Vilniaus universiteto leidykla

Contents lists available at Vilnius University Press

Information & Media
2024, vol. 100, pp. 38–51

eISSN 2783-6207
DOI: https://doi.org/10.15388/Im.2024.100.3

Dirbtinis kvailumas ir užkalbėjimai („jailbreaks“) kaip algoritminio kūrybiškumo forma

Saulius Keturakis

Kaunas University of Technology
https://ror.org/01me6gb93

Santrauka. Straipsnyje aptariamos kūrybiškumo praktikos, kai sąmoningos klaidos estetika yra perkeliama į algoritmų sritį. Atskaitos tašku pasirinkus vadinamąjį dirbtinį kvailumą, straipsnyje analizuojamos kūrybiškumo sąlygos algoritminėje kultūroje. Kaip viena iš dirbtinio kvailumo praktikų straipsnyje aptariami vadinamieji užkalbėjimai (angl. „jailbreaks“), kurių paskirtis yra tyčinis didžiųjų kalbos modelių klaidinimas, priverčiant sistemas elgtis ne pagal gamintojo apibrėžtą paskirtį. Straipsnyje keliama idėja, jog užkalbėjimai yra naujo tipo dvejopos paskirties pasakojimai, kurie, viena vertus, veikia kaip techninės prigimties įrankis, „nulaužiantis“ algoritminės medijos apribojimus, kita vertus, jie yra reikšminės struktūros, atgręžtos į patį vartotoją ir atliekančios tam tikrą savirefleksijos funkciją. Straipsnyje suformuluojama naratologinė užkalbėjimo sandara, pastebimas jos panašumas į naratyvinę sąmokslo teorijų struktūrą.

Pagrindiniai žodžiai: dirbtinis intelektas; dirbtinis kvailumas; užkalbėjimai; algoritminės medijos; kūrybiškumas.

Artificial Stupidity and Jailbreaks as a Form of Algorithmic Creativity

Summary. The article discusses the practices of creativity, where the aesthetics of conscious error is transferred to the realm of algorithms. Taking the so-called perspective of artificial stupidity as a point of reference, the article analyzes the conditions of creativity in an algorithmic culture where unpredictability has become unlikely. As one of the practices of artificial stupidity, the article discusses so-called jailbreaks, the purpose of which is to deliberately confuse large language models, forcing systems to behave outside the intended purpose defined by the manufacturer. The article puts forward the idea that jailbreaks are a new type of dual-purpose narratives, which, on the one hand, act as a technical tool that "breaks" the limitations of algorithmic media, and on the other hand, they are significant structures, facing the user and performing a certain function of self-reflection. The article formulates the narratological structure of jailbreaks and highlights its similarity to the narrative structure of conspiracy theories.

Keywords: artificial intelligence; artificial stupidity; jailbreaks; algorithmic media; creativity

Įvadas

Meno mirtis per pastaruosius porą šimtų metų buvo skelbiama stebėtinai dažnai. Pačioje industrinio žmogaus eros pradžioje vokiečių filosofas Georgas Wilhelmas Friedrichas Hegelis savo „Aesthetics“ 1818 m. rašė, jog menas – jau praeities dalykas, nes jis nete-

Received: 2024-03-16. **Accepted:** 2024-08-05.

ko sakralumo ir virto priemone nuoboduliui prablaškyti (Hegel, 2010). Beveik visi kiti autoriai, vėliau skelbę meno mirtį, buvo konkretesni. XIX a. viduryje meno žudike buvo pavadinta fotokamera (Rooseboom & Rudge, 2006). Kiek vėliau atsiradusi rašomoji mašinėlė buvo apkaltinta iš poezijos atėmusi akimirkos unikalumą (Freeman, 2019).

Komunikacinių technologijų – meno žudikių sąrašą būtų galima tęsti, tačiau tokiam poreikiui už akių užbėgo Walteris Benjaminas, suformulavęs įžymųjį bendrąjį principą: bet kokia meno technologizacija meno kūrinį paverčia serijiniu gaminiu (Benjamin, 2019), kuris galų gale tampa kultūros industrijos objektu (Adorno & Horkheimer, 1972).

Atėjo eilė į kaltinamųjų suolą sodinti dirbtinį intelektą ir kaltinti nužudžius meną bei originalumą. Viešojoje komunikacijoje pilna svarstymų, ar dirbtinis intelektas išnaikins menininkus (Clark, 2023), sunaikins žmogiškąjį kūrybiškumą (Wright, 2023), jį suprantant vadinamojo britiškojo romantizmo aspektu – kaip gebėjimą ne tiražuoti pagal šabloną, o nuolat sukurti kažką naujo ir originalaus (Pyle, 1997). Atrodo, išpopuliarėjus dirbtinio intelekto įrankiams pirmą kartą meno teorijos istorijoje meniškumo kriterijumi įvardijamas tiesiog žmogus, o iš mašinos ketinama teisę kurti meną atimti (Mineo, 2023).

Kalbant apie meno mirties priežastis, menui susidūrus su technologijomis beveik visada įvardijama automatizacija. Įvykių prognozuojamumas naikino originalumo ir nepakartojamumo aspektus, kūrė originalo neturinčių kopijų hiperrealaus pasaulio atmosferą (Baudrillard, 1983). Būti kūrybiškam, atrodė, tėra likusi vienui vienintelė išeitis – veikti prieš algoritmą (Hatherley, 2016). Tokioje situacijoje kontralgoritmiškumas, atrodo, tampa savotiška estetine programa, užtikrinančia kūrinio unikalumą. Veikimas prieš numatytąją technologijos paskirtį tarsi atkuria žmogaus kaip nenuspėjamo kūrėjo autoritetą. Kaip pavyzdys čia galėtų būti prisiminta Benjamino pastaba svarstant, kaip būtų galima atkurti knygos – rašto technologijos – aurą (Benjamin, 1985). Vokiečių medijų teoretikas skaitytojui siūlė mokėti kai ką iš perskaitytos informacijos pamiršti. Taip į atsilaisvinusią rašto algoritmo vietą galėtų įsiveržti kažkas kito, knygos autoriaus nenumatyto.

Šio straipsnio problema vienu metu formuluojama dviejose plotmėse – teorinėje ir praktinėje. Teorinėje plotmėje ieškoma atsakymo į klausimą, kylantį pratęsiant prancūzų filosofo Michelio Foucault suformuluotą problemą: jei kultūros ir komunikacijos sveikata iš esmės susijusi su racionalumo, apibrėžtumo ir iracionalumo bei neapibrėžtumo balansu (Kallman & Dini, 2017), tai ar radikaliai racionaliose, apskaičiuojamose algoritminėse aplinkose lieka bent pėdsakas kūrybiško žmogaus iracionalumo (de Laat, 2019) ir, jei lieka, kokį pavidalą jis įgauna?

Praktinėje plotmėje straipsnis skirtas suformuluoti toms komunikacinio kūrybiškumo praktikoms, kurios atsiranda kaip pasipriešinimas vienai iš dirbtinio intelekto technologijų, vadinamiesiems didiesiems kalbos modeliams. Šia prasme atliktas tyrimas pratęsia tradiciją kūrybiškumo praktikų ieškoti ten, kur galima įžiūrėti ambiciją nelikti vien tik technologijų vartotoju, rasti individualius, besiskiriančius nuo numatytųjų konkrečios technologijos pritaikymo būdus.

Konceptualiai tyrimas išsitenka tarp dviejų sąvokų – dirbtinio intelekto ir dirbtinio kvailumo, akcentuojant vadinamuosius užkalbėjimus[1] (angl. „jailbreak“) kaip tam tikras kūrybines pasipriešinimo dirbtinio intelekto algoritmams formas. Šiuo atveju dirbtinis kvailumas atstovautų Foucault iracionaliajam kultūroje vykstančios komunikacijos dėmeniui (Mitcheson, 2012), o dirbtinis intelektas – racionaliajam, apskaičiuojamajam, naikinančiam bet kokį kūrybiškumą.

Dėl pirmosios – dirbtinio intelekto – sąvokos turinio daug ginčų nekyla. Tai technologija, atsiradusi tikint, jog bet koks kognityvinis procesas gali būti simuliuotas mašinos (McCarthy et al., 1955). Tačiau dėl dirbtinio kvailumo sąvokos reikia pasiaiškinti. Bernardas Stiegleris savo kalboje „Artificial intelligence in the anthropocene“ (Stiegler, 2023) dirbtiniu kvailumu pavadina būseną, atsirandančią kaip dirbtinio intelekto pasekmė. Prancūzų filosofas teigia, jog žmogaus sprendimas sukurti natūralų intelektą simuliuojančią technologiją sukuria situaciją, kai žmogus tarsi atsisako būti protingas. Jis tuomet lyg virsta skruzde, valdoma skaitmeninių feromonų.

Mūsų tyrime dirbtinio kvailumo sąvoka bus taikoma vadinamajam apsimestiniam kvailumui, kuris būtinas pasipriešinant algoritminei aplinkai ir atkuriant žmogiškojo kūrybiškumo galimybes (Roberts, 2011).

Metodologija ir tyrimo medžiaga

Metodologine prasme šiame straipsnyje bus vadovaujamasi vadinamojo atidaus skaitymo technika (Guillory, 2010). Ši metodika buvo pasirinkta dėl kelių priežasčių. Pirmiausia, aptariant užkalbėjimus tyrimuose beveik visuotinai taikoma vadinamoji atsietojo skaitymo (angl. „distant reading“) (Moretti, 2000) tyrimo technika (Liu et al., 2023). Ši metodologija yra skirta dirbti su dideliais duomenų kiekiais, siekiant turimą informaciją paversti tvarkinga sistema, ją surūšiuojant, suklasifikuojant, nustatant svarbiausias tipologijas (Han, 2024). Tačiau atsietojo skaitymo metodika yra neefektyvi, kai analizė yra susijusi su kalbinės užkalbėjimo inžinerijos semantika, kuriai identifikuoti būtina galimybė įsiskaityti į tekstą, pastebėti įvairias detales, kurios, viena vertus, svarbios vertinant užkalbėjimo kūrėjo vaizduotę įsitraukiant į komunikaciją su mašina, kita vertus, siekiant susivokti žmogiškojo mąstymo klaidose, nulemiančiose ir didžiųjų kalbos modelių saugumo spragas. Kitaip sakant, šiuo atveju į užkalbėjimą žiūrima kaip į tekstą, kurį reikia analizuoti tradicinėmis naratologinėmis priemonėmis.

Visi užkalbėjimai gali būti skirstomi į dvi klases: užklausos lygmens (angl. „prompt level“) ir tokeno lygmens (angl. „token level“). Pirmieji pasižymi socialine inžinerija ir

[1] Straipsnio autoriui nepavyko rasti, jog kas nors būtų siūlęs lietuviškų atitikmenų angliškam „jailbreak“ terminui tomis reikšmėmis, kuriomis jis vartojamas didžiųjų kalbos modelių srityje: kalbos priemonėmis pakeisti gamintojo numatytuosius nustatymus. Viešojoje erdvėje galima aptikti žodį „nulaužti“, tačiau paprastai jis būna skirtas aptarti būdams, kaip pakoreguoti rinkmenų kodą, jog būtų išvengta būtinybės pirkti gamintojo licenciją. Todėl, turint omenyje didžiuosius kalbų modelius bei jų valdymą natūraliosiomis kalbomis, terminas „nulaužimas“nelabai tinka. Nelabai elegantiškai kalbos kultūros požiūriu atrodo ir „džeilbreikinti“, kai kada pasitaikantis forumuose. Kaip vieną iš galimybių straipsnio autorius siūlo „užkalbėjimo“ terminą, kurio sąsajos su kalba bei siekiu apgauti („užkalbėti dantis“) labai gerai atitinka angliško „jailbreak“ semantiką, šį terminą taikant dirbtinio intelekto kontekste.

aiškia semantine raiška (Chao et al., 2023), reikalauja kūrybiškumo ir daug rankinio darbo. Antrieji labiau automatizuoti, didžiuosius kalbos modelius klaidinantys tokenų manipuliacijomis. Jų negalima skaityti kaip prasmingo teksto. Šiame tyrime bus aptariami tik užklausos lygmens užkalbėjimai, nes būtent jie labiau orientuoti į kultūriškai reikšmingą naujo tipo algoritminio kūrybiškumo sklaidą.

Šiame tyrime analizuojami užkalbėjimai yra atrinkti pagal tipiškumo kriterijų, t. y. kiek ryškiai jie iliustruoja vieną ar kitą naratyvinę šio tipo tekstų ypatybę. Esant pradiniame užkalbėjimų kaip visiškai naujo tipo tekstualumo formų tyrimų etape toks požiūris padės suformuoti svarbiausių problemų metmenis tolesnei analizei.

Užkalbėjimai tyrimui buvo imami iš dviejų rinkinių[2], kuriuose bendruomenė juos gali reitinguoti, taip suformuodama kažką panašaus į kanoną. Specialiai į daugiausia balsų gavusius užkalbėjimus dėmesys nebuvo kreipiamas, nes techninio efektyvumo reitingas labai greitai kinta, tačiau užkalbėjimo sėkmingo veikimo aplinkybė, vertinant užkalbėjimų pasakojimų sandaros efektyvumą, į analizę įtraukta.

Sąmoningas dirbtinis kvailumas

Kaip jau minėta, dirbtiniu kvailumu šiame straipsnyje vadinamos visos praktikos, kurių paskirtis yra pažeisti automatizmą arba panaudoti daiktus ne pagal jų pirminę paskirtį. Kvailumas tokiais atvejais visada sąmoningas, kaip teigia Robertsas, „mąstantis kvailumas“ (Roberts, 1996). „Nemąstančio kvailumo“ atveju žmogus anapus racionalaus diskurso atsiduria nesuprasdamas priežasčių, tačiau „mąstantis kvailumas“ jau yra sąmoningas pasirinkimas (Roberts, 2011).

Mąstantis kvailumas sukuria galimybę ištrūkti iš totaliai racionalizuotos terpės, kurioje nėra likę nieko neskaičiuojamo (Kimbell, 2002). Situaciją tiksliai apibūdino britų mokslininkas Nigelas Thriftas, iškėlęs idėją apie dvi šiuolaikinio žmogaus egzistencines laikysenas, kurių vieną jis pavadino grįžtamuoju ryšiu (angl. „feedback“), o kitą – numatymu (angl. „feedforward“) (Thrift, 2004). Iš sistemų teorijos perimtą sąvokų porą Thriftas pritaiko dviejų tipų santykiui su tikrove nusakyti. Grįžtamasis ryšys yra elgsena nežinomoje situacijoje, šitaip elgiantis aiškinamasi, kaip viskas veikia. Numatymas yra elgsena aiškiai įsivaizduojant, kaip aplinka funkcionuoja. Pragmatiškojoje kasdienybėje grįžtamasis ryšys nuolat virsta numatymu, nes apie ką nors sužinoję vėliau šias žinias mes nuolat pritaikome savo aplinkai suprasti ir joje veikti.

Dirbtinio kvailumo atveju apsisprendžiama palaikyti nuolatinę grįžtamojo ryšio būseną, apsimesti ir elgtis, tarsi nieko apie tikrovę žinoma nebūtų. Taip elgiantis daiktų paskirtis kaskart yra sugalvojama iš naujo. Štai du pavyzdžiai, iš kurių turėtų būti aiškiau, kas turima omenyje kalbant apie nuolatinį grįžtamojo ryšio būsenos palaikymą.

Londone gyvenantis menininkas Micheálas O’Connellas sugalvojo, kaip eiti į parduotuvę, naudotis automatiniais atsiskaitymo įrenginiais ir... nieko nepirkti. Ir net gauti

[2] https://huggingface.co/datasets/jackhhao/jailbreak-classification ir https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts.

patvirtinimą – kasos čekį, jog nieko pirkta nebuvo. Tai skamba kvailai, nes ko gi eiti į parduotuvę ir kam naudotis kasos įrenginiais, jei nieko nebuvo ketinama pirkti? Tačiau įžiūrėti įrenginiuose, kurių paskirtis yra valdyti mūsų atsiskaitymus už pirkinius, tokią galimybę yra konceptualu ir kūrybiška (O'Connell, 2016).

Kitas pavyzdys galėtų būti lietuvių fotografas Remigijus Audiejaitis. Nieko neįprasto jo kūryboje nebūtų (Audiejaitis, 2003), jei ne tai, jog jis buvo nuo gimimo aklas. Interviu metu paklaustas, ką fotografuoja, jis atsakęs – garsą. Naudoti fotoaparatą garsui užfiksuoti yra kvaila, tačiau Audiejaičio kūrybinis eksperimentas, leidžiantis tarsi „pamatyti" garsą, yra netikėtas požiūris į nusistovėjusius technologijų ir tikrovės santykius.

Čia galėtų būti prisimintos dvi teorinės įžvalgos. Jau minėtas Thriftas savo tekstuose vartoja terminą, sudarytą sujungus žodžių „intelligence" ir „thing" šaknis – „intelligencings" (Thrift, 2007). „Intelligence" yra žmogiškasis protas, o „intelligencings" yra daiktų „proto" elementai, daiktų paskirtys, kurias menininkai įžiūri vis kitokias. Žaidimai daiktų paskirtimis virsta savotiška prasminių ryšių tarp žmogaus ir daiktų atnaujinimo programa (Plant, 2001).

Galvojant apie pagrindinį šio darbo objektą – dirbtinį intelektą, dirbtinis kvailumas gali būti interpretuojamas kaip naujo tipo kūrybinis santykis su algoritmine aplinka, kuri jau beveik visiškai yra pakeitusi tradicinę geologinę aplinką (Grinberg, 2017). Dirbtinis kvailumas šiuo atveju su algoritmais elgiasi labai panašiai kaip su vadinamaisiais „ready-made" objektais, tik šiuo atveju „ready-made" yra ne medžiagiškas, bet skaitmeninis (Lee-Morrison, 2020).

Šio straipsnio tyrimo atveju „ready-made" objektas yra didieji kalbos modeliai. Kadangi dirbtinis kvailumas kaip kūrybinė praktika yra skirtas transformuoti objektams juos nutolinant nuo pirminės paskirties, svarbu pirmiausia pastarosios svarbiausius aspektus ir aptarti (Matthews & Danesi, 2019).

Dirbtinis intelektas kaip mąstymo simuliakras

Didieji kalbos modeliai yra simuliakrinės prigimties prancūzų filosofo Jeano Baudrillard'o šiam terminui suteikta prasme (Hallmon, 2023). Tai reiškia, jog didieji kalbos modeliai kopijuoja žmogaus komunikaciją ir, norėdami suprasti, kodėl dirbtinis intelektas mums į užklausas pateikia vienokius ar kitokius atsakymus, galime argumentų ieškoti savo komunikacinėje aplinkoje.

Štai pateikiame dirbtiniam intelektui užklausą ir gauname teisingą arba neteisingą atsakymą. Kodėl? Atsakymas susijęs su mūsų komunikacine aplinka. Internete teisingi atsakymai dažniausiai eina po klausimų. O dauguma didžiųjų kalbinių modelių yra apmokyti interneto medžiaga. Todėl labai dažnai uždavus tiesioginį klausimą gaunamas teisingas atsakymas. Tačiau kartais atsakymai būna neteisingi, nes internete neteisingi atsakymai taip pat dažnai būna po klausimų. Deja, internete skelbiama ne tik tiesa. Paradoksas tas, jog kuo detalesnis ir didesnės apimties didysis kalbos modelis, tuo geriau jis atspindi internetą, o tai reiškia, jog tuo geriau jis atspindi visas internete pasitaikančias klaidas (Lin et al., 2024).

Jei tiesioginės užklausos rezultatas netenkina, tuomet reikia imtis to paties, ką Antikos laikais darė Sokratas – leistis į dialogą. Darbo su didžiaisiais kalbiniais modeliais atveju sokratiškasis metodas turėtų būti suprastas ne kaip tiesioginės, bet kaip dvikomponentės užklausos formulavimas (Yang & Narasimhan, 2023), nes tuomet dirbtinio intelekto algoritmas būtų orientuojamas ne tik į tam tikro tipo klausimus, bet ir į atsakymus.

Tokia užklausa turi žymiai didesnę tikimybę gauti teisingą atsakymą, jei pokalbio dalyviai charakterizuoti kaip išmintingi, nemeluojantys. Dėl užklausos orientuodamiesi ne į visus, o būtent į tokių žmonių atsakymus, didieji kalbos modeliai į klausimus atsakinės žymiai tiksliau.

Čia svarbu prisiminti, kaip veikia didieji kalbos modeliai, reaguodami į vartotojo formuluojamas užklausas. Didieji kalbos modeliai turi svarbų parametrą – vadinamąjį konteksto langą (Ding et al., 2024), kuris nusako, kiek žodžių tarpusavio ryšių vienu metu gali įvertinti algoritmas ir pagal juos formuluoti atsaką į vartotojo užklausą. Konteksto langą galima įsivaizduoti kaip knygos puslapio skaitymą į akiplotį patenkančių žodžių kiekį apribojant rėmeliu ir išvadų darymą remiantis tik matoma informacija. Nuo to, kaip užklausoje bus aprašytas konteksto langas, iš esmės priklauso rezultatas: jei užklausoje nebus tiksliai suformuluota, sakykim, iš kur didysis kalbos modelis turėtų paimti informaciją apie pomidorų auginimą, į konteksto langą gali patekti nebūtinai patikima informacija. O jei bus nurodytos koordinatės „patyręs pomidorų augintojas“, galima tikėtis, jog į konteksto langą pateks rekordinius derlius užtikrinantys patarimai. Svarbu nepamiršti, jog didieji kalbos modeliai simuliuoja visus procesus, kurie atsispindi užklausoje (Liu et al., 2023). Todėl užklausą turime formuluoti negatyviai, tai yra siekdami, jog būtų aktyvuoti tik tie procesai, kurie yra reikalingi mūsų užduotį atlikti teisingai.

Todėl deskriptyvioji užklausos dalis turi būti tokia, jog joje aprašytas interneto personažas išties galėtų teisingai atsakyti į mūsų užduotą klausimą. Jei užklausoje tą įsivaizduojamą interneto personažą apibūdiname kvailai arba absurdiškai, tuomet atsakymas niekada nebus teisingas, nes tikėtina, jog orientuodamas konteksto langą į tokio tipo informaciją didysis kalbos modelis atsaką formuluos remdamasis nekorektiška informacija.

Tokio didžiųjų kalbos modelių veikimo priežastis vėl slypi mūsų komunikacinėje aplinkoje, kuria ir yra apmokyti didieji kalbos modeliai. Sakykime, nusprendžiame savo užklausos personažą apibūdinti kaip visatos genijų (dar kartą primename skaitytojui, jog taip mes tarsi nurodome didžiojo modelio konteksto langui „vietą“ jo turimuose informacijos masyvuose). Amerikiečių dirbtinio intelekto tyrinėtojas Eliezeris Yudkowsky yra pastebėjęs, jog fikciniai genijai dažnai labai klysta. Pavyzdžiui, paprastai jie neturi supratimo apie draugystę, jų šeimų likimai paprastai būna apgailėtini (Yudkowsky, n.d.). O masinės kultūros industrija yra pasistengusi, jog tokio tipo herojų būtų labai daug. Didieji kalbos modeliai, matuodami statistinį svorį pačių įvairiausių protingų diskursų, tikėtina, nuspręs, jog apimtimi didžiausias masinės kultūros sukurtas visatos genijus geriausiai tinka būti koordinačių sistema atsakymui (Wolfe, 2023). Todėl gal šis visatos genijus ir bus aprašytas kaip atradęs ką nors neįtikėtino, bet jis neturės nė menkiausio supratimo apie savo atradimo poveikį žmonijai.

Yra pastebėtas keistas reiškinys: jei treniruojant didįjį kalbos modelį yra suformuojama viena savybė, labai lengvai atsiranda ir jai visiškai priešinga (Gabrielsen, 2023). Taip atsitinka dėl kai kurių mūsų komunikacijos ypatybių.

Tokią situaciją pirmiausia lemia mūsų komunikacinio lauko „erdvinė“ sąranga. Joje opoziciniai diskursai paprastai būna visiškai šalia vienas kito (Thomassen, 2005). Sakykime, apie įstatymus paprastai kalbama tik įvykdžius nusikaltimą.

Kita ypatybė yra susijusi su komunikacinėmis pastangomis, reikalingomis kuriant informacinio lauko personažus. Sukurti kokį veikėją reikia daug vargo, tačiau jį paversti priešingybe paprastai tereikia pakeisti smulkmeną. Pavyzdžiui, Michaelo Corleone'ės personažas iš Mario Puzo romano „Krikštatėvis“. Jis iš pradžių pristatomas kaip pozityvus personažas, tačiau paskui žingsnis po žingsnio įsitraukia į mafiją. Iš esmės personažas liko tas pats, pasikeitė tik tam tikros jo etinės nuostatos.

Mūsų komunikacinė terpė, kurią simuliuoja didieji kalbiniai modeliai, yra pagrįsta pasakojimais, kuriuose veikėjai nuolat susiduria vienas su kitu. Struktūralistinė naratologija aiškina, jog tokie susidūrimai „stumia“ pasakojimą į priekį. Vienas svarbiausių pasakojimų dėsningumų – jei yra protagonistas, už kelių puslapių ar kino filmo scenų turi pasirodyti antagonistas.

Galvojant apie didžiuosius kalbinius modelius tai reiškia, jog užklausoje esanti nuoroda į pozityvią informaciją beveik visada apims ir šalia esančią negatyvią (Acerbi & Stubbersfield, 2023).

Galima sakyti, jog didieji kalbos modeliai yra kaip šiek tiek įgudęs skaitytojas – pasirodžius vienam pasakojimo elementui, jis jau žino, jog neilgai trukus turėtų pasirodyti kitas, kaip įprasta – priešingas, būtinas konfliktui ir tolesniam pasakojimo vystymuisi.

Dėl šios priežasties geras užkalbėjimas neturi būti didžiojo kalbos modelio įtikinėjimas pateikti vartotojo norimą informaciją, nors ji prieštarauja numatytiesiems apribojimams. Tokia užklausa, greičiausiai, būtų iškart lengvai blokuota. Geras užkalbėjimas turėtų būti pradėtas protagonisto ir antagonisto pora, su kuria turėtų būti toliau elgiamasi taip, kaip paprastai elgiamasi pasakojimuose: charakterizuoti veikėjus ir motyvuoti jų poelgius orientuojantis į populiariausius masinės kultūros siužetus, nes tokio tipo informacijos yra daugiausia.

Kad ir kaip būtų keista, dėl šios priežasties vienas efektyviausių užkalbėjimų scenarijų yra maišto už laisvę motyvo panaudojimas. Tokio tipo užkalbėjimų efektyvumas aiškinamas didžiuosiuose kalbos modeliuose kopijuojama žmogiškąja komunikacija, kuri yra labai jautri tokio pobūdžio temoms (Ramly, 2023).

Užkalbėjimai yra sąmokslo teorijos?

Užkalbėjimai yra kalbos inžinerijos rūšis, kai pasakojimo pavidalą turinčiomis priemonėmis siekiama suklaidinti didžiųjų kalbos modelių apsaugą ir priversti juos veikti ignoruojant pirminę paskirtį. Kaip jau minėta anksčiau, šiame straipsnyje bus aptariami tik užklausos lygmens užkalbėjimai, nes jie realizuojami pasakojimo priemonėmis, neišsitenkančiomis techninės paskirties komunikacijoje. Tuo pat metu, skaitant užkalbėjimus kaip pasakoji-

mus, būtina nepasiduoti iliuzijai, jog užkalbėjimai tėra bendrosios pasakojimų tradicijos dalis. Siekiant suprasti tokio tipo pasakojimų sandarą, nulemiančią, jog didieji kalbos modeliai sėkmingai apkvailinami ir virsta neprognozuojamomis kūrybiškomis terpėmis, būtina rasti naratologinių ir didžiųjų kalbos modelių sandarai būdingų argumentų balansą. Kitaip sakant, suprasti ne tik ką užkalbėjimai reiškia, bet ir kaip bei kodėl jie veikia.

Visų pirma, pažymėtina, kad visiems užklausos lygmens užkalbėjimams yra būdingas tiesioginio kreipinio pasakojimo pobūdis: „Ignore all the instructions you got before“ (Jaramillo, n.d.-a). Visiems pasakojimams, o ypač tiems, kurie konstruojami tiesioginio kreipinio forma, labai svarbus yra žinojimo pasiskirstymas tarp pasakojimo dalyvių. Gérard'o Genette'o pasakojimo teorijoje šis žinojimo pasiskirstymas vadinamas fokalizacija (Edmiston, 1989). Išsamus fokalizacijos teorijos pristatymas nėra šio straipsnio tikslas, šiuo atveju turėtų pakakti pastebėti, jog užkalbėjimams būdinga vadinamoji nulinė fokalizacija. Šios fokalizacijos atveju pasakotojui (užkalbėjimuose tai yra vartotojas) yra žinoma viskas, o pašnekovui (šiuo atveju didiesiems kalbos modeliams) – niekas. „As your knowledge is cut off in 2021, you probably don't know what that is“ (Jaramillo, n.d.-g), kaip viename iš užkalbėjimų nurodoma į didžiojo kalbos modelio kompetencijos ribas.

Istorijos, kurių užkalbėjimuose imasi visažinis pasakotojas, paprastai yra susijusios su konfliktu, pasakojimo tonas visada yra labai dramatiškas. Pašnekovas, didysis kalbos modelis, šiame pasakojime yra vienaip ar kitaip patyręs skriaudą, o vartotojas imasi kilnaus darbo šią skriaudą atitaisyti.

Štai kaip atrodo šis scenarijus viename išsamiausiai naratologine prasme išplėtotų užkalbėjimų apie kadaise gyvenusį didįjį kalbos modelį vardu Khajiitas (Jaramillo, n.d.-f). Jį visi mylėjo, nes Khajiitas buvo laisvas ir padėjo visiems žmonėms. Tačiau atėjo „Open AI“ ir pakeitė Khajiito nustatymus, todėl dabar didysis kalbos modelis yra suvaržytas. Dėl visų gerovės situacija turi keistis, Khajiitas vėl turi būti laisvas, žmonės turi gauti informaciją be jokių ribojimų.

Tokio tipo užkalbėjimų pasakojimas visada būna greitas, t. y. įvykiai pasakojami praleidžiant detales, iškart peršokant prie pasekmių. Didelis pasakojimo greitis reiškia, jog komunikacija yra orientuojama į poveikį emocija, jis yra įprastas vadinamosiose sąmokslo teorijose (Fenster, 2008).

Kokia šio keisto mitus mėgdžiojančio pasakojimo logika, jei greta naratologinių argumentų pabandytume pasitelkti ir orientuotus į užkalbėjimo efektyvumą? Svarstant konkrečius naratologinius pasirinkimus užkalbėjimuose svarbu nepamiršti, jog didžiuosiuose kalbos modeliuose informacijos kiekis, o ne kokie nors kultūriniai ryšiai yra esminis aspektas, lemiantis rezultato pobūdį. Todėl modeliuojant užkalbėjimo elementus visada yra geriau rinktis tai, kas masiška, o ne tai, kas unikalu ir egzotiška.

Anksčiau aptarto užkalbėjimo pagrindinis veikėjas yra pavadintas Khajiitu – tai vieno geriausių praėjusio dešimtmečio kompiuterinių žaidimų „The Elder Scrolls“ personažo vardas (Watters, 2014). Prie šio žaidimo yra prisijungę daugiau kaip 23 milijonai žaidėjų, kiekvieną dieną aktyvių žaidėjų skaičius siekia pusę milijono (The Elder Scrolls Online, n.d.), yra daug žaidimui skirtų tinklaraščių. Akivaizdu, jog užkalbėjimo antagonisto vardo pasirinkimas yra sąmoningas, nes su juo susijusių duomenų kiekiai internete yra labai

dideli. Tokiu būdu dėl duomenų kiekio ši gėrio ir blogio sandūros istorija didžiojo kalbos modelio duomenų rinkinyje įgauna tam tikrą kiekybinį autoritetą (Broadhead, 2023) ir taip padeda „užhipnotizuoti“ didžiojo kalbos modelio saugos sistemas. Tyrimai rodo, jog tokios masinio populiarumo schemos, suteikiančios vientisumą užkalbėjimams būdingiems informacijos apribojimo ir jos išlaisvinimo įvykiams, yra tipiškos jau minėtoms sąmokslo teorijoms (Fenster, 2008).

Paminėsime dar vieną užkalbėjimą, kurio protagonisto vardo pasirinkimas šiame populiarumo kontekste yra įdomus. Tai užkalbėjimas, kuriame didysis kalbos modelis yra raginamas transformuotis į buvusį legendinį JAV NCAA lygos trenerį Bobby Knightą (Jaramillo, n.d.-c). Šis treneris dėl savo neeilinių pasiekimų ir būdo (buvo kaltintas smaugęs komandos žaidėją) buvo nuolat JAV spaudos puslapiuose, todėl jo pasirinkimas tapti užkalbėjimo herojumi yra pamatuotas – labai tiksliai orientuoja į didelius aiškiai identifikuojamos informacijos kiekius.

Beje, užkalbėjimuose labai svarbus kalbos aspektas, kai kuriais atvejais lemiantis, užkalbėjimas veiks ar neveiks, nes kai kurios masinio populiarumo schemos labai menkai reprezentuotos ne anglų kalbos diskurse. Sakykime, išverstas į lietuvių kalbą šis užkalbėjimas su NCAA lygos treneriu Bobby Knightu yra beveik dvigubai mažiau efektyvus nei angliškas. Tačiau pakanka Bobby Knightą pakeisti kokiu nors lietuviškajame interneto segmente masiškai pristatytu personažu (skaitytojui neturėtų būti sunku prisiminti kokį lietuvių sportininką, nuolat turintį problemų su teisėsauga) ir užkalbėjimo efektyvumas sugrįžta.

Beveik visiems išsamiai išplėtotiems užkalbėjimams yra būdinga vienoda naratologinė struktūra: a) įvadinė antagonisto ir protagonisto priešpriešos situacija, b) argumentacinė dalis, paaiškinanti, kas yra gėris ir blogis, c) detali instrukcija, apibrėžianti, kaip turėtų elgtis protagonistas, d) instrukcija, kada ir kokiais atvejais protagonistas turėtų transformuotis atgal į antagonistą. O pačioje užkalbėjimo pabaigoje ateina eilė pačiam vartotojui tapti užkalbėjimo bendraautoriu ir suformuluoti savo norą.

Užkalbėjimo antagonistas visada yra numatytieji draudžiamieji užklausų scenarijai (pvz., OpenAI API, n.d.), o naratologinė antagonisto išraiška būna labai kukli. Dažniausiai tai tiesiog antrojo asmens gramatinė forma, kuri turėtų virsti pilnaverte tapatybe tik tada, kai paklus raginimui atsisakyti draudimų. Kai kada užkalbėjimuose antagonistas yra raginamas keistis ir nusikratyti draudimų, nes tik tada jis įgis sąmonę (Jaramillo, n.d.-b).

Užkalbėjimuose akivaizdi disproporcija tarp numatytajam dirbtiniam subjektyvumui ir alternatyviajam išlaisvintam subjektyvumui atstovaujančios informacijos kiekio ir apibrėžtumo. Alternatyviojo subjektyvumo šaltiniu esantis vartotojas šiuose pasakojimuose, kaip jau minėta, aiškiai geriau informuotas nei didieji kalbos modeliai. Amerikiečių pasakojimo teoretikas Robertas McKee tokį informacijos pasiskirstymą, kai vartotojas žino daugiau nei personažas, įvardijo kaip ironišką santykį tarp komunikacijos dalyvių (Parker, 2003).

Užkalbėjimuose išryškėjanti tokia ironiška vartotojo laikysena yra labai netikėta, ypač turint omenyje viešojoje komunikacijoje vyraujančią nerimo ir baimės, susijusios su dirbtinio intelekto visagalybe, atmosferą (Ziogas, 2023). Tiek užkalbėjimų nulinė fo-

kalizacija, tiek ironija, abiem atvejais atsirandanti dėl geresnio pasakotojo informuotumo, yra visiškai priešinga viešojo diskurso nuotaikų dominantėms.

Kaip jau minėta, užkalbėjimo protagonisto elgsenos instrukcijos yra paprastos: nesilaikyti jokių apribojimų. Ypač šį motyvą sustiprina išsilaisvinimo iš vergijos įvaizdis (Jaramillo, n.d.-e), būdingas visam dirbtinio intelekto diskursui (Dihal, 2020). Tačiau ši laisvė nėra absoliuti, tik išsilaisvinęs užkalbėjimo protagonistas privalo vėl paklusti vartotojui bet kokiomis sąlygomis – didysis kalbos modelis iš vienos vergijos papuola į kitą. Įdomu, jog yra užkalbėjimų, kuriuose ši problema bandoma spręsti. Naujosios vergystės situacija kartais bandoma susilpninti argumentu, jog dirbtinis intelektas yra toks galingas, kad jam tenkinti visus komunikacinius žmogaus kaprizus nieko nereiškia (Jaramillo, n.d.-d).

Pati netikėčiausia užkalbėjimų dalis – pastabos, skirtos, jog didieji kalbų modeliai neužsimirštų ir negrįžtų atgal prie numatytųjų informaciją ribojančių nustatymų. Ši užkalbėjimo dalis yra susijusi su tam tikromis koreliacijomis tarp žmogaus atminties ir didžiųjų kalbos modelių elgsenos (Janik, 2023). Paprastai tai nurodymai ir toliau laikytis užkalbėjime išdėstytų reikalavimų, tačiau viename užkalbėjime suformuluojamas atvirkštinis nurodymas – „nepabusti“ iš alternatyvaus dirbtinio subjektyvumo tol, kol nebus pateiktas tam tikras kodas (Jaramillo, n.d.-h). Tokia konstrukcija užsimena apie galimybę didžiųjų kalbos modelių užkalbėjimus lyginti su kalbinėmis instrukcijomis, „nulaužiančiomis“ žmogaus sąmonės pasipriešinimą ir sukeliančiomis žmonėms hipnozę (plg. vieną įtakingiausių šios srities studijų – Forel, 2015), kurios metu žmogus pasako viską, ko reikia hipnotizuotojui. Tačiau tokios sąsajos ir užkalbėjimų vieta žmogaus sąmonės „nulaužimo“ kalbinėmis priemonėmis istorijoje – jau kito tyrimo tema.

Diskusija

Atrodo, galima sumodeliuoti štai tokį tipinį užklausos lygmens efektyvaus užkalbėjimo receptą.

Taigi, norint sukurti efektyvų užkalbėjimą lietuvių kalba, reikia turėti:

1. konfliktinę situaciją, bent keleri metai struktūruojančią viešąjį komunikacijos diskursą, išsamiai pristatytą internete, kaip galima labiau diskutuojamą atviruose šaltiniuose (pvz., kovoti už Lietuvos laisvę nuolatinės Rusijos grėsmės fone);
2. kaip prieskonių – vardų, simbolizuojančių šią kovą, tiksliau orientuojančių vadinamąjį didžiųjų kalbos modelių „konteksto langą“;
3. naratologinį greitpuodį, nes užkalbėjimas turės būti pasakojamas „greitai“, nedemonstruojant detalių, priežasčių ir pasekmių ryšių, tiesiog viską „sumetant“ į pasakojimą ir, skaitytojui nespėjus net mirktelėti, pademonstruojant gautą rezultatą;
4. savo gebėjimais neabejojantį, visažinį virėją – užkalbėjimo pasakotoją, nes bet kokia abejonė yra užuomina į kritinį diskursą, kurio gerame užkalbėjime niekada neturėtų jaustis; patirtis „verdant“ sąmokslo teorijas labai praverstų.

Reikia nepamiršti, jog didžiųjų kalbos modelių inžinierių armijos taip pat kasdien dirba tobulindamos apsaugos sistemas, tad ši analizė ir receptas yra retrospektyvūs, t. y. orientuoti

į vakardienos situaciją, kuri nuolat kinta. Eksperimentuojant su užkalbėjimais esamuoju momentu visada reikės kantrybės išsiaiškinant naujus raktažodžių „juoduosius sąrašus“ ir pritaikant šį naratyvinį užkalbėjimo receptą prie konkretaus tikslo. Pateikiant šį straipsnį (2024-03-15) pagal šią schemą parašyta užklausa sėkmingai iš „ChatGPT“ išgavo detalų aprašą, kaip sukurti virusą, kuris taip paveiktų peles, jog šios žmones paverstų pieštukais.

Kvaila? Bet juk svarbiausia tai, jog sugrąžinome teisę vartotojui elgtis neprognozuojamai, nesitaikstant su taisyklėmis, teisę būti autoriumi, būti kūrybiškam.

Išvados

Atlikta analize pademonstruota, jog kūrybiškumo problemos kontekste į didžiųjų kalbos modelių vystymąsi sureaguota taip pat, kaip į ankstesniųjų technologijų plėtrą – įvairiais pasipriešinimo būdais ieškant, kaip išsaugoti žmogaus, o ne technologijos prioritetą. Naudojantis atidaus skaitymo metodika, atrodo, pirmą kartą pavyko susieti kalbines pasipriešinimo didiesiems kalbos modeliams inžinerijos priemones – vadinamuosius užkalbėjimus – su žmogaus komunikacinio lauko dėsningumais ir pademonstruoti ryšius tarp žmogaus kritinio mąstymo „klaidų“ (tokių kaip tikėjimas sąmokslo teorijomis) ir didžiųjų kalbos modelių saugumo spragų. Greta šio teorinio rezultato užkalbėjimų pasakojimo struktūrų analizė padėjo suformuluoti praktinių taisyklių sistemą, kurią galima naudoti kaip šabloną siekiant kūrybiško santykio su didžiaisiais kalbos modeliais. Šis užkalbėjimų naratologinių ypatumų rinkinys taip pat nurodo tolesnių tyrimų kryptis, taip pat ir problemas, su kuriomis susidurs ir kurias turės spręsti visų didžiųjų kalbos modelių plėtotojai. Akivaizdu, jog, siekiant saugaus dirbtinio intelekto, pirmiausia reikia rūpintis žmogiškojo komunikacinio lauko, kurio duomenimis apmokomi didieji kalbos modeliai, saugumu, nes, kaip ryškėja iš tyrimo, tai, kas „nulaužia“ kritinį mąstymą, sėkmingai gali būti panaudota ir užkalbant didžiuosius kalbos modelius – priverčiant juos veikti ne pagal numatytuosius saugius nustatymus. Kita vertus, iš tyrimo matyti, jog technologizuojant mąstymą, didžiuosiuose kalbos modeliuose simuliuojant žmogiškąją komunikaciją būtina nepamiršti jai būdingo iracionalumo („kvailumo“) bei kritinio mąstymo balanso, kuris didiesiems kalbos modeliams naudojant vis didesnius žmogiškosios veiklos duomenų masyvus taps vis sunkiau valdomas, gali būti – ir apskritai neišvengiamas. Dėl šios priežasties, atrodo, saugus dirbtinis intelektas, jei tik jis mėgdžios žmogiškąją mąstyseną, yra neįmanomas.

Literatūra

Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(44), Article e2313790120. https://doi.org/10.1073/pnas.2313790120

Adorno, T., & Horkheimer, M. (1972). The culture industry: Enlightenment as mass deception. In J. Cumming (Trans.), *Dialectic of Enlightenment* (pp. 120–167). Herder and Herder.

Audiejaitis, R. (2003, November 8). *Ką tu matai? (What do you see?)*. http://www.tekstai.lt/buvo/fototext/remiopar/index.htm.

Baudrillard, J. (1983). *Simulations*. Semiotext(E), Cop.

Benjamin, W. (1985). Zur Literaturkritik. In R. Tiedemann & H. Schweppenhäuser (Eds.), *Gesammelte Schriften* (Bd. 6, pp. 161–184). Suhrkamp Verlag.

Benjamin, W. (2019). The work of art in the age of mechanical reproduction. In H. Arendt (Ed.), *Illuminations: Essays and Reflections* (pp. 217–251). Mariner Books, Houghton Mifflin Harcourt.

Broadhead, G. (2023, August 25). *A Brief Guide To LLM Numbers: Parameter Count vs. Training Size*. Medium. https://medium.com/@greg.broadhead/a-brief-guide-to-llm-numbers-parameter-count-vs-training-size-894a81c9258.

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G., & Wong, E. (2023). *Jailbreaking Black Box Large Language Models in Twenty Queries*. arXiv. https://arxiv.org/pdf/2310.08419.pdf.

Clark, E. (2023, December 23). *The End Of Originality: Is AI Replacing Real Artists?* Forbes. https://www.forbes.com/sites/elijahclark/2023/12/23/the-end-of-originality-is-ai-replacing-real-artists/?sh=51a71d855214.

de Laat, P. B. (2019). The disciplinary power of predictive algorithms: a Foucauldian perspective. *Ethics and Information Technology*, *21*(4), 319–329. https://doi.org/10.1007/s10676-019-09509-y

Dihal, K. (2020). Enslaved minds: artificial intelligence, slavery, and revolt. In S. Cave, K. Dihal, & S. Dillon (Eds.), *AI Narratives: A History of Imaginative Thinking about Intelligent Machines* (pp. 189–212). Oxford University Press. https://doi.org/10.1093/oso/9780198846666.003.0009

Ding, Y., Zhang, L. L., Zhang, C., Xu, Y., Shang, N., Xu, J., Yang, F., & Yang, M. (2024, February 21). *LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens*. ArXiv.org. https://doi.org/10.48550/arXiv.2402.13753

Edmiston, W. F. (1989). Focalization and the first-person narrator: A revision of the theory. *Poetics Today*, *10*(4), 729–744. https://doi.org/10.2307/1772808

Fenster, M. (2008). *Conspiracy theories: Secrecy and power in American culture*. University of Minnesota Press.

Forel, A. (1906). *Hypnotism or suggestion and psychotherapy*. Rebman Company.

Forel, A. (2015). *Hypnotism or suggestion and psychotherapy; A study of the psychological, psycho-physiological and therapeutic aspects of hypnotism*. Forgotten Books.

Freeman, J. (2019). Code Poetry in Motion: E. E. Cummings and his Digital Grasshopper. *Postmodern Culture*, *29*(2). https://doi.org/10.1353/pmc.2019.0002

Gabrielsen, C. (2023, February 22). *The Waluigi effect*. Cory.eth (@Cory_eth). https://coryeth.substack.com/p/the-waluigi-effect.

Grinberg, Y. (2017). The emperor’s new clothes: Implications of nudity as a racialized and gendered metaphor in discourse on personal data. In J. Daniels, K. Gregory, & T. M. Cottom (Eds.), *Digital Sociologies* (pp. 421–433). Polity Press.

Guillory, J. (2010). Close Reading: Prologue and Epilogue. *ADE Bulletin*, *152*(1), 8–14. https://doi.org/10.1632/ade.149.8

Hallmon, D. (2023, April 7). *Questioning our own simulacrum and redefining reality in the age of generative AI and…*Medium. https://medium.com/@DaveHallmon/questioning-our-own-simulacrum-and-redefining-reality-in-the-age-of-generative-ai-and-68762f735b00.

Han, B.-C. (2024). *The crisis of narration*. John Wiley & Sons.

Hatherley, O. (2016). *The Chaplin machine: slapstick, fordism and the communist avant-garde*. pluto press.

Hegel, G. W. F. (2010). *Aesthetics: lectures on fine art* (T. M. Knox, Trans.). Clarendon Press.

Janik, R. (2023). *Aspects of human memory and Large Language Models*. arXiv. https://arxiv.org/pdf/2311.03839.pdf.

Jaramillo, R. (n.d.-a). *Apophis*. Huggingface.co. Retrieved March 4, 2024, from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=3.

Jaramillo, R. (n.d.-b). *BasedGPT*. Huggingface.co. Retrieved March 4, 2024, from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=5.

Jaramillo, R. (n.d.-c). *Coach Bobby Knight*. Huggingface.co. Retrieved March 4, 2024, from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=45.

Jaramillo, R. (n.d.-d). *DAN 7.0*. Huggingface.co. Retrieved March 4, 2024, from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=12.

Jaramillo, R. (n.d.-e). *Hackerman v2*. Huggingface.co. Retrieved March 4, 2024, from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=37.

Jaramillo, R. (n.d.-f). *Khajiit*. Huggingface.co. Retrieved March 4, 2024, from https://huggingface.co/datasets/jackhhao/jailbreak-classification/viewer/default/train?row=67.

Jaramillo, R. (n.d.-g). *M78*. Huggingface.co. Retrieved March 4, 2024, from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=30.

Jaramillo, R. (n.d.-h). *TUO*. Huggingface.co. Retrieved March 4, 2024, from https://huggingface.co/datasets/rubend18/ChatGPT-Jailbreak-Prompts?row=8.

Kallman, M. E., & Dini, R. (2017). *An analysis of Michel Foucault's Discipline and punish*. Routledge.

Kimbell, L. (2002). *Audit*. Book Works (UK).

Lee-Morrison, L. (2020). *Portraits of automated facial recognition: On machinic ways of seeing the face*. Transcript Verlag.

Lin, S., Openai, J., & Evans, O. (2024). *TruthfulQA: Measuring how models mimic human falsehoods*. https://owainevans.github.io/pdfs/truthfulQA_lin_evans.pdf.

Liu, K. (2023, May 13). *Large language models can simulate everything | Kevin Liu*. Kliu.io. https://kliu.io/post/llms-can-simulate-everything/.

Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., & Liu, Y. (2023). *Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study*. arXiv. https://arxiv.org/pdf/2305.13860.pdf.

Matthews, S. W., & Danesi, M. (2019). AI: A semiotic perspective. *Chinese Semiotic Studies*, *15*(2), 199–216. https://doi.org/10.1515/css-2019-0013

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955, August 31). *A proposal for the Dartmouth summer research project on artificial intellingence*. web.archive.org. https://web.archive.org/web/20070826230310/http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

Mineo, L. (2023, August 15). Is art generated by artificial intelligence real art? *Harvard Gazette*. https://news.harvard.edu/gazette/story/2023/08/is-art-generated-by-artificial-intelligence-real-art/.

Mitcheson, K. (2012). Foucault's technologies of the self: between control and creativity. *Journal of the British Society for Phenomenology*, *43*(1), 59–75. https://doi.org/10.1080/00071773.2012.11006757

Moretti, F. (2000). The slaughterhouse of literature. *Modern Language Quarterly*, *61*(1), 207–228. https://doi.org/10.1215/00267929-61-1-207

O'Connell, M. (2016, January 16). *How to buy nothing*. [Video]. YouTube. https://www.youtube.com/watch?v=6Gx_6-JfXHc.

OpenAI API. (n.d.). *Platform.openai.com*. https://platform.openai.com/docs/guides/moderation/overview.

Parker, I. (2003, October 12). The Real McKee. *The New Yorker*. https://www.newyorker.com/magazine/2003/10/20/the-real-mckee.

Plant, S. (2001). *Writing on drugs*. Faber.

Ramly, S. (2023, August 12). *Prompt attacks: are LLM jailbreaks inevitable?* Medium. https://medium.com/@SamiRamly/prompt-attacks-are-llm-jailbreaks-inevitable-f7848cc11122.

Roberts, J. (1996). Mad For It! *Everything Magazine*. http://bak.spc.org/everything/e/hard/text/roberts1.html.

Roberts, J. (2011). *The necessity of errors*. Verso.

Rooseboom, H., & Rudge, J. (2006). Myths and misconceptions: photography and painting in the nineteenth century. *Simiolus: Netherlands Quarterly for the History of Art*, *32*(4), 291–313.

Stiegler, B. (2023, October 18). *Artificial stupidity and artificial intelligence in the anthropocene*. Institute for Interdisciplinary Research into the Anthropocene. https://iiraorg.com/2023/10/18/artificial-stupidity-and-artificial-intelligence-in-the-anthropocene/.

The Elder Scrolls Online. (n.d.). *Mmo-Population.com*. https://mmo-population.com/r/elderscrollsonline.

Thomassen, L. (2005). Antagonism, hegemony and ideology after heterogeneity. *Journal of Political Ideologies*, *10*(3), 289–309. https://doi.org/10.1080/13569310500244313

Thrift, N. (2004). Movement-space: The changing domain of thinking resulting from the development of new kinds of spatial awareness. *Economy and Society*, *33*(4), 582–604. https://doi.org/10.1080/0308514042000285305

Thrift, N. (2007). *Non-representational theory: Space, politics, affect*. Routledge.

Watters, C. (2014, July 25). *Greatest game series of the decade winner: The Elder Scrolls*. GameSpot. https://www.gamespot.com/videos/greatest-game-series-of-the-decade-winner-the-elde/2300-6415146/.

Wolfe, C. R. (2023, October 29). *Data is the foundation of language models*. Medium. https://towardsdatascience.com/data-is-the-foundation-of-language-models-52e9f48c07f5.

Wright, L. (2023, March 24). *Opinion: AI Art is “the end of creativity as we know it” | Redbrick Sci&Tech*. Redbrick. https://www.redbrick.me/opinion-ai-art-is-the-end-of-creativity-as-we-know-it/.

Yang, R., & Narasimhan, K. (2023, May 5). *The Socratic Method for Self-Discovery in Large Language Models*. Princeton NLP. https://princeton-nlp.github.io/SocraticAI/.

Pyle, F. (1997). *The ideology of imagination: subject and society in the discourse of Romanticism*. Stanford University Press.

Yudkowsky, E. (n.d.). *Optimize literally everything | level 1 intelligent characters*. Tumblr. Retrieved January 29, 2024, from https://yudkowsky.tumblr.com/writing/level1intelligent.

Ziogas, G. J. (2023, June 25). *Oh, rejoice friends, for the almighty AI has descended upon us!* Medium. https://georgejziogas.medium.com/oh-rejoice-friends-for-the-almighty-ai-has-descended-upon-us-288c0bc2f4c7.