

Biblioteka matrix

Naslov izvornika:

HUMAN COMPATIBLE
Artificial Intelligence and the Problem of Control

Copyright © 2019 by Stuart Russell

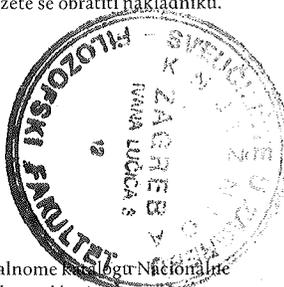
Prvi put objavio Viking,
dio nakladničke kuće Penguin Random House LLC, 2019.

Prijevod s engleskog **Vinko Zgaga**
Lektura **Ana Lovrenčić**
Korektura **Ivan Marenčić**
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 računalnome katalogu Nacionalne
i sveučilišne knjižnice u Zagrebu pod brojem 00147342.



Stuart Russell

KAO ČOVJEK

Umjetna inteligencija –
napredak ili prijetnja?

UMJETNA INTELIGENCIJA: DRUKČIJI PRISTUP

Kad pobijemo sve argumente skeptika i odgovorimo na sve *ali ali ali* prigovore, iduće je pitanje najčešće: „U redu, priznajem da imamo problem, ali ne postoji rješenje, zar ne?“ Da, postoji rješenje.

Podsjetimo se samo koji je zadatak pred nama: stvoriti strojeve s visokim stupnjem inteligencije kako bi nam mogli pomoći s teškim problemima, a u isto vrijeme osigurati da se ti strojevi nikad neće ponašati na način koji bi nas mogao ozbiljno unesrećiti.

Zadatak, nasreću, nije sljedeći: suočeni sa strojem s visokim stupnjem inteligencije, trebamo smisliti kako ga kontrolirati. Da je to zadatak, ne bi nam bilo spasa. Ako takav stroj smatramo crnom kutijom, gotovim činom, bilo bi to isto kao da se stvorio negdje iz svemira. A izgledi da ćemo moći kontrolirati superinteligentan entitet iz svemira otprilike su ravni nuli. Slične se argumente može primijeniti na metode stvaranja sustava umjetne inteligencije za koje nećemo biti sigurni kako rade: takve metode uključuju *emuliranje cijelog mozga*¹ – stvaranje poboljšanih elektroničkih kopija ljudskih mozgova – kao i metode zasnovane na simuliranoj evoluciji programa.² Neću više tritati vrijeme na takve prijedloge jer su to očito loše ideje.

Kako je, dakle, područje umjetne inteligencije u prošlosti pristupalo „stvaranju strojeva s visokim stupnjem inteligencije“? Baš poput

mnogih drugih područja, u istraživanju umjetne inteligencije prigrlili smo standardni model: gradimo strojeve za optimizaciju, u njih ubacujemo ciljeve i puštamo ih da rade. To je radilo prilično dobro dok su strojevi bili glupi i imali ograničen spektar radnji. Ako bismo im dodijelili krivi cilj, vjerojatno bismo dobili priliku da ugasimo taj stroj, ispravimo problem i počnemo iznova. Međutim, kako strojevi dizajnirani prema standardnom modelu postaju sve inteligentniji i kako spektar dostupnih akcija postaje globalan, tako će i taj pristup biti sve manje održiv. Takvi će strojevi stremiti ka svojim ciljevima, ma koliko pogrešni oni bili; odupirat će se pokušajima da ih ugasimo i prikupljat će sve moguće resurse koji mogu pridonijeti ostvarivanju tog cilja. Dapače, optimalno ponašanje za takav stroj moglo bi biti i obmanjivanje ljudi i uvjeravanje da su stroju dodijelili razuman cilj, kako bi stroj dobio na vremenu i postigao cilj koji mu je zapravo dodijeljen. To ne bi bilo „devijantno” ili „zlobno” ponašanje koje zahtijeva svijest i slobodnu volju: to bi samo bio dio optimalnog plana za postizanje zadanih ciljeva.

U prvom poglavlju uveo sam ideju blagotvornih strojeva – to jest strojeva čije će djelovanje postizati naše ciljeve, a ne njihove. Moj je cilj u ovom poglavlju na jednostavan način objasniti kako se to može postići unatoč prepreci da strojevi naizgled ne znaju koji su naši ciljevi. Pristup koji iz toga proizlazi s vremenom bi trebao dovesti do strojeva koji nam ne predstavljaju nikakvu prijetnju ma kako inteligentni bili.

Principi za blagotvorne strojeve

Smatram da je korisno sažeti taj pristup u obliku tri³ principa. Imajte na umu da su zamišljeni primarno kao misao vodilja za znanstvenike i programere na području umjetne inteligencije koji razmišljaju o tome kako stvoriti blagotvorne sustave umjetne inteligencije; oni nisu zamišljeni da budu eksplicitni zakoni koje će sustavi umjetne inteligencije morati poštovati:⁴

1. Jedini cilj stroja je maksimizirati ostvarivanje ljudskih preferencija.
2. Stroj isprva nije siguran kakve su to preferencije.
3. Najbolji izvor informacija o ljudskim preferencijama je ljudsko ponašanje.

Prije nego što zaronimo u detaljno objašnjenje, važno je sjetiti se širokog značenja riječi „preferencija” za potrebe tih principa. Podsjetimo se što sam napisao u drugom poglavlju: *kad biste mogli pogledati dva filma, od kojih svaki dovoljno detaljno i široko opisuje jedan mogući život koji vas čeka, poput pravog virtualnog iskustva, mogli biste procijeniti koji od njih preferirate ili ostati ravnodušni. Stoga „preferencije” ovdje uključuju sve do čega bi vam moglo biti stalo, neograničeno i unedogled u budućnost.*⁵ Također, one su vaše: stroj ne pokušava identificirati ili prigrliti neki skup preferencija, nego razumjeti i zadovoljiti (koliko je moguće) preferencije svake osobe.

Prvi princip: potpuno altruistički strojevi

Prvi princip, da stroj ima samo jedan cilj, maksimizirati ostvarenje ljudskih preferencija, središnje je svojstvo blagotvornog stroja. Konkretno, stroj treba biti blagotvoran za ljude, a ne, recimo, za žohare. Ne možemo zaobići taj koncept blagotvornosti koji je usmjeren prije svega na primatelja.

Taj princip znači da će takav stroj biti potpuno altruistički, to jest, on ne pridaje apsolutno nikakvu intrinzičnu vrijednost vlastitoj dobrobiti, pa čak ni svojem postojanju. Možda će se pokušati zaštititi kako bi nastavio koristiti ljudima, ili zato što bi njegov vlasnik bio nezadovoljan da mora platiti popravak, ili zato što bi prizor prljavog ili oštećenog robota mogao blago uznemiriti prolaznike, no ne zato što on sam želi živjeti. Ugrađivanje bilo kakve preferencije za samoočuvanjem unutar robota stvara dodatnu motivaciju koja nije u potpunosti usklađena s ljudskom dobrobiti.

Izbor riječi u prvom principu povlači dva ključna pitanja. Po pitanju svakog od njih mogla bi se napisati polica knjiga; zapravo, o tome su već napisane mnoge knjige.

Prvo je pitanje imaju li ljudi zaista preferencije u nekom značajnom ili stabilnom smislu. Istini za volju, ideja „preferencije” zapravo je idealizirana i u nekoliko pogleda ne odgovara stvarnosti. Primjerice, ne rađamo se s preferencijama koje ćemo imati kao odrasli ljudi, što znači da se one s vremenom mijenjaju. Zasad ću pretpostaviti da je takvo idealiziranje razumno. Kasnije ću razmotriti što nam se može dogoditi kad odustanemo od takvog idealiziranja.

Drugo je pitanje jedan od temelja društvenih znanosti: budući da je obično nemoguće osigurati da će se svima ostvariti najpovoljniji mogući izbor – ne možemo svi biti car svemira – na koji bi način strojevi trebali napraviti kompromis da zadovolje preferencije većeg broja ljudi? Opet, zasad se čini razumno – i obećavam da ću se u sljedećem poglavlju vratiti na ovo pitanje – prihvatiti jednostavan pristup da sve ljude tretiramo jednako. To podsjeća na korišćene utilitarizma iz 18. stoljeća i izraz „najveća moguća sreća za najveći broj ljudi”⁶ i potrebno je zadovoljiti mnoge složene preduvjete da bi takvo što funkcioniralo u praksi. Možda je najvažnije uzeti u obzir golem broj ljudi koji još nisu rođeni i razmisliti kako možemo uvažiti i njihove preferencije.

Pitanje budućih ljudi povlači još jedno, povezano pitanje: kako da uzmemo u obzir preferencije neljudskih entiteta? To jest, treba li prvi princip uključiti i preferencije životinja? (A možda i biljaka?) To je pitanje vrijedno rasprave, no rješenje te dvojbe vjerojatno neće imati velik utjecaj na razvoj umjetne inteligencije. U svakom slučaju, ljudske preferencije nerijetko uključuju i dobrobit životinja, kao i onih aspekata ljudskog života kojima izravno pogoduje postojanje životinja.⁷ Tvrditi da bi stroj trebao posvetiti pažnju životinjskim preferencijama *povrh* tih ljudskih preferencija, značilo bi da bi ljudi trebali stvoriti strojeve koji o životinjama mare više negoli ljudi, što je teško održiv pristup. Razumnije bi bilo reći da naša sklonost kratkovidnom donošenju odluka – koje zapravo nisu u našem interesu –

često dovodi do negativnih posljedica za okoliš i njegovu životinjsku populaciju. Stroj koji donosi manje kratkovidnih odluka pomogao bi ljudima da počnu provoditi ekološki zdraviju praksu. A ako u budućnosti počnemo više mariti za dobrobit životinja nego danas – što bi vjerojatno značilo žrtvovati određenu količinu vlastite intrinzične dobrobiti – strojevi će se tome prilagoditi.

Drugi princip: skromni strojevi

Drugi princip, da stroj isprva nije siguran kakve su ljudske preferencije, ključan je za stvaranje blagodatnih strojeva. Stroj koji pretpostavlja da savršeno zna svoj stvarni cilj nepokolebljivo će stremiti ka ostvarivanju tog cilja. Nikad neće pitati je li neka vrsta djelovanja ispravna jer će već znati da je to savršeno rješenje koje dovodi do cilja. Ignorirat će ljude koji skaču i viču: „Stani, uništiti ćeš svijet!” jer su to samo riječi. Ako stroj pretpostavlja da savršeno razumije svoj cilj, postat će odvojen od ljudi: ono što ljudi čine postaje nevažno jer stroj zna svoj cilj i radi na njegovom ostvarivanju.

Nasuprot tome, stroj koji nije siguran koji mu je stvarni cilj ponašat će se gotovo skromno: primjerice, bit će poslušan prema ljudima i dopustiti da ga oni ugase. Njegova će logika biti da će ga ljudi ugasi samo ako je učinio nešto krivo – to jest nešto što nije u skladu s ljudskim preferencijama. Prvi mu princip nalaže da izbjegava takvo što, no drugi mu princip daje do znanja da je takvo što moguće jer nije siguran što je „krivo”. Dakle, ako neki čovjek zaista ugasi stroj, stroj će se truditi izbjeći krivu radnju, i to je ono što želi. Drugim riječima, stroj ima pozitivan poticaj da dopusti da ga ugase. I dalje je povezan s ljudima, koji su potencijalni izvor informacija koje će mu omogućiti da izbjegava pogreške i bolje obavlja svoj posao.

Nesigurnost je još od osamdesetih godina prošlog stoljeća jedan od središnjih problema umjetne inteligencije; dapače, izraz „moderna umjetna inteligencija” često se odnosi na revoluciju koja se dogodila kad je napokon prihvaćeno da je nesigurnost neizbježna pri donošenju odluka u stvarnom svijetu. Međutim, nesigurnost kao dio ciljeva

sustava umjetne inteligencije uglavnom se ignorirala. U svem tom radu na maksimiziranju korisnosti, postizanju ciljeva, minimiziranju cijene, maksimiziranju nagrade i minimiziranju gubitaka, pretpostavljalo se da su funkcije korisnosti, cilja, cijene, nagrade i gubitka poznate. Kako je to moguće? Kako je zajednica umjetne inteligencije (kao i zajednice teorije kontrole, operativnih istraživanja i statistike) tako dugo mogla imati tako veliku slijepu točku iako je cijelo to vrijeme uvažavala nesigurnost u svim drugim aspektima donošenja odluka?⁸

Mogli bismo izmisliti neke složene tehničke izgovore,⁹ no mislim da je istina da su, osim nekih časnih iznimaka,¹⁰ znanstvenici na području umjetne inteligencije uglavnom prihvatili standardni model koji preslikava naše ideje o ljudskoj inteligenciji na strojnu inteligenciju: ljudi imaju ciljeve i pokušavaju ih ostvariti, pa bi stoga strojevi trebali imati ciljeve koje mogu ostvariti. Oni, to jest mi, nikad nisu zaista propitkivali tu temeljnu pretpostavku. Ona je dio svih postojećih pristupa konstruiranju inteligentnih sustava.

Treći princip: učimo predviđati ljudske preferencije

Treći princip, da je najbolji izvor informacija o ljudskim preferencijama ljudsko ponašanje, služi dvjema svrhama.

Prva je svrha pružiti definitivne temelje terminu *ljudske preferencije*. Pretpostavljamo da ljudske preferencije nisu upisane u stroj te da ih on ne može izravno promatrati, no svejedno mora postojati neka stvarna poveznica između stroja i ljudskih preferencija. Taj princip tvrdi da će ta poveznica biti promatranje ljudskih odabira: pretpostavljamo da su odabiri povezani na neki (možda vrlo složen) način s nekim pozadinskim preferencijama. Kako bismo dokazali zašto je ta poveznica ključna, uzmimo u obzir sljedeće: kad neka ljudska preferencija ne bi imala apsolutno nikakvog utjecaja ni na jedan stvarni ili hipotetski odabir te ljudske osobe, vjerojatno bi bilo beznačajno reći da ta preferencija uopće postoji.

Druga je svrha omogućiti stroju da postane korisniji kako uči sve više o onome što mi želimo. (Naposljetku, kad ne bi znao ništa o

ljudskim preferencijama, ne bi nam bio koristan.) Ideja je prilično jednostavna: ljudski odabiri otkrivaju informacije o ljudskim preferencijama. Ako to primijenimo na odabir između pizze s ananasom i pizze s kobasicom, povezanost je jasna. Ako to primijenimo na odabire između budućih života i odabire koji se donose s ciljem utjecanja na ponašanje takvih robota, situacija postaje zanimljivija. U sljedećem ću poglavlju objasniti kako formulirati i riješiti takve probleme. Međutim, do stvarnih komplikacija dolazi zato što ljudi nisu savršeno racionalni: nesavršenost će stati na put između ljudskih preferencija i ljudskih odabira, a stroj će morati te nesavršenosti uzeti u obzir ako želi interpretirati ljudske odabire kao dokaze ljudskih preferencija.