Artificial Intelligence Safety and Security. Urednik Roman V. Yampolskiy. Chapman i Hall, 2018. 474 str.

Umjetna inteligencija (UI) sve je prisutnija u našoj svakodnevici i područje se rapidno razvija. Potpuno autonomna vozila već su godinama prisutna na cesti, sofisticirani algoritmi upravljaju trgovanjem dionicama na burzama diljem svijeta, digitalni osobni asistenti uređuju nam rasporede, odgovaraju na naša pitanja i šale se s nama dok računalni algoritmi pišu novinske članke i samostalno upravljaju prometom. Iako se radi o fascinantnim tehnološkim dostignućima, sposobnosti umjetne inteligencije još su uvijek ograničene. Pravi proboj u području predstavljat će kreiranje stroja koji će imati kapacitet razumjeti i naučiti sve što može i čovjek. Jednom kada u dalekoj budućnosti dosegnemo tu razinu, samo korak dalje bit će razvoj tzv. nadinteligencije (engl. *Superintelligence*), dovoljno sposobnog stroja koji bi mogao razviti svijest te bi se mogao redizajnirati ili stvarati inteligentnije nasljednike i tako dovesti do tehnološkog singulariteta, hipotetske točke u kojoj tehnološki razvoj postaje nezaustavljiv i izvan ljudske kontrole. Između današnje slabe umjetne inteligencije i hipotetske pojave nadinteligencije ogroman je prostor u kojem će naši životi postepeno bivati sve više isprepleteni s umjetnom inteligencijom. Ova knjiga promišlja o sigurnosnim ugrozama koje iz toga mogu proizaći i kako minimizirati rizik od negativnih posljedica za čovjeka i društvo.

Artificial Intelligence Safety and Security prvi je višeautorski rad o sigurnosti UI-ja čiji je primarni cilj približiti široj javnosti područje o kojem uska skupina stručnjaka promišlja već desetke godina. Sama knjiga podijeljena je u dva dijela. Prvi dio, slikovito nazvan *Zabrinutosti svjetlonoša,* u kronološkom redoslijedu prikazuje jedanaest prethodno objavljenih temeljnih radova iz područja koji ocrtavaju različite aspekte *problema kontrole UI*, odnosno sprečavanja izrade (nad)inteligencije koja bi nanijela štetu čovjeku. Uključuje radove autora iz brojnih područja i to poduzetnike, inženjere, filozofe, znanstvenike, pisce. Drugi dio predstavlja skup sedamnaest poglavlja u kojima autori iz različitih područja pokušavaju dati svoje teoretske i praktične odgovore na probleme i zabrinutosti iznesene u prvom dijelu. Poglavlja se značajno razlikuju u perspektivi te sežu od visokotehnoloških promišljanja konkretnih računalnih rješenja

do etičkih, filozofskih i politoloških eseja. Drugi dio koncipiran je na način da svaki čitatelj može za sebe odabrati ona poglavlja koja odgovaraju njegovom interesu budući da nisu međusobno povezana.

Prvi dio započinje (ne)slavnim kontroverznim člankom Billa Joya *Why the Future Doesn't Need Us* objavljenom 2000. u Wiredu. Radi se o pesimističnom neoluddističkom prikazu prijetnji koje proizlaze iz razvoja bioinženjeringa, nanotehnologije i robotike. Iako pretenciozan i kontroverzan, stavljen u vremenski kontekst članak je vrlo važan. Otvorio je široku raspravu i inspirirao na tisuće odgovora. Kao svojevrstan kontrast, idući je prikazan optimistični pogled Raymonda Kurzweila, transhumanista i zagovaratelja razvoja tehnologije. Dakako, kao najkontroverzniji Joyev prijedlog Kurzweil odbacuje zagovaranje odricanja od razvoja tehnologije jer je uvjeren kako će se naše obrambeno znanje razvijati paralelno s opasnostima. Iako to u velikoj mjeri stoji i zasigurno ćemo u budućnosti biti pripremljeniji nositi se s izazovima koje donosi razvoj UI-ja, na opasnost od takvog naivnog pristupa razvoju UI-ja upozorava Mark Walker u svojem filozofskom eseju prikazanom u drugom dijelu knjige. Walker prikazuje zakone pesimizma prema kojima ubojitost tehnologije raste eksponencijalno i ofenzivna tehnologija uvijek se razvija i primjenjuje na nekoj populaciji prije nego što se uopće razvije njezin defenzivni pandan. Čovjek će u jednom trenutku razviti tehnologiju koja će imati kapacitet uništiti čovječanstvo prije nego što će razviti odgovarajuću obranu od te tehnologije.

Čitateljima iz područja društvenih znanosti svakako će biti zanimljiv rad utjecajnih stručnjaka u tom području Nicka Bostroma i Eliezera Yudkowskyog koji promišljaju etičke aspekte UI-ja, a posebice moralni status i pravnu osobnost svjesne umjetne inteligencije. Na to se nadovezuju promišljanja o temeljnim ljudskim vrijednostima kojima je potrebno podučiti autonomni UI. Inženjeri koji programiraju algoritme za autonomna vozila već se danas susreću s problemom tramvaja (engl. *Trolley problem*) i šira primjena autonomnih vozila nije izgledna dok ne definiramo odgovor na to etičko pitanje. Max Tegmark na to se nadovezuje tvrdnjom kako, istom logikom, prije nego što bismo mogli razviti prijateljsku nadinteligenciju moramo riješiti pitanje smisla života te cinično zaključuje da ukoliko ustupimo kontrolu nadinteligenciji prije nego što smo odgovorili na pitanje što je smisao života, na to će se pitanje pronaći odgovor bez sudjelovanja čovjeka.

Iz politološke perspektive, u prvom dijelu knjige najrelevantniji je članak američkog diplomata Matta Chesena o ugrozama za pojedinca i društvo koje proizlaze iz strojno upravljanih komunikacija (engl. *MAchine Driven COMmunications* – MADCOM). Već danas računalni algoritmi pišu značajan broj novinskih članaka koje čitamo, upravljaju načinom na koji nam se prikazuje sadržaj na internetu, kroje naše e-iskustvo. Računalni algoritmi određuju kojim ćemo informacijama biti izloženi, kakav će biti emocionalni ton teksta vijesti koja će nam biti prikazana. To čine temeljem

enormne količine podataka kojima raspolažu o nama, a koje smo dobrovoljno predali u elektronski korporativni internetski oblak. Nedavni slučajevi manipulacije izbornim procesima, utjecaji na glasače, poticanje nereda, polarizacija naroda predstavljaju sitnicu u odnosu na razmjere manipulacije koju će propagandistima omogućiti razvoj UI-ja.[1] Ljudi se ne mogu nositi s MADCOM-ovima, barem ne samostalno i bez pomoći prijateljskog UI-ja.

Drugi dio počinje poglavljem Gusa Bekdasha u kojem postulira i elaborira 22 principa koje bi trebalo slijediti pri dizajnu UI-ja kako bi ta tehnologija bila sigurna za ljude. Radi se o vrijednom doprinosu promišljanju o budućnosti s ograničenom uporabnom vrijednosti s obzirom na današnji stupanj razvoja UI-ja. Edward Frenkel pažnju usmjerava prema programerima i inženjerima koji kreiraju algoritme UI-ja. Osim mogućnosti zlonamjerne sabotaže prilikom programiranja UI-ja, Frenkel ističe i kako je velika većina stručnjaka koji danas sudjeluju u osmišljavanju, projektiranju, dizajniranju, kreiranju, testiranju i upravljanju UI-jem tehničkog obrazovanja te mnogi vrlo malo znaju o psihologiji i društvenim procesima. Na to se nadovezuje Mahendra Prasad svojom tvrdnjom da će, uz to što će s porastom broja odluka koje prepuštamo UI-ju rasti potreba za naprednim računalnim modeliranjem, biti potrebno jačati obrazovanje dizajnera UI-ja u etičkom, ekonomskom, političkom i socijalno-filozofskom području. Yampolskiy je u tu knjigu o sigurnosti sustava umjetne inteligencije uvrstio radove velikog broja stručnjaka društvenih i humanističkih znanosti čime je na vrlo odvažan način definirao buduću viziju razvoja umjetne inteligencije kao strogo interdisciplinarno područje, iako ostaje nejasno koja bi sve područja valjalo uključiti. Unatoč tome što postoji još mnogo prostora za napredak i promišljanje utjecaja napretka UI-ja na pojedinca, društvo, demokraciju i ekonomiju, ova knjiga predstavlja dobar početak. U području razvoja UI-ja značajan primat imaju inženjeri, ali to će se morati promijeniti uključivanjem autonomnih strojeva u društvo.

Kroz više poglavlja u drugom dijelu knjige u promišljanjima o mogućim rješenjima problema kontrole UI-ja provlači se potreba za uspostavom nekakve vrste *globalnog upravnog tijela* koje bi imalo regulatorne i izvršne ovlasti radi sprečavanja razvoja i primjene zlonamjernog UI-ja. Phil Torres u svojem se futurističkom poglavlju fokusirao na opasnost koja proizlazi iz zlonamjernih pojedinaca, sociopata, UI-terorista. Budući da je povijest pokazala kako moć kvari ljude, on nudi viziju u kojoj bi potpuni nadzor i kontrola nad ljudima bila prepuštena nekakvoj vrsti prijateljske nadinteligencije, konceptu kojem veći dio ostatka knjige proturječi. Za razliku od njega, Maurizio Tinnirello koristeći teoriju međunarodnih odnosa, odnosno ofanzivni realizam,

[1] O ugrozama za demokraciju koje proizlaze iz nekritičkog i masovnog prikupljanja podataka opsežno sam pisao u svojoj doktorskoj disertaciji: https://repozitorij.fpzg.unizg.hr/islandora/object/fpzg%3A868.

argumentira kako najveća opasnost dolazi od država, regionalnih sila koje suprotno konvencijama i apelima stručnjaka u međusobnoj spirali naoružanja izazvanoj sigurnosnom dilemom razvijaju UI u vojne svrhe. Prema svojevrsnoj deklaraciji koju su potpisale tisuće istraživača iz područja UI-ja, autonomna oružja opisuju se kao treća revolucija u ratovanju, nakon otkrića baruta i nuklearnog oružja. Već su bespilotne letjelice, kojima zapravo daljinski upravljaju ljudi, promijenile koncept ratovanja, a možemo samo zamisliti na koji bi način potpuno autonomna kopnena vojska, topništvo ili zrakoplovstvo radikalno transformirali ratovanje, a potencijalno i čovječanstvo.

Već smo danas suočeni s nezgodama koje je prouzročio UI te možemo očekivati kako će se učestalost i ozbiljnost negativnih posljedica drastično povećavati s razvojem tehnologije. Knjiga koju je sastavio Roman Yampolskiy otvara izuzetno važno poglavlje razvoja UI-ja, pitanje njezine sigurnosti. Odabir tema i autora posebno je značajan jer jasno ocrtava snažnu potrebu za interdisciplinarnim razvojem područja. Vrlo se malo autora referira na to koliko je pojava neke vrste nadinteligencije doista realna i bliska. Procjene autora iz knjige sežu od 2020., preko 2045., pa sve do nekoliko stoljeća u budućnost. Iako je ton prikazanih radova prožet futurizmom koji u pojedinim poglavljima ide do granice znanstvene fantastike, važno je u ovom trenutku razvoja UI-ja početi na sustavan i znanstven način promišljati o toj tehnologiji, donositi ograničenja, postulirati principe i određivati smjer budućeg razvoja. Neće biti velikog i jasnog trenutka u kojem će zlonamjerni UI preuzeti dominaciju nad ljudima nego ćemo postepeno skliznuti najprije u stanje ovisnosti, a kasnije i submisivnosti toj tehnologiji. Ako je suditi prema dosadašnjem odnosu ljudi prema novim tehnologijama, društvenim mrežama i pametnim telefonima te kapacitetu i razini globalne suradnje demonstriranoj u nedavnom suočavanju sa smrtonosnom pandemijom, teško je zadržati optimizam.

Andro Pavuna