TURINYS

PROLOGAS

Štai kaip šią situaciją regi DI.

Klausimas: ką ši kylanti technologijų banga žada žmonijai?

Žmonijos istorijos analuose esama lūžio momentų, kai ant plauko buvo pakibęs jos likimas. Išmokimas panaudoti ugnį, rato, vėliau elektros energijos išradimas – šie momentai transformavo žmonių civilizaciją ir visiems laikams pakeitė istorijos kryptį.

O dabar stovime ant dar vieno tokio momento krašto: susiduriame su kylančia technologijų banga, į kurią įeina ir pažangus DI, ir pažangios biotechnologijos. Dar nesame regėję technologijų, turinčių tokį potencialą transformuoti ir žadančių pakeisti mūsų pasaulį taip, jog tie pokyčiai mums atims žadą ir kels baimę.

Viena vertus, šios technologijos gali duoti didžiulę, pažangią naudą. Pasitelkdami DI galėtume atskleisti visatos paslaptis, išgydyti ligas, kurių ilgą laiką negalime įveikti, bei sukurti naujų meno ir kultūros formų, praplėsiančių mūsų vaizduotės ribas. O naudodamiesi biotechnologijomis galėtume sukurti gyvybės formų, kurios kovotų su ligomis, pakeistų žemės ūkį ir taip sukurtų sveikesnį bei tvaresnį pasaulį.

Kita vertus, tos pačios technologijos gali užtraukti tokį pat didžiulį pavojų. Tuo pačiu DI galime sukurti mums nepavaldžias sistemas ir taip atsidurti mums nebesuprantamų algoritmų malonėje. Biotechnologijomis galėtume pakeisti pačias gyvybės sudedamąsias dalis ir taip potencialiai sukelti nepageidaujamų padarinių tiek mums kaip individams, tiek ištisoms ekosistemoms.

Stovėdami ties šiuo lūžio tašku esame priversti rinktis tarp neprilygstamų galimybių ir neregėtų pavojų ateities. Ant plauko kybo žmonijos likimas, tad sprendimai, kuriuos priimsime ateinančiais metais ir dešimtmečiais, lems, ar sėkmingai įveiksime šių technologijų keliamus iššūkius, ar nukentėsime nuo jų sukeltų pavojų.

Vis dėlto atėjus šiam neaiškumo kupinam momentui aišku viena: pažangių technologijų amžius jau čia pat, tad turime būti pasirengę susidurti su jo kelsimais iššūkiais.

Šį pirmiau pateiktą tekstą parašė DI. Likusi knygos dalis parašyta žmogaus, tiesa, ją greitai galės parašyti ir DI. Štai kas mūsų laukia.

Netrukus jis pasklis dar plačiau, ir kone visose srityse jūsų darbas dėl patirties su DI technologijomis taps našesnis, spartesnis, lengvesnis ir sklandesnis.

DI jau čia. Tačiau jis toli gražu nėra užbaigtas.

VISKĄ ATLIKITE AUTOMATIŠKAI: DIDELIŲ KALBOS MODELIŲ IŠKILIMAS

Dar visai neseniai atrodė, kad dirbtiniam intelektui per sunku apdoroti natūralią kalbą, kad jis negali susidoroti su kalbų įvairove, niuansais. Tada, 2022-ųjų lapkritį, DI tyrimų įmonė „OpenAI“ išleido DI sistemą „ChatGPT“. Po savaitės ja naudojosi per milijoną vartotojų, apie ją kalbėta kaip apie be galo reikšmingą dalyką – tokią naudingą technologiją, kuri netrukus gali užimti „Google“ paieškos vietą.

Paprastai tariant, „ChatGPT“ yra pokalbių botas. Tačiau jis gerokai pažangesnis ir išmanantis daugiau sričių negu kuri nors kita anksčiau viešai prieinama sistema. Tereikia užduoti jam klausimą ir jis iškart sklandžiai atsakys. Paprašykite jį parašyti rašinį, spaudos pranešimą ar verslo planą karaliaus Jokūbo Biblijos arba XX a. 9-ojo dešimtmečio reperio stiliumi ir jis tai padarys vos per kelias sekundes. Liepkite jam parašyti fizikos kurso programą, dietą ar programinį kodą „Python“ programavimo kalba, ir jis tai parašys.

Žmonės yra protingi iš dalies todėl, kad žvelgia į praeitį siekdami nuspėti ateitį. Šia prasme, intelektą galima suvokti kaip gebėjimą sudaryti aibę galimų scenarijų, pagal kuriuos veiktų aplinkinis pasaulis, ir tada imtis protingai veikti atsižvelgiant į sudarytas prognozes. 2017 m. nedidelė mokslininkų grupė įmonėje „Google“ bandė išspręsti supaprastintą šios problemos variantą: kaip padaryti, kad DI sistema daugiausia dėmesio skirtų tik svarbiausioms duomenų dalims, kad galėtų atlikti tiksliausias ir itin kvalifikuotas ateities prognozes. Jų darbas paklojo pagrindą didelių kalbos mo-

delių (DKM) (*large language models, LLM*), įskaitant „ChatGPT“, srityje, kurioje jų darbas sukėlė tikrą perversmą.

DKM pasinaudoja tuo, jog kalbiniai duomenys pateikiami nuoseklia tvarka. Kiekvienas informacijos vienetas vienaip ar kitaip susijęs su ankstesniais duomenų vienetais. Tuomet kalbos modelis perskaito daugybę sakinių, sužino abstrakčią jų reikšmę ir, naudodamasis įgytomis žiniomis, pateikia kito sakinio prognozę. Tačiau sunkiausia sukurti algoritmą, kuris „žinotų, kur“ kiekviename sakinyje „ieškoti“ signalų. Kokie yra raktiniai žodžiai, reikšmingiausios sakinio dalys ir kaip jie vienas su kitu susiję? DI srityje tai vadinama „dėmesiu“.

Didelis kalbos modelis, apdorojęs sakinį, sudaro savotišką „dėmesio žemėlapį“. Pirmiausia dažnai pasitaikančių raidžių ar skyrybos ženklų grupę suskirsto į „simbolius“, kurie panašūs į skiemenis, bet iš tikrųjų tėra dažnai pasikartojančios raidės, – taip kalbos modeliui lengviau apdoroti informaciją. Svarbu pabrėžti, kad žmonės, žinoma, irgi taip daro su žodžiais, tačiau kalbos modelis nevartoja mūsų žodyno. Užuot vartojęs mums žinomus žodžius, jis sudaro naują žodyną iš dažnai pasikartojančių simbolių, tai jam padeda tarp milijardų dokumentų pastebėti sekas. Dėmesio žemėlapyje kiekvienas simbolis yra kaip nors susijęs su visais ankstesniais simboliais, ir šios sąsajos į kalbos modelį įvestam sakiniui suteikia informacijos apie simbolio svarbą sakinyje. Iš esmės DKM išmoksta atkreipti dėmesį į reikiamus žodžius.

Pavyzdžiui, įvedus sakinį „Rytoj Brazilijoje bus gana didelė audra“, kalbos modelis greičiausiai sukurs simbolius iš raidžių „us“, esančių žodyje „bus“, bei „oj“, esančių žodyje „rytoj“, kadangi jos dažnai pasitaiko ir kituose žodžiuose. Analizuodamas visą sakinį kalbos modelis supras, kad jame esminiai žodžiai yra „audra“, „rytoj“ ir „Brazilija“, prieis išvadą, kad Brazilija – tai vieta, kad audra bus rytoj ir t. t. Tada pagal šią informaciją kalbos modelis bandys nuspėti, kokie simboliai turėtų būti tolesnėje sekoje, kokia išvestis

būtų logiška įvesties atžvilgiu. Kitaip tariant, jis pats užbaigia galimą tolesnį tekstą.

Tokios sistemos vadinamos transformeriais. Nuo pat tada, kai 2017 m. „Google“ mokslininkai išleido pirmąjį mokslinį straipsnį apie juos, transformeriai ėmė tobulėti pribloškiančiu tempu. Netrukus „OpenAI“ išleido „GPT-2“ (*generative pre-trained transformer**). Tuo metu tai buvo didžiulis kalbos modelis. „GPT-2“, turįs 1,5 milijardo parametrų (parametrų skaičius yra pagrindinis DI sistemų masto ir sudėtingumo rodiklis), buvo apmokytas pagal 8 milijonus skaitmeninio teksto puslapių. Tačiau tikrą kalbos modelių reikšmės mastą žmonės ėmė suprasti tik 2020 m. vasarą, „OpenAI“ išleidus „GPT-3“. Ši versija turėjo neįtikėtinus 175 milijardus parametrų, tuomet tai buvo didžiausias kada nors sukurtas neuroninis tinklas – šimtąkart didesnis negu vos prieš metus išleistas jo pirmtakas. „GPT-3“ tikrai įspūdingas kalbos modelis, tačiau jo mastas jau įprastas, o į jį panašaus modelio apmokymo kaštai per pastaruosius dvejus metus sumažėjo dešimteriopai.

2023 m. kovą išleidus „GPT-4“ sulaukta tokių pat įspūdingų rezultatų. Kaip ir jo ankstesniųjų versijų, „GPT-4“ galima paprašyti sukurti poezijos Emily'ės Dickinson stiliumi, ir jis paklūsta; galima paprašyti pratęsti atsitiktinę „Žiedų valdovo“ ištrauką, ir netrukus jūs jau skaitote visai įtikinamą Tolkieno imitaciją; paprašykite sudaryti verslo planą startuoliui ir gavę rezultatą pasijusite, lyg kalbėtumėte su daugybe įmonių direktorių. Be to, šis modelis gali tobulai išlaikyti standartizuotus testus, pradedant advokatūros egzaminu ir baigiant stojamaisiais egzaminais.

Jis taip pat geba apdoroti vaizdus ir programinį kodą, kurti trimačius žaidimus, veikiančius jūsų naršyklėje, kurti programas išmaniesiems telefonams, ištaisyti klaidas jūsų programiniame kode, pastebėti trūkumus sutartyse ir pasiūlyti cheminę sudėtį naujiems vaistams ir netgi patarti, kaip juos pakeisti, kad jie nebūtų užpaten-

* Generatyvus iš anksto apmokytas transformeris.

tuoti. Jis geba sukurti internetinę svetainę iš jūsų ranka nupieštų paveikslėlių ir suvokti įmantrias žmogiškąsias ypatybes sudėtingose scenose; parodykite jam šaldytuvą, ir jis sugalvos receptą iš jame esančių produktų; parašykite jam pristatymo juodraštį, o jis jį nugludins, pateiks profesionaliai atrodantį variantą. Atrodo, kad „GPT-4“ „supranta“ erdvinius ir priežastinius daiktų ryšius, mediciną, teisę ir žmonių psichologiją. Praėjus vos keletui dienų po modelio išleidimo, žmonės pasitelkdami jį sukūrė DI sistemų, padedančių teisminiuose procesuose, jis padėjo auginti vaikus ir netgi realiuoju laiku patarė, kaip rengtis. Po kelių savaičių modelio vartotojai sukūrė papildinių, kurių padedamas „GPT-4“ galėtų atlikti sudėtingas užduotis, tarkime, kurti aplikacijas mobiliesiems įrenginiams arba atlikti rinkos tyrimą ir pateikti išsamią ataskaitą.

Ir visa tai tik pradžia. Mes dar tik pradedame suvokti, kokį didžiulį poveikį netrukus padarys dideli kalbos modeliai. DQN ir „AlphaGo“ buvo pirmieji ženklai, rodantys, kas mūsų laukia ateityje, o „ChatGPT“ ir dideli kalbos modeliai buvo pirmieji lūžtančios bangos pranašai. 1996 m. internetu naudojosi 36 milijonai žmonių, šiemet* juo naudosis daugiau nei 5 milijardai. Turėtume numatyti, kad tokia pat trajektorija sklis ir šie įrankiai, tik daug greičiau. Mano manymu, per kitus penkerius metus DI taps toks pat visuotinis kaip internetas: toks pat prieinamas ir netgi dar reikšmingesnis.

SMEGENŲ MASTO MODELIAI

Mano aprašomų DI sistemų veiklos mastas milžiniškas. Štai vienas pavyzdys.

Didžiąją dalį DI pažangos XXI a. 2-ojo dešimtmečio viduryje leido padaryti efektyvus „prižiūrėtas“ gilusis mokymasis. Taikant tokį metodą, DI modeliai mokosi iš duomenų, kuriems tikslias eti-

* Knyga originalo kalba pirmąkart išleista 2023 m. rugsėjį.

srityje daroma nemenka pažanga. Jie greitai mane apsėdo, perskaičiau šimtus mokslinių straipsnių šia tema, visa galva pasinėriau į šią sparčiai plėtojamą sritį. 2020 m. vasarą jau buvau įsitikinęs, kad kompiuterijos ateitis bus susijusi su pokalbiais. Galima teigti, kad bet kokia sąveika su kompiuteriu jau yra tam tikras pokalbis, tik jis mygtukais, klavišais, pikseliais padeda žmonių mintis išversti į kompiuteriui suprantamą kodą. Greitai tai jau nebesudarys kliūties. Kompiuteriai netrukus supras *mūsų* kalbą. Ši galimybė kėlė ir tebekelia jaudulį.

Gerokai anksčiau nei viešumoje nuskambėjo „ChatGPT“ pasirodymas, buvau vienas „Google“ komandos narių, dirbęs prie naujo didelio kalbos modelio, mūsų praminto „LaMDA“ – *Language Model for Dialogue Applications* (Kalbos modelis dialogų aplikacijoms). „LaMDA“ yra pažangus DKM, sukurtas būti puikiu pašnekovu. Iš pradžių jis buvo sunkiai suprantamas, nenuoseklus, dažnai susipainiodavo. Bet jame slypėjo ir genialumo grūdas. Po kelių dienų lioviausi pirmiausia naudotis paieškos sistema. Pirma pasišnekučiuodavau su „LaMDA“, norėdamas ką nors išsiaiškinti, o vėliau patikrindavau, ar gaudavau faktiškai teisingą informaciją. Pamenu, vieną vakarą sėdėjau namuose ir svarsčiau, ką pagaminti vakarienei. „Paklausk *LaMDA*“, – pamaniau. Ir vos po kelių akimirkų pasinėrėme į ilgą, užsitęsusį pokalbį apie įvairius spagečių su Bolonijos padažu receptus: aptarėme skirtingus spagečius, padažus iš įvairių regionų, ar šventvagiška pridėti grybų. Tuo metu troškau būtent tokio banalaus, bet įtraukiančio pokalbio, ir jis man atvėrė akis.

Ilgainiui ėmiau vis dažniau naudoti „LaMDA“. Vieną sekmadienio popietę nusprendžiau, kad atėjo metas įsigyti naują spausdintuvą. „LaMDA“ pateikė puikių pasiūlymų, įvardijo skirtingų modelių privalumus ir trūkumus, padėjo apsispręsti, kokio spausdintuvo man reikėjo. Galiausiai nusipirkau prabangų spausdintuvą, galintį spausdinti nuotraukas. Tai mane paragino į „LaMDA“ integruoti

DVYLIKTAS SKYRIUS

Didžioji dilema

KATASTROFA: DIDŽIAUSIA NESĖKMĖ

Žmonijos istorija iš dalies yra katastrofos istorija. Joje gausu pandemijų. Dvi jų pražudė 30 procentų pasaulio gyventojų: VI a. Justiniano maras ir XIV a. juodoji mirtis. 1300 m. Anglijoje gyveno 7 milijonai žmonių, o 1450 m. po kelių maro epidemijos bangų šalies populiacija sumažėjo iki 2 milijonų.

Žinoma, katastrofas gali sukelti ir žmonės. Antai Pirmasis pasaulinis karas pražudė maždaug 1 procentą pasaulio gyventojų, Antrasis pasaulinis – 3 procentus. Arba, pavyzdžiui, Čingischano ir mongolų armijos smurto banga, XIII a. nuvilnijusi per Kiniją ir Centrinę Aziją bei pasiglemžusi iki 10 procentų pasaulio populiacijos. Sukūrus atominę bombą, žmonija įgijo tokią mirtiną jėgą, kuri pajėgi keletą kartų išnaikinti visus planetos gyventojus. Katastrofos, kurios anksčiau užtrukdavo metus ar dešimtmečius, dabar gali įvykti per kelias minutes, nuspaudus vieną mygtuką.

Kylanti banga žmonijai žada dar vieną tokio lygio šuolį, praplėsiantį rizikos ribas ir suteiksiantį daugiau galimybių tiems, kurie norės panaudoti katastrofinę galią. Šiame skyriuje žengsime toliau silpnumo, grėsmių, valstybės funkcionavimo keliais ir įsivaizduosime, kas nutiks – anksčiau ar vėliau – nesugebėjus suvaldyti kylančios bangos.

Dauguma šios bangos technologijų bus panaudota geriems tikslams. Nors daugiau dėmesio skyriau su jomis susijusioms rizikoms, tačiau svarbu nepamiršti, kad jos kasdien pagerins milijardų žmonių gyvenimą.

Šiame skyriuje aptarsime kraštutinius atvejus, kurių beveik niekas nenori matyti, ypač žmonės, dirbantys su įrankiais, kurie ir lems tuos atvejus. Vis dėlto vien todėl, kad ateityje tokių atvejų bus labai nedaug, nereiškia, kad juos galime ignoruoti. Jau matėme, kad piktavaliai veikėjai gali pridaryti daug žalos, sukelti didžiulį nestabilumą. O dabar įsivaizduokite laikus, kai bet kuri bent kiek kompetentinga laboratorija ar programišius gebės sintetinti sudėtingas DNR grandines. Ar ilgai reikės laukti, kol įvyks katastrofa?

Visur išplitus vienoms galingiausių visų laikų technologijų, šie kraštutiniai atvejai taps tikėtinesni. Galiausiai kur nors įvyks klaida, tačiau tokiu mastu ir greičiu, kuris prilygs panaudotiems pajėgumams. Kylančios bangos keturių ypatybių esmė ta, kad neturint tvirtų sulaikymo būdų, veikiančių visais lygmenimis, galimybė įvyksiant katastrofą, kaip antai dirbtinai sukelta pandemija, tampa kaip niekad reali.

Tai nepriimtina, bet vis vien susiduriame su didžiąja dilema: patikimiausi sulaikymo būdai taip pat nepriimtini, jie žmoniją veda į autoritarizmą, antiutopiją.

Viena vertus, visuomenės gali atsigręžti į tokias technologijų sukurtas stebėjimo sistemas, kokias aptarėme praeitame skyriuje: tai natūralus atsakas, sudėtingus mechanizmus panaudojantis prieš nepastovias ar nekontroliuojamas technologijas. Saugumas už laisvę. Arba žmonija gali visai nutolti nuo naujausių technologijų. Nors tokia reakcija mažai tikėtina, tai irgi nėra gera išeitis. Vienintelis darinys, iš principo pajėgus išlaviruoti iš šio egzistencinio akligatvio, yra ta pati šiuo metu byranti tautinių valstybių sistema, ardoma tų pačių jėgų, kurias jai privalu suvaldyti.

Ilgainiui šių technologijų pasekmės žmoniją privers laviruoti tarp kelio į katastrofą ir kelio į antiutopiją. Tai esminė mūsų laikų dilema.

Technologija turi potencialą pagerinti žmonėms gyvenimą, jos nauda gerokai nusveria kaštus ir minusus. Tačiau blogi pasirinkimai reiškia, kad šis potencialas bus apverstas aukštyn kojomis.

Nuolat girdėdami kalbas apie pasaulio pabaigą žmonės – įskaitant mane – liaujasi į jas rimtai žiūrėję. Tada į jas galite pradėti žiūrėti įtariai, skeptiškai. Kalbos apie katastrofiškus įvykius dažnai išjuokiamos: pasigirsta kaltinimai katastrofizmu, perdėtu negatyvumu, skambinimu pavojaus varpais, kad per daug dėmesio skiriama tolimoms ir mažai tikėtinoms rizikoms, nors mūsų dėmesio reikalauja akivaizdūs, neatidėliotini pavojai. Nepaliaujamą technokatastrofizmą, kaip ir tokį pat technooptimizmą nesunku atmesti kaip iškreiptą, suklaidintą ir istoriniais pavyzdžiais nepagrįstą entuziazmą.

Vis dėlto vien tai, kad perspėjimas kalba apie dramatiškas pasekmes, nėra geras pagrindas jį automatiškai atmesti. Kad katastrofa tikrai įvyktų, užtenka vien pesimizmo ignoravimo, sveikinančio katastrofos tikimybę. Ir nors perspėjimus atmesti kaip išpūstas kelių keistuolių kalbas atrodo tinkama, racionali reakcija, tačiau toks požiūris tiesia kelią nesėkmei.

Nėra jokios abejonės, kad su technologijomis susijusi rizika mus pastato į neaiškią padėtį. Šiaip ar taip, visos tendencijos rodo, jog susidursime su aibe rizikų. Šis spėjimas pagrįstas nuolat vienas kitą papildančiais moksliniais ir technologiniais patobulinimais. Todėl, mano nuomone, tie, kurie į katastrofą numoja ranka, neatsižvelgia į akivaizdžius objektyvius faktus. Juk kalbame ne apie motociklų ar skalbimo mašinų sklaidą.

KATASTROFŲ ĮVAIROVĖ

Norėdami išsiaiškinti, kokioms katastrofoms turėtume ruoštis, tiesiog jas išveskite iš piktavalių veikėjų išpuolių, aptartų dešimtame skyriuje. Štai keletas galimų scenarijų pavyzdžių.

Teroristai automatiniais ginklais su veido atpažinimo funkcija aprūpina autonominių dronų, gebančių staigiai modifikuotis, – ar kalbėtume apie ginklo atatranką, šaudymą trumpomis serijomis, ar judėjimą pirmyn – spiečių, sudarytą iš šimtų ar tūkstančių įrenginių. Šiuos dronus jie paleidžia į kurio nors didžiulio miesto centrą, jiems nurodyta žudyti pagal konkrečius kriterijus. Sausakimšą piko valandą šie dronai veiktų priblokiamai tiksliai, mieste judėtų tinkamiausiu keliu. Vos po kelių minučių įvyktų ataka, kuri būtų gerokai didesnio masto negu, pavyzdžiui, 2008 m. išpuolis Mumbajuje, kai ginkluoti teroristai lakstė po žymiausias miesto vietas, kaip antai centrinėje traukinių stotyje.

Kitame scenarijuje masinis žudikas nusprendžia dronais, purškimo įrenginiais ir pagal užsakymą sukurtu patogenu smogti didžiuliam politiniam mitingui. Netrukus po išpuolio suserga mitingo dalyviai, vėliau jų šeimos. Pagrindinis kalbėtojas mitinge, daugelio mylimas ir daugelio neapkenčiamas politinis atpirkimo ožys, tampa viena pirmųjų aukų. Susidarius karštligiškai, susiskaldžiusiai aplinkai, išpuolis įžiebia smurtingą kerštavimą visoje šalyje, kyla chaosas.

Dar kitame scenarijuje DI sistemai pateikęs keletą nurodymų žmonių kalba, piktų kėslų turintis sąmokslininkas Amerikoje paskleidžia didžiulį kiekį kruopščiai parengtos, žmones priešinančios dezinformacijos. Visuomenė didžiąją dalį jos atmeta. Galiausiai vienas bandymas pavyksta: policininkas Čikagoje nužudo žmogų. Istorija visiškai sufabrikuota, tačiau chaosas gatvėse, visuotinis pasipiktinimas – tikri. Išpuolį surengusiems asmenims pavyko rasti veiksmingą modelį. Ir kai pagaliau paaiškės, kad vaizdo įrašas pa-

dirbtas, per šalį jau bus nusiritusios daugybės aukų pareikalavusios riaušės, o nauji dezinformacijos gūsiai ir toliau kurstys kilusį gaisrą.

Arba įsivaizduokite visus šiuos įvykius vykstant vienu metu. Arba kad jie sudaro ne vieną įvykį ir vyksta ne viename mieste, o šimtuose skirtingų vietų. Turint tokius įrankius nereikia ilgai mąstyti, kad suvoktume, jog galių suteikimas piktavaliams veikėjams sudaro sąlygas katastrofai. Nūdienos DI sistemos vos susilaiko nepateikusios instrukcijų, kaip užnuodyti vandentiekio vandenį ar sukurti neaptinkamą bombą. Kol kas jos dar negeba pačios apibrėžti savų tikslų ir jų siekti. Tačiau, kaip matėme, netrukus turėsime ir plačiai paplitusius, ir ne tokius saugius naujausius ir galingiausius DI modelius.

Kalbant apie visas katastrofiškas rizikas, atsirandančias iš kylančios bangos, daugiausia dėmesio sulaukė dirbtinis intelektas. Bet egzistuoja begalė kitų rizikų. Pavyzdžiui, visiškai automatizavus kariuomenes, įsitraukti į konfliktą bus daug paprasčiau negu dabar. Gali būti, kad karas kils atsitiktinai ir dėl niekad nepaaiškėsiančių priežasčių: DI suras kokį nors elgesio modelį arba grėsmę ir akimirksniu sureaguos, atsakydamas triuškinama jėga. Gana pasakyti tiek, kad ateityje karo prigimtis gali būti mums svetima, kad karas greitai eskaluosis ir savo destruktyvumu pranoks visus kitus karus.

Žmonijai jau yra tekę susidurti su dirbtinai sukeltomis pandemijomis ir atsitiktinai nutekėjusių virusų pavojais, jai tekę regėti, kas nutinka, kai milijonai savimodifikacijos entuziastų gali eksperimentuoti su gyvybės genetiniu kodu. Negalima atmesti galimybės įvyksiant ne tokį akivaizdų biologinį pavojų keliantį įvykį, tokį, kuris, tarkime, smogtų tam tikrai populiacijai ar sabotuotų kurią nors ekosistemą. Įsivaizduokite, kad sustabdyti kokaino prekybą užsimanę aktyvistai sukuria naują vabalą, kuris kenktų tik kokos plantacijoms ir taip pakeistų iš oro purškiamas chemines medžia-

gas. Arba, tarkime, karingai nusiteikę veganai nusprendžia sutrikdyti visą mėsos tiekimo grandinę, sukeldami pragaištingų lauktų ir nelauktų padarinių. Bet kuris šių užmojų gali tapti nevaldomas.

Žinome, kokią žalą gali padaryti iš laboratorijos ištrūkęs virusas nuolat didėjančio silpnumo kontekste, tačiau nesuvaldytas jis prilygtų ankstesniems marams. COVID omikrono variantu ketvirtis amerikiečių užsikrėtė per šimtą dienų po pirmo nustatyto atvejo. O kas nutiktų, jeigu susidurtume su pandemija, kuri tokia pat lengvai užkrečiama, tačiau kurios mirtingumo rodiklis būtų, tarkime, 20 procentų? O kas, jeigu tai būtų kokia nors oru plintanti ŽIV atmaina, kuriai būdingas kelerių metų trukmės besimptomis inkubacinis laikotarpis? Naujoviškas žmonių platinamas virusas, kurio reprodukcijos dažnis, tarkime, 4 procentai (gerokai mažesnis už vėjaraupių ar tymų), o mirtingumo rodiklis 50 procentų (kur kas mažesnis už Ebolos ar paukščių gripo), per kelis mėnesius pražudytų per milijardą žmonių, netgi pritaikius karantino priemones. Kas nutiktų vienu metu paleidus keletą tokių patogenų? Tai būtų šis tas gerokai daugiau nei silpnumą skatinantis veiksnys – tai būtų neįsivaizduojama katastrofa.

•

Neskaitant Holivudo klišių, scenarijų, kuriame DI gali sukelti egzistencinę katastrofą, taip pat garsino akademinių mokslininkų subkultūra. Įsivaizduokite galingiausią mašiną, kuri kokiu nors būdu dėl savo paslaptingų tikslų sunaikina pasaulį – ne kokį piktavalį DI, visoje planetoje sėjantį chaosą, koks vaizduojamas filmuose, o BDI, aklai optimizuojantį, siekiantį neaiškaus žmonių gerovę ignoruojančio tikslo.

Štai vienas populiariausių mintinių eksperimentų: pakankamai galingam DI skyrus užduotį gaminti sąvaržėles, bet atidžiai nepatikslinus galutinio tikslo, ilgainiui jis gali sąvaržėlėmis paversti visą pasaulį ir netgi viso kosmoso turinį. Užtenka tik pradėti gal-

voti apie įvairias įmanomas baigtis, ir jau prikuriame aibę nerimą keliančių įvykių scenarijų. DI saugumą tyrinėjantys mokslininkai (pagrįstai) nerimauja, kad sukūrus ką nors panašaus į BDI, žmonija nebebus savo likimo kalvė. Tokiu atveju pirmą kartą per visą istoriją žmonės prarastų savo kaip dominuojančios rūšies vietą stebimoje visatoje. Kad ir kokių sumanių inžinierių turėtume, kad ir kokie galingi būtų mūsų apsaugos mechanizmai, pritaikyti visoms įmanomoms situacijoms, tačiau saugumo užtikrinti neįmanoma. Net jeigu jis būtų visiškai suderintas su žmonių interesais, pakankamai galingas DI potencialiai gali perrašyti savo programą ir nusikratyti neva įdiegtomis saugumo ir suderinamumo sistemomis.

Sekant tokia logika, dažnai girdžiu žmones švaistantis perspėjimais: „BDI kelia didžiausią mūsų laikų pavojų žmonijai! Jis sunaikins pasaulį!“ Tačiau prispausti nupiešti vaizdą, kaip atrodo toji pasaulio pabaiga, taip postringaujantys žmonės vengia atsakyti, pateikia neaiškius atsakymus, miglotai suvokia konkretų pavojų. Jų tvirtinimu, DI su visais matematinėms operacijoms skirtais ištekliais gali liautis klausęs žmonių ir visą pasaulį paversti milžinišku kompiuteriu. Kai DI tampa vis galingesnis, norint išvengti baisiausių scenarijų, žmonėms reikia gerai viską apsvarstyti ir sugalvoti, kaip galima sumažinti grėsmę. Tačiau iki tol gali nutikti aibė įvairių blogybių.

Per kitą dešimtmetį DI taps visų laikų didžiausiu galią stiprinančiu veiksniu. Štai kodėl jis gali sudaryti sąlygas istorinio masto galios perdalijimui. Didžiausias žmonių pažangos spartintuvas, kokį tik galima įsivaizduoti, dirbtinis intelektas, taip pat duos galimybę pridaryti žalos: ar kilus karams, nelaimingiems atsitikimams, ar juo pasinaudojus atsitiktinėms teroristinėms grupuotėms, autoritarinėms valdžioms, per didelę galią įgijusioms korporacijoms, paprasčiausioms vagystėms ar sąmoningam sabotažui atlikti. Tik pagalvokite apie PDI, gebantį be jokių sunkumų išlaikyti šiuolaikinį Turingo testą, bet naudojamą katastrofiškiems tikslams.

Svarbu nepamiršti, kad pažangūs dirbtiniai intelektai ir sintetinė biologija taps prieinami ne tik grupėms, ieškančioms naujų energijos šaltinių ar gyvenimą pakeisiančių vaistų, bet ir kitam Tedui Kaczynskiui*.

DI kartu ir vertingas, ir pavojingas būtent todėl, kad tai mūsų geriausių ir blogiausių ypatybių tąsa. Kadangi šios technologijos pagrindą sudaro mokymasis, ji gali adaptuoti, nagrinėti ir generuoti naujoviškas strategijas ir idėjas, kurios gali būti itin nutolusios nuo anksčiau mūsų ar netgi kitų DI svarstytų strategijų bei idėjų. Užtenka tik paprašyti pasiūlyti būdų, kaip sutrikdyti gėlo vandens tiekimą, sukelti akcijų biržos griūtį, branduolinį karą arba kaip sukurti pavojingiausią virusą, ir jis pateiks atsakymą. Nedelsdamas. Didesnį nerimą nei spekuliatyvūs sąvaržėlių maksimizuotojai ar koks nors keistas piktavalis demonas man kelia galimybė, kad per kitus dešimt metų šis įrankis didesnę galią suteiks jau egzistuojančioms jėgoms.

Įsivaizduokite scenarijų, kuriame dirbtiniai intelektai kontroliuoja energijos infrastruktūrą, medijos programas, elektros energijos stotis, lėktuvus ar pagrindinių finansinių įstaigų prekybos vertybiniais popieriais sąskaitas. Kaip atrodys įsilaužimas į skaitmeninę sistemą, surengtas kito DI, kai robotai bus tapę visuotini, o kariuomenės apsiginklavusios mirtinais autonominiais ginklais – turės pilnus sandėlius technologijų, galinčių vieno mygtuko paspaudimu imti autonomiškai vykdyti masines žudynes? Arba pagalvokite apie paprastesnes nesėkmes, ne išpuolius, bet paprasčiausias klaidas. Kas nutiks dirbtiniams intelektams padarius klaidą esminėse infrastruktūrose arba sugedus plačiai naudojamai medicinos sistemai? Nesunku įžvelgti, kaip daugybė pajėgių ir pusiau autonomiškų agentų, siautėjančių laisvėje, ar netgi tų, kurie

* Tedas Kaczynskis, dar žinomas kaip Unabomberis, Kalifornijos universitete Berklyje trumpai dėstęs matematikos vunderkindas, vėliau 1978–1995 m. įvairiems asmenims siuntęs savadarbes bombas. Jos pražudė 3 asmenis ir dar 23 sužeidė. Nusikaltimus vykdė motyvuojamas su technologijų pažanga ir gamtosauga susijusių idėjų.

siekia gerų, tačiau blogai suformuotų tikslų, gali imti sėti didžiulį chaosą. Be to, mes dar nežinome, kokių pasekmių DI turės įvairioms sritims kaip žemdirbystė, chemija, chirurgija ir finansai. Tai dalis problemos: nežinome, kokių klaidų DI gali pridaryti ir kaip toli gali siekti jų pasekmės.

Mes neturime vadovo, kuriame būtų nurodyta, kaip saugiai kurti kylančios bangos technologijas. Negalime kurti pavojingų, vis galingesnėmis tampančių sistemų, kuriomis atliktume įvairius bandymus, pirma jų nesukūrę saugių. Negalime numatyti, kaip greitai DI ims save tobulinti arba kas nutiks laboratorijoje įvykus nelaimingam atsitikimui su kokia nors dar neišrasta biotechnologija. Nežinome, kas būtų žmogaus sąmonę tiesiogiai prijungus prie kompiuterio arba kaip DI valdomas kibernetinis ginklas gali paveikti kritiškai svarbią infrastruktūrą, arba kokių rezultatų tikrovėje turės modifikuoti genai. Į pasaulį paleidus greitai tobulėjančius, save surenkančius automatonus ar naujus patogeninius mikroorganizmus, kelio atgal nebebus. Pasiekus tam tikrą tašką, pavojų užtraukti gali net smalsumas ar mėgėjiškas eksperimentavimas. Net jeigu manote katastrofos tikimybę esant nedidelę, vis dėlto faktas, kad mes veikiame aklai, turėtų priversti akimirką stabtelti.

Taip pat negana kurti saugias ir sulaikomas technologijas. Išspręsti DI suderinimo su žmonių interesais problemą nereiškia, kad užtenka tai padaryti tik kartą – šį derinimą reikės atlikti *kas kartą* kur nors ir kada nors sukūrus pakankamai galingą DI. Negana vienoje laboratorijoje išspręsti nutekėjimo problemą, ją reikės amžiais spręsti visose visų šalių laboratorijose, net kai tos šalys patirs didžiulį politinį spaudimą. Technologijoms pasiekus aukščiausią lygį negana, kad jų kūrėjai jas sukurtų saugias, kad ir kaip tai būtų sudėtinga. Saugumui užsitikrinti privalu visais atvejais laikytis saugumo standartų – tai milžiniškas lūkestis, atsižvelgiant į tai, kaip greitai ir plačiai jau dabar sklinda technologijos.

Taip nutinka, kai bet kas gali išrasti visą žmoniją paveikiančius įrankius ar jais naudotis. Ir aš kalbu ne vien apie priėjimą prie spausdinimo staklių ar garo variklio, kad ir kokie įspūdingi šie išradimai. Turiu omenyje visiškai naujo pobūdžio pajėgumus: naujus cheminius junginius, naujas gyvybės formas, rūšis.

Nesulaikius kylančios bangos, tik laiko klausimas, kada susidursime su galimu nelaimingu atsitikimu, klaida, piktiems tikslams panaudota technologija, žmonių nekontroliuojama evoliucija, įvairiausiais nenuspėjamais padariniais. Vienu ar kitu metu, vienu ar kitu pavidalu kažkur įvyks kokios nors technologijos gedimas. Ir jis nebus toks kaip Bopalo ar Černobylio katastrofos – jis palies visą pasaulį. Štai toks bus technologijų, sukurtų turint geriausių ketinimų, palikimas.

Tačiau yra žmonių, turinčių kitokių ketinimų.

KULTAI, LUNATIKAI IR SAVIŽUDIŠKOS VALSTYBĖS

Dažniausiai rizikos, kylančios iš tokių sričių kaip genetinėmis modifikacijomis grįsti tyrimai, yra sankcionuoto darbo, vedamo gerų tikslų, rezultatas. Kitaip tariant, jos yra milžiniški neigiami gerų ketinimų padariniai, nenumatytos pasekmės. Deja, dalį organizacijų motyvuoja visai priešingi dalykai.

Antai XX a. 9-ajame dešimtmetyje Japonijoje susiformavo apokaliptinis kultas „Aum Shinrikyo“ („Aukščiausioji tiesa“). Grupuotė susiformavo vyro, save vadinusio Shoko Asahara, jogos studijoje. Iš gyvenimu nusivylusių žmonių susidaręs kultas radikalizavosi, daugėjantys nariai įsikalė į galvą artėjant pasaulio pabaigą, kad jie vieni ją išgyvens ir kad jiems derėtų ją paspartinti. Asaharos kultas išaugo maždaug iki 40–60 tūkstančių narių. Lyderiui pavyko įkalbėti grupelę leitenantų panaudoti biologinius ir cheminius ginklus. Apskaičiuota, kad „Aum Shinrikyo“ populiarumui pasiekus piką kultas valdė daugiau nei 1 milijardo dolerių vertės turtą ir kad jam