Superintelligence: Paths, Dangers, Strategies

Nick Bostrom
(Oxford University Press, 2014., 415 str.)

Knjiga *Superintelligence: Paths, Dangers, Strategies* predstavlja jedan od ključnih radova koji se bave temom umjetne inteligencije te svim dodatnim problemima koji bi mogli proizaći iz njezinog razvoja. Autor Nick Bostrom jedan je od važnijih mislioca koji se bave spomenutom tematikom, a njegovo interesno područje uključuje propitivanje rizika razvoja novih i potencijalno opasnih tehnologija. On je ujedno i voditelj *Future of Humanity Institute* (FHI) na Oxfordu, a na istoimenom je sveučilištu zaposlen i kao profesor. Knjiga je podijeljena na petnaest poglavlja od kojih se svako bavi određenim aspektima razvoja futurističkog koncepta koji Bostrom naziva superinteligencijom.

Bostrom u prvom poglavlju knjige čitatelja pobliže upoznaje s tematikom umjetne inteligencije gdje se kao ključan događaj ističe ljetna radionica 1956. na Dartmouth Collegeu. Ta je radionica zapravo bila skup na koji je John McCarthy, kao jedan od začetnika discipline umjetne inteligencije pozvao relevantne pojedince iz tog područja. Koncept umjetne inteligencije u to vrijeme nije bio naročito prisutan u znanstvenim krugovima, a vjerojatno još i manje izvan njih. Bostrom ističe kako je taj skup predstavljao svojevrsnu prekretnicu jer je nakon njega uslijedio porast interesa za pitanja umjetne inteligencije, a to se manifestiralo razvojem programa poput ELIZA i SHRDLU. Međutim, inicijalni entuzijazam prema ovoj novoj tehnologiji pokleknuo je 70-ih javljanjem sve češćih hardverskih poteškoća. Ideja umjetne inteligencije ponovno je revitalizirana 80-ih kada je Japan lansirao petu generaciju računala, ali Bostrom ističe kako ni ta peta generacija nije bila dostatna za ispunjavanje hardverskih zahtjeva umjetne inteligencije. Prema prikazanim primjerima vidimo kako je koncept umjetne inteligencije prolazio kroz faze rasta i pada interesa, što nije spriječilo očuvanje njezine ideje. Postupno su računala postajala sofisticiranija i naprednija pa čak i spretnija od čovjeka u nekim zadacima. Ključan je trenutak u tom pogledu predstavljala 1997. godina, kada je IBM-ovo računalo *Deep Blue* porazilo tadašnjeg prvaka u šahu Garryja Kasparova. Od tog događaja prošlo je preko 20 godina što je bio dovoljan vremenski odmak za postizanje značajnog tehnološkog skoka u računalnoj tehnologiji. S tim se tehnološkim napretcima tematika umjetne inteligencije ponovno revitalizirala te postala danas šire poznat pojam. Bostrom zaključuje kako je zbog napretka umjetne inteligencije došlo vrijeme za uvođenje novog koncepta kojeg on naziva superinteligencijom, a koji je rezultat postavljanja sljedećeg pitanja: „Što kada razvijemo inteligenciju koja je sposobna intelektualno prestići čovjeka i postati naprednija u svakom aspektu?“ Nick Bostrom takav entitet koji iznimno nadilazi ljudske kognitivne performanse u doslovno svakoj interesnoj domeni naziva superinteligencijom. Također, kroz naredna poglavlja Bostrom ulazi u detaljnu analizu vezanu uz ključna pitanja koja proizlaze iz potencijalnog razvoja superinteligencije. Za razliku od uobičajenih, nerijetko pretjerano optimističnih percepcija prema takvoj tehnologiji, Bostrom prilazi tematici iz druge perspektive tako što teoretizira o mogućim neočekivanim posljedicama i rizicima upuštanja u razvoj ovakvih tehnologija. Prema njegovom mišljenju, one predstavljaju prekretnicu koja će odlučiti budućnost čovječanstva. U tom pogledu, Bostrom nerijetko ima pesimističan i oprezan stav prema, kako kaže, nepromišljenom upuštanju u razvoj ovakvih rizičnih tehnologija koje imaju potencijal značajno promijeniti ljudsko društvo.

Drugo poglavlje Bostrom posvećuje mogućim smjerovima razvoja koji bi doveli do superinteligencije. Kao prvi smjer on navodi umjetnu inteligenciju koja bi se mogla potencijalno doraditi do razine superinteligencije. Međutim, spominju se i drugi potencijalni smjerovi razvoja, poput emulacije mozga koja bi bila rezultat mogućnosti skeniranja i repliciranja ljudskog mozga. Ističe još i mogućnost poboljšanja biološke kognicije, što bi podrazumijevalo jačanje ljudskih kognitivnih mogućnosti. Spominje još i stvaranje sučelja mozga i računala koje bi uključivalo korištenje implantata za stvaranje međuodnosa između ljudskog mozga i računala. Kao posljednji smjer navodi mreže i organizacije koje bi nastale umrežavanjem ljudskih mozgova te stvaranjem zajedničkog uma u formi kolektivne superinteligencije. Iako bi svaki od tih smjerova potencijalno rezultirao razvojem specifičnog oblika superinteligencije, valja istaknuti da se Bostrom kroz knjigu najviše bavi tipom superinteligencije temeljenim na umjetnoj inteligenciji.

Treće poglavlje bavi se mogućim oblicima superinteligencije koje Bostrom dijeli na 3 tipa superinteligencije. Prva je brzinska, druga je kolektivna, a treća je kvalitativna superinteligencija. Podjela je isključivo usmjerena na način kojim neka superinteligencija nadilazi ljudski um. Brzinska bi imala znatno veću sposobnost procesuiranja podataka od čovjeka. Stoga, ako prosječni čovjek može u sat vremena pročitati jedan znanstveni članak, brzinska superinteligencija bi za to vrijeme pročitala petnaest članaka. Kolektivna superinteligencija podrazumijeva spajanje skupa manjih inteligencija koje bi zajedno raspolagale sa znatno većim intelektualnim kapacitetima. Bostrom kao primjer toga navodi potencijalno umrežavanje ljudskih umova koje je slično ideji mreža i organizacija u drugome poglavlju. Kvalitativna superinteligencija podrazumijeva sustav koji bi bio brz otprilike kao ljudski um, ali znatno kvalitativno pametniji. S obzirom na to da je koncept pomalo nejasan, Bostrom ističe kako bi se radilo o umu koji je intelektualno nadmoćan prema ljudima kao što su ljudi prema životinjama.

Četvrto poglavlje bavi se kinetikom eksplozije inteligencije, a radi se o analiziranju mogućnosti brzine razvoja superinteligencije. Odnosno, Bostroma zanima koliko će vremena trebati da inteligencija prestigne ljudske mogućnosti nakon što postane pametna kao čovjek. Ponuđeni odgovori se svode na sporo, srednje i brzo gdje je posljednja mogućnost očito i najopasnija za čovječanstvo. Brzina uspona takve tehnologije kao i uvjeti pod kojima će ona biti razvijena značajno će ovisiti o budućim uvjetima ljudske vrste. Sa sličnom tematikom nastavlja se u petom poglavlju nazvanom *Odlučna strateška prednost*. U tom poglavlju Bostrom preispituje kakve okolnosti bi stvorio razvoj specifičnih smjerova superinteligencije. Bostroma u tom pogledu najviše brine već prethodno spomenuti brzi uspon koji bi bio dodatno problematičan ukoliko bi superinteligenciju razvio jedan entitet, kojeg on naziva *singleton*. *Singleton* može biti jedna tvrtka, pojedinac ili čak država. Kada bi neka specifična država prva razvila superinteligenciju, vjerojatna posljedica toga bi, smatra Bostrom, bila globalna dominacija te države. Za razliku od ranijih strateških, tehnoloških utrka, poput razvoja interkontinentalnih raketa i nuklearnog oružja, postoji mogućnost da bi razvoj superinteligencije rezultirao tehnološkim usponom *singletona* koji nitko ne bi mogao sustići.

Preispitivanje rezultata kognitivnih sposobnosti superinteligencije nastavlja se kroz šesto poglavlje gdje Bostrom analizira metodu kojom bi superinteligencija oduzela moć ljudima ako bi joj to bilo po volji te kakve mjere predostrožnosti trebaju biti poduzete. Bez ikakve sumnje, takav bi entitet imao na raspolaganju iznimnu moć, što bi se vjerojatno manifestiralo njegovom sposobnošću kontrole ljudske tehnologije. Bilo da se radilo o socijalnoj manipulaciji, hakiranju ili sličnim metodama, scenarij preuzimanja moći nije toliko nevjerojatan, posebno ako je hipotetska superinteligencija slična istinskom društvenom agentu koji donosi odluke, a ne samo alatu kakvim danas doživljavamo računala.

Sedmo poglavlje bavi se ciljevima superinteligencije, odnosno njihovom formulacijom. Bostrom zaključuje kako je malo vjerojatno da bi superinteligentni entitet mogao funkcionirati na sličnim temeljima kao ljudi, odnosno s nepreciznim smislom egzistencije. Pri stvaranju superinteligencije bilo bi nužno uspostavljanje preciznog skupa ciljeva i funkcija, posebno ako je ona razvijena u formi umjetne inteligencije. Međutim, time bi se stvorila nepredvidiva situacija u kojoj bi bilo teško procijeniti na koji način

će taj entitet interpretirati svoju funkciju. Stoga je zadatak ljudi procijeniti kako najbolje formulirati cilj superinteligencije. Bostrom, zatim, temeljito analizira neke od mogućih ciljeva superinteligencije, poput akvizicije resursa, samoočuvanja, tehnološkog napretka i dodatnog kognitivnog poboljšanja.

U osmome poglavlju Bostrom se dotiče pitanja na koje bi mnogi teoretičari zavjere odmah dali pozitivan odgovor, a to je: „Je li konačni ishod egzistencijalna katastrofa?“ To je donekle i povratak na pitanje koliku ugrozu superinteligentni entitet predstavlja čovječanstvu. Ovdje Bostrom zauzima pesimistično stajalište, oslanjajući se na tezu ortogonalnosti prema kojoj bilo kakva inteligencija može imati bilo kakav vrijednosni sustav. Ako je superinteligencija vrijednosno racionalna, jedan od mogućih rezultata je da nas doživljava kao što mi doživljavamo mrave pored autoceste. Naravno, sve ovisi o tome koji je zadatak superinteligencije, pa ako je to akvizicija resursa ili tehnološki napredak, možemo pretpostaviti da bi nas mogla doživljavati kao smetnju. Kroz poglavlje Bostrom prikazuje različite scenarije prema kojima bi se superinteligencija mogla slučajno preokrenuti protiv čovječanstva, a dva glavna scenarija su izdaja i greška u sustavu. Scenarij izdaje rezultirao bi promjenom prioriteta superinteligencije nakon postizanja predviđene razine autonomije, dok bi scenarij greške u sustavu podrazumijevao krivu interpretaciju zadataka ili neispravno shvaćanje zadanih vrijednosti prema kojima superinteligencija treba postupati. Sličnom se problematikom bavio Isaac Asimov kada je izmislio tri zakona robotike i onda propitivao na koje načine bi umjetna inteligencija mogla neispravno interpretirati njihovo značenje.

Deveto poglavlje bavi se tematikom kontrole, odnosno pronalaženjem načina kojim bi se mogli izbjeći mogući fatalni scenariji proizašli iz grešaka koje je Bostrom problematizirao u prethodnom poglavlju. Dva glavna problema koja on uočava temelje se na kontroli mogućnosti i odabiru motivacije. Vezano uz kontrolu mogućnosti, Bostrom nudi niz restriktivnih rješenja s ciljem stavljanja superinteligencije unutar okvira koji ljudima odgovaraju. Na pitanje motivacije, on odgovara nudeći više mogućnosti od kojih mu je najzanimljivija indirektna normativnost, temeljena na ideji nepostojanja preciznog vrijednosnog sustava ili cilja kojeg ljudi mogu smisliti za superinteligenciju, te da sustav bude osmišljen tako da ona razmišlja što bi ljudi htjeli kada bi raspolagali njezinim kognitivnim kapacitetima.

Deseto poglavlje bavi se još jednom podjelom oblika superinteligencije, ali ovaj put pitanjem rješavanja problema kontrole koji je ponuđen u devetom poglavlju. Bostrom te oblike dijeli u proroke *(oracles)*, duhove i suverene *(genies and sovreigns)* i alate, ističući njihove nedostatke i prednosti. Zadatak proročkog sustava bio bi samo odgovaranje na pitanja, a kao primjer toga navodi se neka potencijalno naprednija inačica današnjeg *Googlea*. Duhovi i suvereni bi, s druge strane, bili fokusirani na upravljanje ili ispunjavanje zadataka, a samim time bi imali značajnu razinu kontrole nad svojim odlukama. Stoga su oni, zbog njihove visoke razine agencije, po pitanju kontrole najproblematičniji tip superinteligencije. Treći su tip alati koji su samo programi za rješavanje problema bez dodatnih antropomorfnih svojstava o kojima Bostrom teoretizira u knjizi. Alati stoga ne bi bili agenti, nego limitirani sustavi analogni današnjim programima.

Jedanaesto poglavlje vraća se natrag na probleme već navedene u sedmom i osmom poglavlju. Primarno na one vezane uz scenarije u kojima bi jedna osoba, tvrtka, država imala kontrolu nad superinteligencijom. Ovo poglavlje se bavi multipolarnim scenarijima te dodatno teoretizira kako bi se izmijenio položaj prosječnog čovjeka u društvu sa superinteligencijom. Za sociologe je ovo možda najinterestantnije poglavlje jer problematizira pitanja ekonomije, rada i sličnih tema u društvu gdje je sve savršeno matematički programirano.

Dvanaesto poglavlje bavi se problemom vrijednosti, odnosno kako usaditi sustav vrijednosti u superinteligentni entitet. To je zadatak koji Bostrom smatra znatno kompleksnijim nego što se čini. Ako zamislimo čovječanstvo kao jedinku, ono samo po sebi nema unificirani vrijednosni sustav zbog čega bi pokušaj stvaranja istog za umjetnu inteligenciju predstavljalo iznimno težak zadatak. Bostrom priznaje kako su filozofi po tom pitanju pokleknuli te se ni sami nisu usuglasili oko jednog moralnog, odnosno vrijed-

nosnog sustava. Iz toga proizlazi pitanje kako ćemo usaditi umjetnoj inteligenciji nešto što sami nismo uspjeli riješiti? Kroz ovo poglavlje Bostrom nudi niz potencijalnih rješenja za problem usađivanja vrijednosti umjetnoj inteligenciji. Na to se nadovezuje trinaesto poglavlje koje se bavi sličnom tematikom, odnosno preispitivanjem možemo li se pouzdati u svoje vlastite odluke. Naše je društvo još pogođeno i problemima spolne nejednakosti, rasizmom, siromaštvom i drugima. Kako se onda mi možemo postaviti u položaj moralnog autoriteta koji određuje superinteligenciji što točno treba učiniti?
Coherent Extrapolated Volition jedan je od odgovora koje Bostrom nudi, a veže se uz prethodno spomenut koncept indirektne normativnosti, odnosno prebacivanja rješavanja vrijednosnih problema na superinteligenciju, što je ideja već spomenuta u ranijim poglavljima, ali se ovdje detaljnije razrađuje. Četrnaesto poglavlje uglavnom se bavi odnosima među ljudima koji razvijaju umjetnu inteligenciju i mogućim posljedicama proizašlih iz tih odnosa te sukobima raznih interesnih skupina prisutnih u pojedinim razinama razvoja. Knjiga potom završava s petnaestim poglavljem u kojem Bostrom pokušava dati okvirni zaključak te se ponovno vraća na neka pitanja već postavljena u prethodnim poglavljima, uz kratko sažimanje poante čitave knjige.

Knjiga *Superintelligence* nudi veoma mračna predviđanja o ljudskoj budućnosti čime značajno spušta visoku razinu optimizma koja vlada među stručnjacima koji se bave umjetnom inteligencijom. Razvoj oblika inteligencije kakvog Bostrom predviđa vjerojatno je moguć tek u znatno daljnjoj budućnosti nego što to on procjenjuje, ali pitanja i analize koje postavlja su svejedno aktualne. Čak i ako izbjegnemo scenarij katastrofe, razvoj takve tehnologije značajno bi utjecao na promjenu društvenih odnosa i položaja čovjeka u društvu. Kao i uz razvoj svih prethodnih tehnologija ovakvog potencijala, dramatične društvene transformacije su moguće i stoga je pozitivno što je Bostrom svojom knjigom oživio raspravu o toj temi. Slagali se mi s njim ili ne, Bostromova knjiga nudi zanimljive teorije o budućnosti i izvrsne konceptualne analize kakve prije nisu postojale. Stoga, bez obzira na futurističku tematiku koja možda u svim aspektima nije strogo znanstvene prirode, u njoj svejedno nalazimo niz zanimljivih pitanja na koja će buduće generacije morati dati odgovore, a među njima svakako i sociolozi.

Mladen Mirat