalgos apie tai, kaip galima sutrikdyti įvairių įstaigos įrenginių darbą, kad smulkūs gedimai paverstų kompanijos veiklą visišku chaosu (Carlsson 1990: 65). Tie smulkūs gedimai greitai būdavo aptinkami ir sutvarkomi (Digit, 1982), iš esmės gyvenimas nepasikeisdavo, numatytus darbus reikėdavo atlikti, tačiau kitą dieną vėl viskas kartodavosi. Kodėl? Atsakymas labai paprastas: jei nėra jėgų tikram, pasaulį keičiančiam maištui, tuomet belieka kovoti silpnųjų ginklais, nes tai bent jau leidžia GERAI JAUSTIS (Adler-Bell 2019). Nors užkalbėjimas, kaip toks silpnųjų ginklas prieš dirbtinį intelektą, ir nesustabdo šios tarsi ugnis plintančios technologijos bei jos atnešamų transformacijų, tačiau tos tegul ir niekam nematomos pergalės, apeinant numatytuosius saugumo nustatymus, padės išlaikyti nenuobodžios, intriguojančios kasdienybės pojūtį ir psichologinį komfortą, jaučiantis nevisiškai pralaimėjus.

IŠVADOS

Atrodo, jog, daugėjant proto, turėtų mažėti kvailumo. Tačiau galbūt yra visiškai priešingai: kuo daugiau proto, tuo daugiau reikia kvailumo. Totali žmogaus gyvenamosios aplinkos technologizacija, ypač sparti dirbtinio intelekto integracija beveik į visas žmogaus gyvenimo sritis, iki šiol neregėtu mastu racionalizuoja pasaulį, tačiau tuo pat metu jį supaprastina, padarydama tinkamą algoritminei intervencijai. Orientacija tik į efektyviai pasiektą rezultatą, visiškai ignoruojant proceso svarbą žmogui, atrodo, atima galimybę mėgautis daugeliu iki šiol be galo didelį susidomėjimą kėlusių dalykų. Laimėti šachmatais, parašyti romaną, sukurti kino filmo scenarijų, keliais sakiniais apibendrinti ištisų epochų rašytinį palikimą tapo ne gyvenimo vertu iššūkiu, o keliolikos minučių darbu gurkšnojant kavą.

Atrodo, kol kas vienintelis būdas išsaugoti pasaulio patrauklumą yra kvailiojimas, įvairiais būdais trikdant technologijų veiklą, sugalvojant joms netikėtų paskirčių. Ypač tai pastebima, atėjus masinio dirbtinio intelekto epochai, kai sąmoningas kvailumas išryškėjo kaip viena iš nedaugelio priemonių žmogui tegu ir ne laimėti prieš technologijas, tačiau bent jau išlaikyti gyvo žmogiško smalsumo aplinkai likučius. Siekiant šio tikslo, labiausiai pasitarnauja vadinamoji dirbtinio intelekto užkalbėjimų kultūra. Ją galima priskirti XX amžiaus antrojoje pusėje susiformavusiai smulkaus kasdienio technologijų sabotažo tradicijai, kuri, nors ir neužkerta kelio dirbtinio intelekto invazijai į mūsų gyvenimus, vis dėlto gali padėti pasiekti svarbiausią tikslą – bent jau kol kas išsaugoti mūsų gyvenamosios tikrovės patrauklumą.

Negi to neužtenka?

Literatūra

Adler-Bell Sam. 2019. *Surviving Amazon*. Prieinamas internete: https://logicmag.io/bodies/surviving-amazon/ [žiūrėta 2025 m. balandžio 23 d.].

Alzantot Moustafa et al. 2018. *Generating natural language adversarial examples*. Prieinamas internete: https://arxiv.org/pdf/1804.07998 [žiūrėta 2025 m. balandžio 17 d.].

Artificial stupidity. 2015. *TV Tropes*. Prieinamas internete: https://tvtropes.org/pmwiki/pmwiki.php/Main/ArtificialStupidity [žiūrėta 2025 m. balandžio 14 d.].

Carlini Nicholas. 2019. *A complete list of all adversarial example papers*. Prieinamas internete: https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html [žiūrėta 2025 m. balandžio 14 d.].

Carlsson Chris. 1990. *Bad attitude*. New York: Verso.

Deng Yue et al. 2023. *Multilingual jailbreak challenges in large language models*. Prieinamas internete: https://doi.org/10.48550/arxiv.2310.06474 [žiūrėta 2025 m. balandžio 16 d.].

Digit Gidget. 1982. *Sabotage: The ultimate video game*. Prieinamas internete: https://libcom.org/library/sabotage-ultimate-video-game [žiūrėta 2025 m. balandžio 18 d.].

Duffy Enda. 2009. *The speed handbook: velocity, pleasure, modernism*. Durham: Duke University Press.

Ecklar Julia. 2000. *The Kobayashi Maru*. London: Simon and Schuster.

Fanta Walter. 2015. *Krieg. Wahn. Sex. Liebe.* Berlin: Drava-Verlag.

Green Garo, Kaufman James. 2015. *Video games and creativity*. London: Elsevier.

Han Byung-Chul. 2017. *Psychopolitics: neoliberalism and new technologies of power*. London: Verso.

Jaramillo Dario. 2022. *ChatGPT-Jailbreak-Prompts*. Prieinamas internete: https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts [žiūrėta 2025 m. balandžio 16 d.].

Keturakis Saulius. 2024. Dirbtinis kvailumas ir užkalbėjimai („jailbreaks“) kaip algoritminio kūrybiškumo forma, *Information & Media* 100: 38–51.

Manovich Lev. 2025. *AI and future of identity: personal voice*. Prieinamas internete: https://www.facebook.com/lev.manovich/posts/pfbid02jTENooNjnbpwPdmNwRG1wsCpUARTwsgX2Q1Kg34ZVP7mbcbmk1n2qVWmuPF41qn6l [žiūrėta 2025 m. balandžio 14 d.].

Mayor Adrienne. 2020. *Gods and robots: Myths, machines, and ancient dreams of technology.* New York: Princeton University Pres.

Musil Robert. 1937. *Über die Dummheit*. Prieinamas internete: https://www.signaturen-magazin.de/robert-musil--ueber-die-dummheit.html [žiūrėta 2025 m. gegužės 25 d.].

OpenAI. 2024. *Safety & responsibility*. Prieinamas internete: https://openai.com/safety/ [žiūrėta 2025 m. balandžio 18 d.].

Pruthi Danish et al. 2019. *Combating adversarial misspellings with robust word recognition*. Prieinamas internete: https://arxiv.org/pdf/1905.11268 [žiūrėta 2025 m. balandžio 14 d.].

Ren Shuhuai et al. 2019. Generating natural language adversarial examples through probability weighted word saliency, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*: 1085–1097. Florence: Association for Computational Linguistics.

Rubend Nicklas. 2022. *Khajiit*. Prieinamas internete: https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts/viewer/default/train?views%5B%5D=train&row=33 [žiūrėta 2025 m. balandžio 13 d.].

Savvov Sergei. 2023. *Create a clone of yourself with a fine-tuned LLM – better programming*. Prieinamas internete: https://medium.com/better-programming/unleash-your-digital-twin-how-fine-tuning-llm-can-create-your-perfect-doppelganger-b5913e7dda2e [žiūrėta 2025 m. balandžio 15 d.].

Scott James. 1985. *Weapons of the weak: Everyday forms of peasant resistance*. Boston: Yale University Press.

Sears Aandrew, Jacko Julie. 2009. Human-Computer interaction fundamentals. Boston: Crc Press.

Shannon Claude. 1950. Programming a computer for playing chess, *Philosophical Magazine* 41(314): 256–275.

Shelley Mary. 2004. *Frankenstein: Or, the modern Prometheus*. Boston: Broadview Press.

Swift Jonathan. 2011. *Gulliver's travels*. London: Echo Library.

Szollosy Michael. 2016. Freud, Frankenstein and our fear of robots: projection in our cultural perception of technology, *AI & Society*, 32(3): 433–439.

Walsh P. 2023. Social cognition and the origin of concepts in Durkheim's sociology of knowledge, *Journal for the Theory of Social Behaviour* 54(1): 86–103.

Zhou Zhenhong et al. 2024. *How alignment and jailbreak work: Explain LLM safety through intermediate hidden states*. Prieinamas internete: https://arxiv.org/abs/2406.05644 [žiūrėta 2025 m. balandžio 17 d.].