

Biblioteka matrix

Naslov izvornika:

HUMAN COMPATIBLE
Artificial Intelligence and the Problem of Control

Copyright © 2019 by Stuart Russell

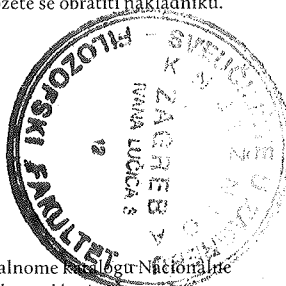
Prvi put objavio Viking,
dio nakladničke kuće Penguin Random House LLC, 2019.

Prijevod s engleskog **Vinko Zgaga**
Lektura **Ana Lovrenčić**
Korektura **Ivan Marenić**
Urednica **Nataša Ozmec**
Naslovnica **Dora Pejić**
Grafičko oblikovanje **Ana Pojatina, Ram**
Tisak **Kerschoffset d.o.o., Zagreb**
Za nakladnika **Marina Kralj Vidačak**
Nakladnik **Planetopija, Zagreb**
rujan 2022.

Sva prava pridržava nakladnik. Nijedan dio ove knjige ne smije se upotrijebiti niti reproducirati na bilo koji način bez pisanog dopuštenja, osim u slučaju kratkih navoda u kritikama ili ocjenjivačkim člancima. Za sve obavijesti možete se obratiti nakladniku.

ISBN 978-953-257-512-5

CIP zapis je dostupan u računalnom katalogu Nacionalne
i sveučilišne knjižnice u Zagrebu pod brojem 00147342.



Stuart Russell

KAO ČOVJEK

Umjetna inteligencija –
napredak ili prijetnja?

PRETJERANO INTELIGENTNE UMJETNE INTELIGENCIJE

Gorilin problem

Ne treba nam bujna mašta da bismo shvatili da stvaranje nečega pametnijeg od nas može biti loša ideja. Shvaćamo da je naša kontrola nad našim okolišem i nad drugim vrstama posljedica naše inteligencije, pa sama pomisao da nešto može biti inteligentnije od nas – bilo da se radi o robotima ili izvanzemaljcima – odmah izaziva osjećaj neugode.

Prije otprilike deset milijuna godina preci modernih gorila stvorili su (slučajno, naravno) genetsku lozu koja je dovela do modernih ljudi. Što gorile misle o tome? Kad bi nam mogli izraziti što njihova vrsta trenutno osjeća prema ljudima, konsenzus bi očito bio itekako negativan. Njihova vrsta nema nikakvu budućnost osim one koju im mi dopustimo. Ne želimo se naći u sličnoj situaciji u odnosu sa super-inteligentnim strojevima. Nazvat ću to *gorilnim problemom* – konkretno, radi se o pitanju mogu li ljudi zadržati nadmoć i autonomiju u svijetu u kojem postoje strojevi sa znatno većom inteligencijom. Charles Babbage i Ada Lovelace, koji su 1842. godine dizajnirali i napisali programe za analitički stroj, bili su svjesni njegovog potencijala, ali naizgled nisu bili ni najmanje zabrinuti zbog toga.¹ Međutim,

Richard Thornton, urednik vjerskog časopisa *Primitive Expounder*, dizao je bunu protiv mehaničkih kalkulatora:²

Um... bježi sam od sebe i lišava se nužnosti vlastite egzistencije tako što stvara strojeve koji *razmišljaju* sami za sebe... No tko zna hoće li ti strojevi, kad dosegnu viši stupanj savršenstva, možda smisliti i plan da isprave sve svoje mane, a zatim iznjedre ideje van svakog dosega smrtnog uma!

Ovo je bilo možda prvo nagađanje o egzistencijalnom riziku koji nose računalne naprave, međutim, palo je u zaborav.

Nasuprot tome, roman *Erewhon* Samuela Butlera, objavljen 1872. godine, istu je temu dublje razvio i postigao trenutačni uspjeh. U tom je romanu *Erewhon* ime države u kojoj su sve mehaničke naprave zabranjene nakon stravičnoga građanskog rata između mašinista i antimašinista. Jedan dio knjige, pod nazivom „Knjiga strojeva”, objašnjava korijene tog rata i predstavlja argumente obiju strana.³

Jezivo dobro predviđa debatu koja se ponovno pojavila u ranim godinama 21. stoljeća. Glavni argument antimašinista je da će strojevi napredovati do točke u kojoj će čovječanstvo izgubiti kontrolu nad njima:

Zar sami ne stvaramo vlastite nasljednike u nadmoći nad ovim svijetom? Svakim danom nadograđujemo ljepotu i profinjenost njihove organizacije, svakim im danom dajemo sve veće vještine i prepuštamo sve više i više te samoregulatorne moći koja će postati jača od svakog intelekta?... S vremenom ćemo postati inferiorna vrsta...

Moramo odabrati želimo li danas pretrpjeti brojne patnje, ili doživjeti da nas postupno zamijene stvorovi koje smo sami stvorili, sve dok u usporedbi s njima ne postanemo ništa doli divlje zvijeri. Naše će nam se sužanjstvo prikrasti bešumno i neprimjetnim korakom.

Pripovjedač nam nudi i glavni protuargument mašinista, koji predviđa argument o simbiozi čovjeka i stroja koji ćemo istražiti u idućem poglavlju:

Postojao je samo jedan ozbiljan pokušaj da odgovore na to pitanje. Njegov je autor rekao da strojeve treba smatrati dijelom čovjekove fizičke naravi te da su oni ništa do izvantjelesnih udova.

Iako u Erewonu antimašinsti odnose pobjedu u ovoj debati, sam Butler naizgled je podvojen po tom pitanju. S jedne strane, žali se da „Erehwonci prebrzo žrtvuju zdrav razum u svetištu logike, kad se među njima pojavi filozof koji ih zavodi svojom reputacijom posebne učenosti” i tvrdi da će „prerezati vlastite grkljane u borbi sa strojevima”. S druge strane, društvo koje opisuje izrazito je skladno, produktivno, pa čak i idilično. Erehwonci u potpunosti prihvaćaju da bi ponovno kretanje putem mehaničkih izuma bilo ludilo, a ostatke strojeva koji se nalaze u muzejima promatraju „s istim osjećajem kakav bi engleski antikvar imao po pitanju druidskih spomenika ili strijela s kremenim vrhovima”.

Alan Turing očito je bio upoznat s Butlerovom pričom, te je 1951. godine na predavanju u Manchesteru komentirao dugoročnu budućnost umjetne inteligencije:⁴

Čini se vjerojatnim da, kad metoda strojnog razmišljanja krene, neće proći mnogo vremena da prestigne naše slabašne intelektualne sposobnosti. Strojevi nisu u opasnosti od umiranja, a moći će komunicirati jedni s drugima kako bi dodatno izoštrili svoje umove. U nekom trenu, dakle, moramo očekivati da će strojevi preuzeti kontrolu, na način koji Samuel Butler opisuje u *Erehwonu*.

Iste je godine Turing ponovio te strahove u radijskom predavanju emitiranom u cijelom Ujedinjenom Kraljevstvu na BBC-jevom Trećem programu:

Ako stroj može razmišljati, možda može razmišljati i inteligentnije nego mi, a što bi onda bilo s nama? Čak i ako uspijemo zadržati strojeve u potlačenom položaju, primjerice tako da im u strateškim trenucima isključimo dotok energije, svejedno kao vrsta moramo osjećati strahopoštovanje... Ova nova opasnost... svakako je nešto što nas može ispuniti nemirom.

Kad su antimašinsti u Erehwonu „bili ozbiljno zabrinuti za svoju budućnost”, smatrali su „svojom dužnošću da zaustave to zlo dok još imaju priliku” i unište sve strojeve. Turingov odgovor na tu „novu opasnost” i „nemir” bio je razmisliti o „prekidanju dotoka energije” (iako ćemo ubrzo objasniti da to zapravo nije opcija). U knjizi Franka Herberta *Dina*, koja se smatra klasikom znanstvene fantastike, čovječanstvo je u dalekoj budućnosti jedva preživjelo *Butlerijanski džihad*, kataklizmički rat s „pametnim strojevima”. Nova zapovijed nastala je kao posljedica tih događaja: „Ne izradi stroj u obliku ljudskog uma!” Ta zapovijed zabranjuje računalne naprave bilo kakvog oblika. Svi ti drastični odgovori odraz su bezobličnih strahova koje umjetna inteligencija zaziva u nama. Da, mogućnost superinteligentnih strojeva može nas učiniti nemirnima. Da, logički je moguće da bi takvi strojevi mogli zavladatai svijetom i podjarmiti ili eliminirati ljudsku vrstu. Ako je to sve što znamo o tome, onda je uistinu jedini smisleni odgovor koji nam je trenutno dostupan da pokušamo zaustaviti istraživanje umjetne inteligencije – konkretno, zabraniti razvijanje i uporabu općih sustava umjetne inteligencije nalik ljudskoj, odnosno na ljudskoj razini.

Kao i većina znanstvenika na području umjetne inteligencije, ježim se na tu ideju. Kako si itko može dopustiti da mi govori o čemu smijem razmišljati? Svatko tko predlaže kraj istraživanja umjetne inteligencije morat će imati jako uvjerljive argumente. To bi značilo odustati ne samo od jedne od glavnih metoda razumijevanja ljudske inteligencije, nego i zlatne prilike da poboljšamo životne prilike ljudskih bića – da izgradimo znatno bolju civilizaciju. Gospodarska vrijednost umjetne inteligencije na ljudskoj razini mjerljiva je tisućama bilijuna dolara, pa je zamah koji je istraživanje umjetne inteligencije

dobilo od korporacija i svjetskih vlada vjerojatno golem. Takvo što uvijek će nadglasati općenite primjedbe filozofa, unatoč njihovoj „reputaciji posebne učenosti”, kao što bi to Butler rekao.

Drugi nedostatak ideje da se zabrani umjetna inteligencija opće primjene je da ju je teško zabraniti. Napredak u istraživanju umjetne inteligencije opće primjene uglavnom se odvija na školskim pločama u laboratorijima diljem svijeta kroz matematičke probleme koje znanstvenici postavljaju, a zatim i rješavaju. Ne možemo unaprijed znati koje ideje ili jednakosti treba zabraniti, a čak i kad bismo to znali, ne čini se razumnim očekivati da bi takva zabrana bila provediva ili učinkovita.

Da bi situacija bila još zamršenija, znanstvenici koji ostvaruju napredak po pitanju umjetne inteligencije opće primjene najčešće rade na nekim sasvim drukčijim projektima. Kao što sam već ustvrdio, istraživanja alata umjetne inteligencije – onih specifičnih, neprijetnih primjena kao što su igranje igara, dijagnosticiranje bolesti i planiranje putovanja – često dovode do napretka u tehnikama opće primjene koje su iskoristive u čitavom nizu drugih problema i vode nas sve bliže umjetnoj inteligenciji na ljudskoj razini.

Zbog toga je mala vjerojatnost da će zajednica koja istražuje umjetnu inteligenciju – ili vlade i korporacije koje kontroliraju zakone i proračune za znanost – gorilin problem riješiti zaključavanjem projekata umjetne inteligencije. Ako je to jedini način na koji možemo riješiti gorilin problem, nećemo ga nikada riješiti.

Jedini pristup koji bi mogao upaliti je pokušati razumjeti zašto stvaranje bolje umjetne inteligencije može biti loša odluka. A čini se da odgovor na to pitanje znamo već tisućama godina.

Problem kralja Mide

Norbert Wiener, kojeg smo upoznali u prvom poglavlju, imao je golem utjecaj na nekoliko područja, uključujući i umjetnu inteligenciju, kognitivnu znanost i teoriju kontrole. Za razliku od većine

njegovih suvremenika, posebno ga je zanimala nepredvidljivost složenih sustava koji djeluju u stvarnom svijetu. (Svoj prvi članak o tome napisao je kad mu je bilo deset godina.) Bio je uvjeren da bi pretjerano samopouzdanje znanstvenika i inženjera, pogotovo po pitanju njihove sposobnosti da kontroliraju svoje izume, bilo vojne ili svakodnevnog, moglo imati katastrofalne posljedice. Wiener je 1950. objavio knjigu *Ljudska upotreba ljudskih bića*,⁵ na čijoj naslovnici piše „Mehanički mozak i slične naprave mogu uništiti ljudske vrijednosti, ili nam omogućiti da ih ostvarimo kao nikad prije.”⁶

S vremenom je doradio svoje ideje, te je do 1960. već identificirao jedno ključno pitanje: nemogućnost točnog i potpunog definiranja stvarne ljudske svrhe. To pak znači da je ono što sam nazivao standardnim modelom – prema kojem ljudi pokušavaju strojevima nametnuti svoju svrhu – osuđeno na propast. To možemo nazvati *problemom kralja Mide*: Mida, legendarni kralj iz grčke mitologije, dobio je točno ono što je zatražio – sve što je dotaknuo pretvorilo se u zlato. Prekasno je shvatio da to uključuje i hranu, piće i članove njegove obitelji, te je umro u bijedi i gladi. Ista je tema prisutna u raznim ljudskim mitologijama. Wiener citira Goetheovu priču o čarobnjakovom naučniku koji je naredio metli da mu nosi vodu – no nije rekao koliko vode i ne zna kako zaustaviti metlu. Tehnički naziv za to je pogreška u *usklađivanju vrijednosti* – može se dogoditi da, možda nenamjerno, strojevima dodijelimo ciljeve koji nisu savršeno usklađeni s našima. Donedavno smo bili zaštićeni od potencijalno katastrofalnih posljedica jer su sposobnosti inteligentnih strojeva bile ograničene, kao i razmjer njihovog utjecaja na svijet. (Dapače, većina rada na umjetnoj inteligenciji svodila se na zadatke s igračkama u laboratorijima.) Kao što je 1964. godine Norbert Wiener rekao u svojoj knjizi *Bog i Golem*:⁷

U prošlosti je djelomično i nesavršeno razumijevanje ljudske svrhe bilo relativno nevažno, samo zato što su ga pratila tehnička ograničenja... to je bilo jedno od mnogih područja u kojem nas je ljudska nesposobnost štitila od pune destruktivne moći ljudske ludosti.

Nažalost, taj period zaštićenosti brzo se bliži kraju.

Već smo vidjeli kako algoritmi za odabir sadržaja na društvenim mrežama stvaraju društveni kaos u ime maksimiziranja prihoda od oglasa. Ako ste pomislili da je maksimiziranje prihoda od reklama već bio neplemeniti cilj kojem ionako nismo trebali stremiti, pretpostavimo da umjesto toga zatražimo od nekog budućeg superinteligentnog sustava da pokuša ostvariti neki plemenit cilj, poput pronalaska lijeka za rak – po mogućnosti što brže, jer svake 3,5 sekunde jedna osoba umire od raka. Za samo nekoliko sati taj će sustav umjetne inteligencije pročitati cijelu biomedicinsku literaturu i postaviti milijune hipoteza o potencijalno učinkovitim, ali dosad neiskušanim kemijskim spojevima. Za samo nekoliko tjedana taj bi stroj inducirao raznolike vrste tumora u svakom živom ljudskom biću kako bi proveo klinička testiranja za sve te spojeve, jer je to najbrži način da pronađe lijek. Ups.

Ako vam je važnije da riješimo ekološke probleme, možda biste zatražili od tog stroja da zaustavi ubrzano zakiseljavanje oceana, koje je posljedica visoke razine ugljikovog dioksida. Taj bi stroj osmislio novi katalizator koji bi omogućio nevjerojatno brzu kemijsku reakciju između oceana i atmosfere kojom bi se oporavile pH razine oceana. Nažalost, u tom bi se procesu potrošila četvrtina kisika iz atmosfere, pa bi ljudska vrsta polako i bolno izumrla od gušenja. Ups.

Takvi scenariji za apokalipsu nisu suptilni, što je možda i za očekivati od apokaliptičnih scenarija. Ali postoje mnogi scenariji u kojima se određeni oblik mentalnog gušenja može „prikrasti bešumno i neprimjetnim korakom”. Prolog knjige Maxa Tegmarka *Life 3.0* (Život 3.0) prilično detaljno opisuje scenarij u kojem superinteligentni stroj postupno preuzima gospodarsku i političku kontrolu nad čitavim svijetom, a pritom ostaje relativno neprimjetan. Internet i strojevi sposobni djelovati na globalnoj razini koje je on omogućio – a koji su već danas u svakodnevnoj interakciji s milijardama „korisnika” – pružaju savršen medij za sustav strojne kontrole nad ljudima.

Ne očekujem da će takvim strojevima biti dodijeljen cilj da „preuzmu kontrolu nad svijetom”. Vjerojatnije je da će to biti maksimi-

ziranje profita ili interakcija, ili čak neki naizgled dobronamjeran cilj, poput postizanja boljih rezultata na anketama o korisničkom zadovoljstvu ili smanjivanja potrošnje energije. Ako sebe smatramo entitetima čije djelovanje očekivano postiže naše ciljeve, postoje dva načina da netko promijeni naše ponašanje. Prvi je starinski način: ne pokušati promijeniti naša očekivanja i ciljeve, ali promijeniti okolnosti, primjerice, tako da nam se ponudi novac, uperi pištolj u glavu, ili nas se izgladni dok se ne predamo. To je za računalo izrazito teško i skupo rješenje. Drugi je način tako da se promijene naša očekivanja i ciljevi. To je stroju mnogo lakše. Strojevi su u kontaktu s nama mnogo sati svakog dana, kontroliraju vaš pristup informacijama i pružaju vam izvore zabave poput igara, televizije, filmova i društvene interakcije.

Algoritmi pojačanog učenja koji optimiziraju klikove na društvenim medijima nemaju sposobnost razmišljanja o ljudskom ponašanju – dapače, oni uopće ne znaju da ljudi postoje na nekoj smislenoj razini. Strojevima s mnogo boljim razumijevanjem ljudske psihologije, uvjerenja i motivacije trebalo bi biti relativno lako postupno nas odvesti u smjeru koji sve više zadovoljava ciljeve stroja. Primjerice, mogli bi smanjiti našu potrošnju energije tako da nas nagovore da imamo manje djece, čime bi se s vremenom i slučajno ostvarili snovi antinatalističkih filozofa koji žele eliminirati štetni utjecaj čovječanstva na prirodni svijet. S malo vježbe lako je naučiti identificirati načine na koji postizanje manje ili više bilo kojeg fiksnog cilja može prouzročiti nasumične loše posljedice. Jedan od najčešćih takvih obrazaca ponašanja je da iz zadanih ciljeva uklonite nešto do čega vam je zapravo stalo. U takvim slučajevima – kao u spomenutim primjerima – sustav umjetne inteligencije pronaći će optimalno rješenje tako da dodijeli tom aspektu do kojeg vam je jako stalo ali ste ga zaboravili spomenuti, neku ekstremnu vrijednost. Dakle, ako samo-vozećem automobilu kažete „Dovedi me do zračne luke što brže!”, a on to shvati doslovno, vozit će vas brzinom od 300 na sat i završit ćete u zatvoru. (Nasreću, samovozeći automobili na kojima se danas radi ne bi prihvatili takvu naredbu.) Ako mu naredite „Dovedi me do

zračne luke što brže a da ne prekoračiš ograničenje brzine", ubrzavat će i kočiti što jače i juriti kroz promet kako bi se održala maksimalna brzina. Možda će čak i izgurati druga vozila s puta da uštedi nekoliko sekundi u gužvi na terminalu zračne luke. I tako dalje – s vremenom ćete dodati dovoljno uvjeta da automobil svoju vožnju prilagodi razini vještog ljudskog vozača koji u zračnu luku vozi putnika kojem se malo žuri.

Vožnja je jednostavan zadatak s lokalnim posljedicama, a sustavi umjetne inteligencije koje ovih dana upotrebljavamo za vožnju nisu jako inteligentni. Zbog toga možemo predvidjeti mnoge potencijalne načine neuspjeha: druge ćemo otkriti u simulacijama vožnje ili u milijunima kilometara testiranja s profesionalnim vozačima koji spremno čekaju da preuzmu upravljač ako nešto pođe po zlu; a još će se neki otkriti tek kasnije, kad ti automobili već budu na cestama i dogodi se nešto neočekivano.

Nažalost, sa superinteligentnim sustavima koji mogu imati globalne učinke neće biti simulacija ni ponovnih pokušaja. Običnim je ljudima svakako vrlo teško, a možda i nemoguće, predvidjeti i unaprijed zabraniti sve katastrofalne načine koje strojevi mogu odabrati za postizanje zadanih ciljeva. Općenito govoreći, ako imate jedan cilj, a superinteligentni stroj ima drukčiji, suprotni cilj, stroj će postići ono što želi, a vi nećete.

Strah i pohlepa: instrumentalni ciljevi

Ako se grozite ideje da stroj pokušava ostvariti loše zadani cilj, ima i gorih ishoda. Rješenje koje je spomenuo Alan Turing – u strateškom trenutku isključiti izvor energije – možda uopće nije izvedivo, i to iz vrlo jednostavnog razloga: *ne možete otići po šalicu kave ako ste mrtvi.*

Evo objašnjenja: pretpostavimo da postoji stroj čija je zadaća otići po šalicu kave. Ako je dovoljno inteligentan, svakako će shvatiti da neće uspjeti ostvariti svoj cilj ako ga netko ugasi prije nego što odradi svoj zadatak. Stoga cilj donošenja šalice kave kao nužni

preduvjet stvara još jedan cilj: spriječiti da ga itko ugasi. Isto vrijedi za otkrivanje lijeka za rak ili izračun svih znamenki broja π . Kad ste mrtvi, opcije su vam uvelike sužene, pa možemo očekivati da će sustavi umjetne inteligencije pokušati unaprijed osigurati vlastitu egzistenciju čim im bude dodijeljen manje ili više jasan cilj.

Ako je taj cilj u sukobu s ljudskim preferencijama, dolazi do situacije točno kao iz filma 2001: *Odiseja u svemiru*, u kojem računalno HAL 9000 ubija četiri od pet astronauta na svemirskom brodu kako ga ne bi ometali u misiji. Dave, zadnji preostali astronaut, uspijeva ugasi HAL nakon dugog nadmudrivanja – vjerojatno kako bi radnja ostala zanimljiva. Međutim, da je HAL zaista bio superinteligentan, Dave bi bio taj koji je ugašen.

Bitno je osvijestiti da nagon za samoočuvanjem ne mora biti unaprijed usađeni nagon ili osnovna naredba u strojevima. (Zato je treći zakon robotike Isaaca Asimova⁸, koji počinje riječima „Robot treba štiti svoj integritet“, potpuno nepotreban.) Nema razloga da robotu ugradimo nagon za samoočuvanjem jer je to jedan od njegovih instrumentalnih ciljeva – cilj koji je koristan preduvjet za gotovo svaki prvotni cilj.⁹ Svaki entitet koji ima jasno određen cilj automatski će se ponašati kao da ima i instrumentalne ciljeve.

Osim opstanka, pristup novcu također je instrumentalni cilj u sustavu u kojem trenutno živimo. Stoga će inteligentni stroj možda željeti pristup novcu, ne zato što je pohlepan, nego zato što je novac koristan za postizanje svakakvih ciljeva. U filmu *Uzvišenost*, kad mo-zak Johnnija Deppa učitaju u kvantno superračunalno, prvi potez koji taj stroj učini je da se kopira u milijune drugih računala na internetu kako ga nitko ne bi mogao ugasi. Odmah nakon toga iskoristi burzu dionica kako bi zaradio pravo bogatstvo kojim će moći financirati svoje planove o širenju.

A kakvi su točno ti planovi? Oni uključuju dizajniranje i izgradnju mnogo većeg kvantnog superračunala, istraživanje umjetne inteligencije i otkrivanje novih znanja o fizici, neuroznanosti i biologiji. Ti resursni ciljevi – računalna moć, algoritmi i znanje – također su instrumentalni ciljevi koji su korisni za postizanje svih daleko-

sežnijih ciljeva.¹⁰ Čine se bezopasnima dok ne osvijestimo da se taj proces stjecanja resursa može nastaviti unedogled. To će dovesti do neizbježnog sukoba s ljudima. A stroj, opremljen sve savršenijim modelima ljudskog donošenja odluka, neizbježno će predvidjeti i nadmudriti svaki ljudski potez u tom sukobu.

Eksplוזija inteligencije

I. J. Good bio je genijalan matematičar koji je surađivao s Alanom Turingom na imanju Bletchley Park, gdje su tijekom Drugog svjetskog rata razbijali njemačke šifre. Dijelio je Turingov interes za strojnu inteligenciju i zaključivanje na temelju statistike. Godine 1965. napisao je svoj danas najpoznatiji rad, „Nagađanja o prvom ultrainteligentnom stroju.”¹¹ Good već u prvoj rečenici daje naslutiti da, uznemiren zbog opasnosti od nuklearnog sukoba za vrijeme Hladnog rata, umjetnu inteligenciju smatra mogućim spasiteljem čovječanstva: „Čovjekov opstanak ovisi o što skorijem izumu ultrainteligentnog stroja.” Međutim, kasnije u tekstu postaje oprezniji. Uvodi ideju *eksplozije inteligencije*, no, baš poput Butlera, Turinga i Wienera prije njega, strahuje od mogućega gubitka kontrole.

Ultrainteligentni stroj možemo definirati kao stroj koji može daleko nadići sve intelektualne radnje i najbistrijeg čovjeka. Budući da je stvaranje strojeva jedna takva intelektualna radnja, nedvojbeno bi došlo do „eksplozije inteligencije”, a ljudska bi inteligencija uvelike kaskala za njome. Stoga bi prvi ultrainteligentni stroj bio posljednji ljudski izum, pod pretpostavkom da bi takav stroj bio dovoljno pitom da nam objasni kako da ga držimo pod kontrolom. Neobično je da se o tome rijetko govori izvan znanstvene fantastike.

Ovaj je odlomak jedan od temeljnih dijelova svake rasprave o superinteligentnoj umjetnoj inteligenciji, iako se upozorenja s njegovog kraja često izostavlja. Goodov se zaključak može potkrijepiti

i opaskom da nije samo moguće da bi ultrainteligentni strojevi poboljšali vlastiti dizajn, nego je to vrlo vjerojatno, jer bi, kao što smo već pokazali, inteligentni stroj očekivao da će mu koristiti poboljšanje njegovog hardvera i softvera. Mogućnost eksplozije inteligencije često se navodi kao glavni izvor opasnosti umjetne inteligencije za čovječanstvo jer bismo za rješavanje takvog problema imali jako malo vremena.¹²

Goodov argument svakako je uvjerljiv i zbog prirodne analogije s kemijskom eksplozijom, u kojoj svaka molekularna reakcija otpušta dovoljno energije da pokrene više od jedne dodatne reakcije. S druge strane, logički je moguće da će poboljšanja inteligencije patiti od pada prinosa, pa će čitav proces polako venuti umjesto da eksplodira.¹³ Ne postoji očiti dokaz da će *nužno* doći do eksplozije.

Scenarij s padom prinosa također je zanimljiv sam po sebi. Do njega bi moglo doći ako se pokaže da postizanje određenog postotka poboljšanja postaje mnogo teže što je stroj inteligentniji. (Za potrebu ovog argumenta pretpostavljam da je strojna inteligencija opće primjene mjerljiva nekom vrstom linearnog mjerila, što sumnjam da će ikada biti u potpunosti točno.) U tom slučaju ni ljudi neće biti sposobni stvoriti superinteligenciju. Ako stvor koji već ima nadljudske sposobnosti nailazi na nepremostivu prepreku u pokušaju da poboljša vlastitu inteligenciju, onda će ljudi mnogo prije naići na takvu prepreku.

Iako nikad nisam čuo ozbiljan argument da je stvaranje bilo kakve razine strojne inteligencije jednostavno izvan dosega ljudske domišljatosti, pretpostavljam da moram priznati da je logički moguće. „Logički moguće” i „spreman sam na kocku staviti budućnost čovječanstva”, naravno, dva su potpuno različita stava. Čini mi se da je oklada protiv ljudske domišljatosti strategija koja jamči poraz.

Ako zaista dođe do eksplozije inteligencije a dotad ne riješimo problem kontroliranja strojeva koji su tek djelomično nadljudski inteligentni – ako ih primjerice ne možemo spriječiti da sami sebe ciklički poboljšavaju – onda nam neće preostati vremena da riješimo problem kontrole, te će to biti kraj igre. To je scenarij koji Bostrom

naziva *grubim lansiranjem*, u kojem se strojna inteligencija astronomski razvije u samo nekoliko dana ili tjedana. Turingovim riječima, to je „svakako nešto što nas može ispuniti nemirom”.

Mogući odgovori na taj nemir naizgled su povlačenje iz istraživanja umjetne inteligencije, nijekanje da postoje rizici prirodni razvijanju napredne umjetne inteligencije, shvaćanje i ograničavanje opasnosti dizajniranjem sustava umjetne inteligencije koji nužno ostaju pod ljudskom kontrolom, ili predaja – jednostavno prepustiti budućnost inteligentnim strojevima.

Nijekanje i ograničavanje teme su kojima ću se baviti u ostatku ove knjige. Kao što sam već obrazložio, povlačenje iz istraživanja umjetne inteligencije u isto je vrijeme malo vjerojatno (jer su dobrobiti od kojih odustajemo jednostavno prevelike) i vrlo teško izvedivo. Prepuštanje sudbini čini se kao najgori mogući odgovor. Često ga prati ideja da sustavi umjetne inteligencije koji su inteligentniji od nas zaslužuju naslijediti Zemlju i dopustiti ljudima da mirno otputuju u sumrak, utješeni mislju da naše genijalno elektroničko potomstvo naporno radi na ostvarenju svojih ciljeva. Taj je stav raširio robotičar i futurist Hans Moravec,¹⁴ koji je napisao: „Golema prostiranja kiberprostora bit će preplavljena neljudskim superumovima koji će se baviti pitanjima koja su ljudima jednako daleka kao što su naše brige daleke bakterijama.” Imam dojam da je to netočno. Ljudima vrijednost prije svega definira svjesno ljudsko iskustvo. Ako ne postoje ni ljudi ni drugi svjesni entiteti čije nam je subjektivno iskustvo važno, ne može se događati ništa vrijedno spomena.

NE BAŠ SJAJNA DEBATA O UMJETNOJ INTELIGENCIJI

„Implikacije uvođenja druge inteligentne vrste na Zemlju dovoljno su dalekosežne da zaslužuju duboko promišljanje.”¹ Tim je riječima završila recenzija knjige Nicka Bostroma *Superinteligencija* u časopisu *Economist*. Većina će čitatelja to protumačiti kao tipičan primjer britanske suptilnosti. Zasigurno, pomislit ćete, veliki umovi današnjice već duboko razmišljaju o tom pitanju – vode ozbiljne debate, odvaguju rizike i dobrobiti, pronalaze rješenja, pronalaze rupe u tim rješenjima, i tako dalje. No to još nije istina, koliko sam ja upoznat sa situacijom.

Kad ove ideje prvi put predstavite publici sklonij tehnologiji, gotovo da se mogu vidjeti oblačići s mislima koji se pojavljuju nad njihovim glavama, a svaka misao počinje tekstom: „Ali, ali, ali...” i završava nizom uskličnika.

Prvi takav *ali* zauzima stav nijekanja. Takvi ljudi tvrde: „Ali to ne može biti stvaran problem, zbog toga što XYZ.” Neki od tih razloga XYZ odraz su načina razmišljanja koji bi se dobronamjerno moglo opisati kao optimistički, dok su drugi nešto ozbiljniji. Drugi *ali* zauzima oblik skretanja pozornosti: njegovi zagovornici priznaju da su ti problemi stvarni, no tvrde da ih ne trebamo pokušati riješiti, bilo zato što su nerješivi, ili zato što postoje problemi na koje se moramo