Da li treba da gradimo veštačku inteligenciju?

Domaći rad u okviru kursa
Istorija i filozofija računarstva
Univerzitet u Beogradu
Matematički fakultet

Vladimir Mandić
mi18465@alas.matf.bg.ac.rs

23.04.2022

1 Uvod

U ovom poglavlavlju prelazimo na drugo od naša dva etička pitanja: da li treba da gradimo „veštačku inteligenciju“ - to jest softver („softbotove“) ili hardver (robote) koji mogu misliti? Ovo pitanje ima najmanje dva aspekta: Prvo, da li je etički ili moralno stvoriti računar koji bi mogao da misli ili da doživi emocije? Drugo, kakav bi bio odnos takvih kreacija prema nama?

Kada sam prvi put predavao filozofiju CS, oko 2006. godine, o pitanju da li bi trebalo da gradimo veštačku inteligenciju jedva da se ikada razgovaralo. Tokom godina dok sam predavao različite verzije kursa, prikupljao sam članke koji su bili relevantni za sve ove teme. Deo pripreme ove knjige uključivao je pregled tih radova i uključivanje nekih njihovih pogleda. Ja bih to uradio tako što bih ih organizovao hronološkim redom. Za većinu tema, bio je približno isti broj radova u svakoj dekadi od 1970-ih do 2010-ih godina. Za temu ovog poglavlja, međutim, nisam imao takvih „novih“ radova pre 2000. godine (ne uključujući obavezna i preporučena dela ovog poglavlja, od kojih je jedno beletristično delo); bilo ih je 8 iz 2000-ih; a bilo ih je i skoro duplo više samo u prvoj polovini 2010-ih. To ukazuje na skoro eksponencijalni rast interesovanja za etiku veštačke inteligencije, kako u akademskoj tako i u popularnoj štampi. Bez sumnje, to je delimično i zbog činjenice da se roboti i „inteligentni“ računari približavaju svakodnevnoj stvarnosti (mislite na Siri ili Aleka) i tako je pitanje postalo hitnije. Ovo je razlog više da postoji filozofsko razmišljanje o budućim tehnologijama mnogo pre nego što te tehnologije budu implementirane.

Kratka priča Stanislava Lema „Non Serviam“ (1971) tiče se onoga što se danas zove „veštački život“ (ili „A-Life“) [1]. A-Life je pokušaj da se život istraži kao računarski proces razvijanjem računarskih programa koji generišu i razvijaju virtuelne entitete koji imaju neka ili sva apstraktna svojstva povezana sa

biološkim živim bićima.

U Lemovoj priči, istraživač A-Life-a konstruiše računarski svet inteligentnih entiteta i prati njihovu evoluciju i razvoj jezika i filozofije. Ovi „personoidi“ raspravljaju o postojanju Boga na isti način kao i filozofi. Razlika (ako uopšte i jeste razlika) je u tome što istraživač (i čitalac) shvata da je on, istraživač, njihov bog; da, iako ih je stvorio, on nije sveznajući ni svemoćan; i, što je još gore, kada mu ponestane novčanih sredstava, on će morati bukvalno da izvuče utikač na računaru i time ih uništi.

Da li bi takav eksperiment uošte trebalo da počne? Šta bi se dogodilo da AI programi zaista prođu Tjuringov test i počnu da interaguju sa nama (i mi sa njima) na svakodnevnom nivou? Da li bismo imali moralnu ili zakonsku odgovornost prema njima? Da li bi oni imali nekakvu odgovornost prema nama? Da li bi bili zaista svesni ili bi bili samo filozofski zombiji? Iako je to trenutno prvenstevno stvar naučne fantastike, takođe je i predmet mnogih filozofskih razmišljanja. Razmotrićemo neka od ovih pitanja u ovom poglavlju.

2 Da li je AI moguć u principu?

Jedno od najranijih filozofskih istraživanja ovih pitanja je esej Majkla R. LaChat-a koji se pojavio u AI Magazinu (LaChat,1986) [2]. LaChat je tvrdio da je vredno razmotriti moralne implikacije stvaranja veštačke inteligencije - veštačke osobe. Jedan od razloga je taj što bi se to moglo dogoditi, tako da treba da budemo spremni na to. Drugi razlog je taj što, čak i ako se ispostavi da je to malo verovatno, takva rasprava rasvetljava šta to znači biti ličnost, što je u svakom slučaju važan cilj.

U poglavljima 2.8 i 12.4.6, raspravljali smo o klasičnom filozofskom problemu um-telo (ili um-mozak) dualizam. Ovo je, otprilike, stav da su um i mozak dve različite vrste entiteta koje na neki način interaguju. Jedan od načina da se to reši je da kažete da se um može smatrati apstrakcijom (kao što se govori u poglavljima 9 i 14) koja može biti višestruko implementirana. Jedna implementacija bi bila u mozgu; druga može biti u računaru. Ako se računarska teorija spoznaje (saznanja) može razviti, onda se njeni algoritmi mogu implementirati u ne-ljudske računare, a takvi kompjuterski programi (ili računari koji ih pokreću) bi tada bili kandidati da se smatraju „veštačkom inteligencijom“.

Po LaChat-ovom mišljenju, AI je u principu moguća ako je moguće da postoji „funkcionalni izomorfizam“ između (1) neuronske mreže koja čini naš mozak (tj. stanja i procesi mozga) i (2) bilo koje druge fizičke implementacije funkcionalnog (tj. psihološkog) ponašanja koje ta neuronska mreža implementira. Drugim rečima, psihologija je apstrakcija koja se može implementirati ili u mozgovima ili u drugim fizičkim medijima.

„Funkcionalizam“ u filozofiji uma je otprilike stav da je spoznaja jedna od funkcija mozga; kao slogan, um je ono što mozak radi. Kao što je filozof Hilari Putnam (1960) prvi predložio, Tjuringova mašina je u istoj relaciji sa kompjuterskim stanjima i procesima kao što su mentalna stanja i procesi u relaciji sa

stanjima i procesima mozga (ponekad je rezimirano kao „um je za mozak isto što je softver za hardver“) [3]. Funkcionalizam, kao način rešavanja mind-brain problema, ima prednost u tome što dozvoljava svim mentalnim stanjima i procesima da budu implementirani u nekim fizičkim stanjima i procesima;Ovo je princip „višestruke realizacije“.

Postoje, naravno, problemi, kako za funkcionalizam posebno, tako i za AI uopšte. Jedan je problem ličnosti. LaChat koristi izraz „lična (veštačka) inteligencija“, znači, otprilike, AI agent (robot ili samo neki softver) koji se može smatrati ličnošću. Da li bi „lična inteligencija... mogla imati ličnost?“. LaChat misli da je ovo „skoro nemoguće“, ali bilo je značajnog kompjuterskog rada na emocijama - sigurno jedna važna osobina ličnosti - tako da ovo ne bih isključio.

Drugi problem za funkcionalizam tiče se bola i drugih „kvalija“, odnosno kvalitativnih „osećaja“ i „iskustava“ kao što su boje i zvuci. Jedan problem je što nije jasno kako su psihološka iskustva kvalija implementirana u mozgove ili bilo koji drugi fizički medij. Srodni problem je da li računari mogu da dožive kvalije, pa čak i kada bi mogli, kako bismo mi to znali. Ovo je ogromna tema daleko izvan našeg trenutnog obima, ali za kratko razmatranje da kvaliji nisu na odmet za veštačku inteligenciju, pogledajte digresiju na kraju ovog poglavlja o računaru koji, možda, oseća bol.

3 Šta je to ličnost?

Kako bismo znali da li smo postigli „ličnu veštačku inteligenciju“? Jednosmerno, naravno, možda tako što će proći Tjuringov test. LAChat nudi drugačiji kriterijum: tako što ćemo videti da li agent veštačke inteligencije zadovoljava nezavisnu definicije „ličnosti“. Tako da mi sada trebamo pitati: Šta je to ličnost?

Pitanje koje vrste entiteta se računaju kao „ličnosti“ nije ograničeno na AI. Pitanje koje se naročito nameće u debati o abortusu: da pojednostavimo stvari, ako su fetusi osobe, i ako je ubijanje osoba nemoralno, onda je abortus nemoralan. Takođe se nameće i u životinjskoj etici kao i u pravu i u politici: Da li su delfini dovoljno inteligentni da ih možemo smatrati ličnostima? Šta je sa vanzemaljcima? Ili korporacijama? Poenta je da postoji razlika između biološke kategorije biti čovek i etičke ili pravne kategorije biti ličnost. Pitanje je: kako se apstraktno može okarekterisati ličnost, odnosno na način nezavisan od implementacije?

Jedna od najranijih filozofskih rasprava o ličnosti nastala je zbog engleskog filozofa Džona Loka, koji je živeo pre oko 350 godina (1632 - 1704). U svom Eseju o ljudskom razumevanju, Lok je napravio razliku između „ideja“ „čoveka“i „ličnosti“ (Lok, 1694) [4].

Sa mogućim izuzetkom svesti — pa čak i o tome se može diskutovati — sve ove karakteristike bi se mogle primeniti na veštačku inteligenciju.

Umesto Lokove definicije, LaChat koristi analizu ličnosti od bioetičara Džozefa Flečera (1972) [5]. Prema Flečerevoj analizi, x je osoba ako i samo ako x ima sledeće potizivne i negativne karakteristike:

Pozitivne karakteristike osobe:

1. minimalna inteligencija
    - Ovo može značiti, na primer, koeficijent inteligencije veći od 30 ili 40 (ako verujete da IQ meri „inteligenciju“). To jest, biti minimalno inteligentan ne znači puki biološki život, verovatno da bakterija ne bi bila minimalno inteligentna. Na primer, minimalna inteligencija može uključivati neki nivo racionalnosti, ili možda čak i upotrebu jezika.(Prema Hofstadteru 2007, ono što Flečer naziva „minimalna inteligencija“ bi se evolutivno odnosilo samo na oblike života „više“ od komarca [6]; videti takodje Tye 2017 [7]; Roelofs i Buchanan 2018.[8])
2. osećaj sebe
    - To jest, osobe moraju biti samosvesne i pokazivati samokontrolu.
3. osećaj za vreme
    - Osobe moraju imati osećaj za prošlost, dakle neku vrstu kulture; osećaj za budućnost, tako da imaju sposobnost da prave planove; i osećaj protoka vremena.
4. društvena uloga
    - Osobe moraju imati sposobnost da se odnose prema drugima, da imaju brigu o drugima, i da komuniciraju sa drugima (otuda potreba za jezikom kao delom minimalne racionalnosti).
5. radoznalost
    - Odnosno, osobe ne smeju biti ravnodušne
6. promenljivost
    - Osobe moraju biti kreativne i biti u stanju da promene svoje mišljenje
7. idiosinkrazija, ili jedinstvenost
    - Osobe nisu „indigo kopije“ bilo koje druge osobe.
8. nekortikalna funkcija
    - Moždana kora je mesto gde se odvijaju sva „kognitivna dejstva“ u mozgu, dakle, za Flečera, ličnost mora imati nešto čija je funkcija ekvivalentna korteksu. (Za više informacija o neokortikalnoj funkciji pogledati rad Silvije Kardozo iz 1997.[9])

Negativne karakteristike osobe:

1. ni suštinski neveštački ni suštinski antiveštački

- Ova klauzula dozvoljava višestruku realizaciju i ne ograničava ličnost na biološke entitete

2. nije suštinski seksualni
    - To jest, entitet koji nije proizveden seksualnom reprodukcijom (kao što je klonirani entitet – tačnije – robot) može biti ličnost.
3. nije suštinski snop prava
    - Flečer tvrdi da ne postoje „suštinska prava“; dakle pojam prava se ne može koristiti za karakterizaciju ličnosti.
4. nije suštinski poklonik
    - Ne morate biti religiozni da biste bili ličnost.

Lokov i Flečerov pokušaj nisu jedini pokušaji da se definiše „ličnost“. Tomas Vajt (2007 [10], 2013 [11]), etičar koji je pisao o delfinima i kitovima, nudi još jedan:

1. „biti živ“
2. „biti svestan“
3. ima „sposobnost da doživljava pozitivna i negativna osećanja (zadovoljstvo i bol)“
4. imati „emocije“
5. imati „samosvesnost i ličnost“
6. ispoljavanje „samokontrolisanog ponašanja“
7. „prepoznavanje i postupanje prema drugim osobama na odgovarajući način“
8. posedovanje „serije intelektualnih sposobnosti višeg reda (apstraktno mišljenje, učenje, rešavanje kompleksnih problema i komuniciranje na način koji sugeriše misao)

Nije nerazumno misliti da bi AI agent mogao dostići nivo programiranja koji bi mu mogao dati neke ili sve ove karakteristike. I zato su pitanja koja se nameću, da li lični AI ima ikakva prava i da li bi mi trebalo da imamo neke odgovornosti prema njemu, razumna. Pa hajde da ih razmotrimo.

4 Prava

Da li „lični“ AI ima prava? Odnosno, da li veštačka inteligencija koja prolazi Tjuringov test ili zadovoljava definiciju ličnosti ima prava?

Na primer, da li bi imala pravo da ne bude rob? Na prvi pogled, možda mislite da bi. Ali zar to nije ono što većina robota treba da bude? Na kraju krajeva, većina industrijskih i robota za ličnu asistenciju koji su sada u upotrebi su robovi u smislu da moraju da rade šta im mi kažemo da rade, a oni nisu plaćeni za svoj rad. Dakle, ako prođu Tjuringov test ili test ličnosti, da li imaju pravo da ne rade ono za šta smo ih mi stvorili da rade? Filozof Stiv Petersen (2007) je sugerisao da oni nemaju to pravo — da je „robotsko služenje dozvoljeno“ [12].

Pod „robotskim ropstvom“, Petersen ne podrazumeva dobrovoljnu pomoć, gde vi radite nešto ili pomažete nekome zato što želite, a ne zato što ste plaćeni za to. Niti misli na ropstvo u smislu prinudnog rada koji je suprotan vašoj volji. Pod „robotskim ropstvom“ on misli na robote koji su u početku programirani da žele da nam služe, posebno, da žele da rade zadatke koje ljudi smatraju ili neprijatnim ili nezgodnim. Na primer, zamislite robota koji je programiran da voli da pere veš. Ovo podseća na kastu „epsilon“ u Vrlom novom svetu Oldosa Hakslija, koji su genetski programirani da imaju ograničene želje — oni koji su predodređeni da budu operateri liftova ne žele ništa drugo nego da upravljaju liftovima [13].

Odgovore na ovakva pitanja najbolje je dati sa stanovišta određenih etničkih teorija, koje su izvan našeg dometa. Ali evo dve mogućnosti koje Petersen razmatra.

Aristotel je verovao da ljudi imaju suštinska svojstva. Dakle aristotelovski etičar bi mogao da tvrdi da inženjering ljudi jeste pogrešan jer ljudi imaju suštinsku funkciju ili svrhu i bilo bi pogrešno udaljiti ih od toga. U ovom slučaju nema paralele sa robotima. U stvari osnovna funkcija robota bi mogla biti da pere veš.

Kant je verovao da su ljudi autonomni u smislu da slede svoja moralna pravila koja moraju biti univerzalno generalizovana. Dakle, kantovski etičar bi mogao da raspravlja da, ako je robot za pranje veša takođe autonoman, onda bi bilo pogrešno sprečiti takvog robota da pere veš, a ne bi bilo štetno ni pustiti ga da samostalno radi ono što želi da radi. S druge strane, ako roboti nisu autonomni, onda ne možemo učiniti loše robotu tako što ćemo ga naterati da pere naš veš, ništa više nego što možemo učiniti pogrešno prema mašini za pranje veša.

5 Odgovornosti

Da li bismo mi ljudi (i programeri) imali bilo kakvu odgovornost prema ličnim AI koje bismo mogli da sretnemo, posedujemo ili stvorimo? Da li bi konstrukcija lične AI bila nemoralan eksperiment? Neki naučni eksperimenti se smatraju nemoralnim, odnosno krše određena (ljudska) prava. Postojanje

institucionalnih odbora za pregled na univerzitetima svedoči o tome.

Ponekad, hvale vredan cilj može imati negativne sporedne efekte. Ali šta ako su troškovi – odnosno negativne posledice – vrednog cilja preskupe? (Uporedite ovo pitanje sa tim da li postoje „pravedni“ ratovi.) Rani istraživač kibernetike Norbert Viner borio se sa ovim pitanjem:

Ako se pridržavamo svih ovih tabua, možemo steći veliku reputaciju konzervativaca i zdravih mislilaca, ali mi ćemo veoma malo doprineti daljem napredovanju znanja. To je deo naučnika - inteligentnog književnika i poštenog duhovnika - da se zalaže za jeretička i zabranjena mišljena eksprimentalno, čak i ako konačno treba da ih odbaci(Viner, 1964, str. 5) [14].

Čini se da je osnovni etički princip ovde ono što LaChat naziva „nezlonamernošću“, ili nenaškodljivošću. Ovo je strože od „dobrotvorstva“ ili činjenja dobrog, jer dobročinstvo (činjenje dobrog) može dozvoliti ili zahtevati nanošenje štete nekolicini zarad dobrobiti mnogih (barem, prema etičkom stanovištu zvanom „utilitarizam“), dok bi nezlonamernost ograničila činjenje dobra kako bi se izbeglo nanošenje štete.

Da li je stvaranje lične veštačke inteligencije korisno ili ne za samu veštačku inteligenciju? Da li sam čin njenog stvaranja šteti onome što je stvoreno? Jedan od načina da razmislite o ovome je da se zapitate da li je svesni život „bolji“ nego nikakav život. Ako nije, onda stvaranje veštačkog života nije „terapeutski ekseriment“, pa stoga nije dozvoljeno od strane odbora za pregled. Zašto? Zato što je predmet eksperimenta - veštačka osoba koju će eksperiment stvoriti ako bude uspešan (ili, možda čak i više, ako je samo delimično uspešan) - ne postoji pre nego što eksperiment počne. Ovde se približavamo teoriji egzistencijalizma, jednoj od onih čiji su principi sažeti u sloganu „egzistencija prethodi suštini“.

Aristotel je imao suprotan stav: suština prethodi postojanju. To jest, vi ste određena ličnost i ne možete promeniti ovu činjenicu. Vaša „suština“ je „suštinska“ - nije promenljiva. Ali egzistencijalistički slogan znači ko si, kakva si ličnost - tvoja suština - nešto što se može odrediti tek nakon što se rodiš (nakon što nastaneš). Štaviše, vaša suština nije nepromenljiva jer svojim postupcima možete promeniti ko ste.

Sa egzistencijalističkog pogleda, prvo postojite, a zatim određujete šta želite biti. Prema Aristotelovskom gledištu, suština je nešto poput apstrakcije, koja mora biti implementirana (ili „realizovana“). U veštačkoj inteligenciji možemo - zaista - moramo - planirati suštinu entiteta pre njegovog nastanka (pre njegove implementacije). U svakom slučaju, ne možemo garantovati da će sve ispasti u redu. Dakle, stvaranje AI je verovatno nemoralno!

6 Lična AI i moral

Pojavljuju se sasvim drugačija razmatranja, bez presedana, osim možda u kontekstu odgajanja dece, kada pitamo šta bi bilo da sami sistemi veštačke inteligencije budu moralni agenti - to jest, da budu u stanju da preuzmu etičku

odgovornost za svoje postupke.... Takvi sistemi moraju biti sposobni za moralno rasuđivanje... —Brajan Kantvel Smit(2019, str. 125) [15].

Razmotrili smo da li je moralno stvarati ličnu veštačku inteligenciju. Pretpostavimo da uspemo u tome. Da li bi ta veštačka inteligencija koju sami stvaramo bila moralna? Da li bi to imalo neke odgovornosti prema nama?

Ako su AI programirani, onda bi se moglo reći da nisu slobodni, dakle da su amoralni. Ovo je drugačije od nemoralnog! Biti „amoralan“ samo znači da je moral nebitan za to šta ste ili ko ste. Da pojednostavim, dobri ljudi su moralni, loši ljudi su nemoralni, olovka je amoralna. Aktuelno pitanje je da li su lični AI amoralni ili ne.

Ovde smo naleteli na jedno od velikih pitanja filozofije: Postoji li slobodna volja? Da li je ljudi imaju? Da li je roboti imaju? Nećemo pokušavati da istražimo ovo pitanje ovde, već samo napominjemo da je najmanje jedan istraživač AI, Dru McDermot, tvrdio da slobodna volja može biti neophodna iluzija koja proizilazi iz naše samosvesti (McDermot, 2001) [16].

Drugačiju perspektivu zauzeo je Erik Ditrih (2001, 2007) [17] [18]. On tvrdi da bi se roboti mogli programirati tako da budu bolji od ljudi (možda zato što njihova suština prethodi njihovom postojanju). Dakle, mogli bismo da smanjimo količinu zla u svetu tako što ćemo izgraditi moralne robote i dozvoliti im da neslede Zemlju.

Literatura

[1] Stanislaw Lem. Non serviam. *S. Lem, A Perfect Vacuum, trans. by M. Kandel (New York: Harcourt Brace Jovanovich, 1979)*, 1971.

[2] Michael LaChat. Artificial intelligence and ethics: an exercise in the moral imagination. *Ai Magazine*, 7(2):70–79, 1986.

[3] Hilari Putnam. *Minds and Machines, Dimensions of Mind: A Symposium.* PhD thesis, ed. S. Hook. New York: New York University Press, 1960.

[4] John Locke. 1694. an essay concerning human understanding. book ii, chapter xxvii. clarendon, 1975.

[5] Joseph Fletcher. Indicators of humanhood: a tentative profile of man. *Hastings Center Report*, pages 1–4, 1972.

[6] Douglas Hofstadter. *I am a strange loop.* Basic books, 2007.

[7] Michael Tye. *Tense bees and shell-shocked crabs: are animals conscious?* Oxford University Press, 2017.

[8] Luke Roelofs and Jed Buchanan. Panpsychism, intuitions, and the great chain of being. *Philosophical Studies*, 176(11):2991–3017, 2018.

[9] Silvia Cardoso. Specialized functions of the cerebral cortex, 1997. https://cerebromente.org.br/n01/arquitet/cortex_i.htm.

[10] Thomas White. *In defense of dolphins: The new moral frontier*. Oxford: Blackwell, 2007.

[11] Thomas White. A primer on nonhuman personhood, cetacean rights and 'flourishing', 2013. http://indefenseofdolphins.com/wp-content/uploads/2013/07/primer.pdf.

[12] Stephen Petersen. The ethics of robot servitude. *Journal of Experimental & Theoretical Artificial Intelligence*, 19(1):43–54, 2007. https://www.stevepetersen.net/petersen-ethics-robot-servitude.pdf.

[13] Aldous Huxley. *Brave New World*. 1932. https://www.huxley.net/bnw/two.html.

[14] Norbert Wiener. God and golem, inc.: A comment on certain points where cybernetics impinges on religion, 1964. https://www.scribd.com/document/2962205/God-and-Golem-Inc-Wiener and http://simson.net/ref/1963/GodAndGolemInc.pdf.

[15] Brian Cantwell. *The promise of artificial intelligence: Reckoning and judgment*. Cambridge, MA: MIT Press, 2019.

[16] Drew McDermott. *Mind and Mechanism*. Cambridge, MA: MIT Press, 2001.

[17] Eric Dietrich. Homo sapiens 2.0: Why we should build the better robots of our nature. *Journal of Experimental & Theoretical Artificial Intelligence*, 13(4):323–328, 2001.

[18] Eric Dietrich. After the humans are gone. *Journal of Experimental & Theoretical Artificial Intelligence*, 19(1):55–67, 2007. https://bingweb.binghamton.edu/~dietrich/Papers/apocalyptic-philosophy/AHG-PN.01%20copy%202.pdf.