Da li kompjuteri treba da donose odluke umesto nas?

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

Marko, Irina
mi17243@alas.matf.bg.ac.rs

13. jun 2022.

Sažetak

U ovom tekstu će biti prevedena jedna od glava knjige "Filozofija informatike"od autora Vilijama J. Rapaporta

Sadržaj

1 Uvod

2004. godine, kada sam prvi put predavao predmet na kome se zasniva ovaj tekst, nije se mnogo raspravljalo o pitanju da li treba verovati odlukama koje donose kompjuteri. Ali od pojave samovozećih automobila, to je postalo hitnije pitanje, sa neposrednim, stvarnim, praktičnim implikacijama, kao i moralnim i pravnim posledicama.

Pre nego što razmotrimo etičko pitanje da li računari treba da donose odluke umesto nas i jasno povezano pitanje koje je naslovljeno u eseju Džejmsa Mura iz 1979. godine : Postoje li odluke koje računari nikada ne bi trebalo da donose?, postoje dva prethodna pitanja: Šta je „odluka“ ? i : Da li kompjuteri uopšte „donose odluke“?

2 Šta je odluka?

Otprilike, odluka je izbor napravljen između nekoliko alternativa, obično iz nekog razloga (Eilon, 1969) [1]. Počnimo sa razmatranjem tri vrste odluka:

1. Odluka može biti rezultat proizvoljnog izbora, kao što je bacanje novčića, ako padne glava "Idemo u bioskop", ako je rep "Ostajemo kod kuće da gledamo tv".
2. Odluka može biti rešenje čisto logičkog ili matematičkog problema koji zahteva neku kalkulaciju.
3. Odluka bi mogla biti rezultat istraživanja prednosti i nedostataka različitih alternativa, racionalnog procenjivanja ovih prednosti i nedostataka, a zatim izbora jedne od alternativa na osnovu ove procene.

3 Da li kompjuteri donose odluke?

Na prvi pogled je jednostavan odgovor na pitanje da li kompjuteri donose odluke: Da; kompjuteri mogu lako doneti prvu vrstu odluke umesto nas. Štaviše, svaki put kada kompjuter reši logički ili matematički problem, on donosi odluku druge vrste.

Mogu li kompjuteri da donose odluke treće vrste? Sigurno je, čini se, odgovor, opet, „da“: kompjuteri mogu da igraju strateške igre, kao što su dame, šah i Go, i mogu da ih igraju tako dobro da mogu da pobede (ljudske) svetske šampione. Takve igre uključuju izbore između alternativnih poteza koji se moraju proceniti, pri čemu se, nada se, donosi najbolji (ili najmanje najgori) izbor. Računari to mogu.

Naravno, nije samo fizički kompjuter taj koji donosi odluku. Moguće je da je kompjuterski program koji se izvršava od strane kompjutera koji donosi odluku, iako ću nastaviti da govorim kao da je kompjuter taj koji odlučuje. I, naravno, nije samo kompjuterski program taj koji donosi odluku. Kompjuterske programe pišu ljudi. (Čak i kompjuterski programi koje pišu računari su rezultat kompjuterskih programa koje su napisali ljudi.) I ljudi, naravno, mogu pogrešiti na različite načine, nenamerno ili na neki drugi način. Ove greške mogu naslediti programi koje pišu. Robin K. Hil je tvrdio da kompjuteri ne donose odluke, upravo zato što njihove programe pišu ljudi. Ljudi su ti koji donose odluke koje su naknadno kodirane u programima. I Mulainatan (2019) [2] tvrdi da je „pristrasne algoritme lakše popraviti nego pristrasne ljude“.

Ali šta se dešava kada je programer čovek van slike, a računar koji pokreće taj program je ono na šta se oslanjamo? U bilo kojoj situaciji, kada računar mora da reaguje ili da donese ili preporuči odluku, on će to učiniti autonomno i u svetlu tadašnje trenutne situacije, bez konsultacija sa programerom (ili u mogućnosti da konsultuje programera).

Obično se ljudska delegacija ovlašćenja za donošenje odluka na računare dešava u slučajevima kada su velike količine podataka uključene u donošenje odluke ili u kojima se odluke moraju donositi brzo i automatski. I, naravno, s obzirom na to da je jedan od ciljeva CS-a da odredi koji su zadaci iz stvarnog sveta izračunljivi , otkrivanje koje su odluke izračunljive je aspekt toga. U svakom slučaju, ljudi mogu dati takvu moć kompjuterima. U svakom slučaju, ljudi mogu delegirati takvu moć kompjuterima. Dakle, drugi način da izrazimo naše pitanje je: Koje oblasti našeg života treba da budu kontrolisane kompjuterom, a koje oblasti treba da budu prepuštene kontroli ljudi? Da li postoje odluke koje ne-ljudski računari ne bi mogli da donesu tako dobro kao ljudi? Na primer, mogu postojati situacije u kojima postoje senzorna ograničenja koja sprečavaju kompjuterske odluke da budu potpuno racionalne. Ili mogu postojati situacije u kojima odluka zahteva izvesnu (verovatno neuračunljivu) empatiju. S druge strane, mogu postojati situacije u kojima bi računar mogao imati prednost u odnosu na ljude.

Da li sada postoje odluke koje kompjuter ne bi mogao da donese tako dobro kao i čovek(empirisjko je pitanje). Mnoge, ako ne i većinu, primedbi na ograničenja računara najbolje je posmatrati kao istraživačke probleme: ako neko kaže da računari ne mogu da rade X, trebalo bi da pokušamo da napravimo one koji rade X. Ovo je ključno: ljudi treba da budu kritički mislioci. Postoji logička zabluda koja se zove Apel na autoritet: Samo zato što autoritet kaže da je nešto istina, iz toga logički ne sledi da je istina. Iako nas logičari ponekad upozoravaju na ovu zabludu, prihvatljivo je žaliti se autoritetu (čak i kompjuteru!) sve dok je konačna odluka vaša. Možete — i morate — odlučiti da li ćete verovati autoritetu ili da verujete računaru. Takođe bi trebalo da budete u mogućnosti (i voljni!) da dovedete u pitanje autoritet—ili kompjuter!—kako biste razumeli razloge za odluku.

Dakle, čak i ako dozvolimo kompjuterima da donose (određene) odluke umesto nas, i dalje je važno da možemo da razumemo te odluke. Kada je moj sin tek učio da vozi, nisam želeo da se oslanja na automatizovani sistem „dinamičkog tempoma“ u vozilu, jer sam želeo da zna kako i kada da uspori ili ubrza na super-autoputu. Kada je znao kako to da uradi, mogao je da se osloni na kompjuter automobila koji će doneti te odluke umesto njega, jer, ako bi veštine kompjuterskog odlučivanja pošle po zlu ili bi bile nedostupne. Drugo pitanje koje proizilazi iz činjenice da kompjuterske programe pišu ljudi jeste da li su, s obzirom na povremenu iracionalnost ljudskog ponašanja, kompjuterski donete odluke zaista racionalne, čemu se sada okrećemo.

4 Da li su kompjuterske odluke racionalne?

Kada odluka ima uticaj na naše živote, želeli bismo da proces donošenja odluka bude racionalan, bilo da odluku donosi čovek ili je „donosi“ ljudski napisan program. Mogu li računari (i programi koje izvršavaju) biti potpuno racionalni? Svakako se čini da neki računari umesto nas mogu doneti racionalne odluke. Čini se da su vrste donošenja odluka opisane u

prethodnom odeljku čisto racionalne. I zar algoritmi zasnovani na pravilima nisu čisto racionalni?

Razmotrite algoritam koji ne uključuje bilo kakvu nasumične ili interaktivne procedure proizvedene od strane neracionalnog proročišta. Verovatno, ako odluku donosi računar koji prati takav algoritam, onda je ta odluka čisto racionalna. Pod 'racionalnim' ne mislim nužno da je to čisto logična odluka. Može, na primer, da uključuje empirijske podatke, koji mogu biti na neki način pogrešni: može biti nepotpun, može biti statistički netačan, može biti pristrasan, i tako na.

Drugi potencijalni problem je ako algoritam zahteva eksponencijalno vreme ili je NP-kompletan, ili čak i ako bi samo trebalo duže da se donese odluka od vremena potrebnog za akciju. U tom slučaju, ili ako ne postoji takav algoritam, morali bismo da se oslonimo na „zadovoljavajuću" heuristiku u smislu algoritma čiji je izlaz „dovoljno blizu" „tačnom" rešenju. Ali ovo je i dalje neka vrsta racionalnosti — ono što je Simon nazvao „ograničenom" racionalnošću.

Da li kompjuteri treba da donose odluke umesto nas je ekvivalentno tome da li naše odluke treba da se donose algoritamski. A to sugeriše da je to ekvivalentno tome da li naše odluke treba da se donose racionalno. Ako postoji algoritam za donošenje date odluke, zašto se onda ne osloniti na njega? Uostalom, zar to ne bi bilo racionalno? Neko bi čak mogao da tvrdi da ne postoji takva stvar kao što je kompjuterska etika. Sva pitanja o moralnosti korišćenja računara da bi se nešto uradila su zaista pitanja o moralnosti korišćenja algoritama. Sve dok su algoritmi racionalni, pitanja o moralnosti njihovog korišćenja su zaista pitanja o moralnosti racionalnosti, i čini se neuverljivim tvrditi da ne treba da budemo racionalni.

Moorovo (1979) [3] pitanje, „Postoje li odluke koje računari nikada ne bi trebalo da donose?", zaista ne bi trebalo da ima nikakve veze sa računarima! Pitanje bi zaista trebalo da bude: da li postoje odluke koje ne treba donositi na racionalnoj osnovi?

Ali onda postaje važno pitanje: da li su algoritmi zaista racionalni? A kako bismo saznali? Pre nego što pogledamo ova pitanja, pretpostavimo, za trenutak, da je algoritam za donošenje odluka racionalan. Sledeće pitanje je: da li treba da dozvolimo da ona odlučuje umesto nas?

5 Da li kompjuteri treba da donose odluke umesto nas?

Paragraf duboko ukorenjen iz naučnog članaka iz 2004. sugeriše da je ljudima teško da prihvate racionalne preporuke, čak i ako dolaze od drugih ljudi, a ne od računara. Članak izveštava o dokazima da se određena popularna i uobičajena hirurška procedura upravo pokazala da nema koristi: „Dr. Hilis je rekao da je pokušao da objasni pacijentima dokaze, ali bez uspeha. „Na kraju ćete dostići nivo frustracije", rekao je. „Mislim da su razgovarali sa nekim ko ih je ubedio da će im ovaj postupak spasiti život" (Kolata, 2004) [4]. Možda fundamentalno pitanje nije da li kompjuteri treba da donose racionalne odluke ili preporuke, već da li ili zašto ljudi treba ili ne prihvataju racionalne savete! Postoji nekoliko razloga zašto bismo možda želeli da dozvolimo računaru da donosi odluku umesto nas: Računari su mnogo brži od nas u proceni opcija, mogu da procene više opcija nego što bismo mi mogli (u istom vremenskom periodu), bolji su u procenjujući složenije opcije, mogu imati pristup relevantnijim podacima. I, u mnogim situacijama u savremenom svetu, možda jednostavno

nemamo drugu opciju osim da dozvolimo računarima da donose odluke umesto nas. Dakle, da li je dobra ideja ili loša ideja da im to dozvolite, to je jednostavna činjenica da to čine.

I, na kraju krajeva, da li se ovo razlikuje od toga da dozvolimo nekom drugom da odlučuje umesto nas — nekom ko je mudriji, ili bolje upućeniji, ili neutralniji od nas? Ako nije drugačije, onda — u oba slučaja — i dalje postoji pitanje koje treba uvek postavljati: da li treba da verujemo odluci tog drugog agenta? Pre nego što pogledamo ovo, postoji srednja pozicija koju treba da razmotrimo.

6 Da li kompjuteri treba da donose odluke sa nama?

Mur sugeriše da, ako računari mogu doneti određene odluke barem jednako dobro kao ljudi, onda bi trebalo da im dozvolimo da to učine, a na nama ljudima bi bilo da prihvatimo ili odbacimo odluku računara. Na kraju krajeva, kada tražimo savet stručnjaka za medicinska ili pravna pitanja, slobodni smo da prihvatimo ili odbijemo taj savet. Zašto isto ne bi važilo i za kompjutersko odlučivanje? Drugim rečima, umesto da jednostavno dozvolimo računarima (ili drugim ljudima) da donose odluke umesto nas, trebalo bi da sarađujemo u procesu donošenja odluka, tretirajući računar (ili stručnjaka za ljude) kao koristan izvor informacija i sugestija da nam pomogne da donesemo konačnu odluku. Ali ovo nije uvek moguce. Mogu (i najverovatnije će biti) situacije u kojima nemamo vremena da procenimo sve opcije pre nego što se donese odluka. Situaciju koju Smit pominje detaljno je istražio antropolog i kognitivni naučnik Edvin Hačins. Hačinsova teorija „distribuirane kognicije“ [5] koristi primere velikih pomorskih brodova koji plove i mlaznih pilota koji rade u svojim kokpitima. U oba ova slučaja, ni mašine same (uključujući, naravno, kompjutere) niti ljudi sami donose odluke ili rade posao, već njihova kombinacija—zaista, u slučaju velikih mornaričkih brodova, to su timovi ljudi, kompjutera i drugih tehnologija. Hačins sugeriše da ova kombinacija čini „distribuisani“ um. Slično tome, filozofi Endi Klark i Dejvid Čalmers (1998) [6] razvili su teoriju „proširene spoznaje“, prema kojoj naš (ljudski) um nije omeđen našom lobanjom ili kožom, već se „proširuje“ u spoljašnji svet kako bi uključuju stvari kao što su sveske, referentni radovi i računari.

Ali mora li biti slučaj da složeni sistemi donošenja odluka budu takvi „hibridi” ili „timski napori”? Smit, Hačins i Klark i Čalmers razvili su svoje teorije mnogo pre pojave samovozećih automobila. Čak i od ovog pisanja (2019), ostaje da se vidi da li će samovozećim automobilima i dalje biti potrebna ljudska intervencija (zapamtite: liftovima koji se sami voze ne treba mnogo toga!): Steven E. Shladover (2016) [7] tvrdi da će nivo koji se zove „uslovna automatizacija“, u kojem računari i ljudi rade zajedno, biti teže postići od potpuno automatizovanog nivoa koji se zove „visoka automatizacija“. Ipak, takvi „hibridni“ ili „prošireni“ sistemi će verovatno ostati realnost.

7 Da li treba da verujemo odlukama koje kompjuteri donose?

Bilo da dozvoljavamo kompjuterima da donose odluke umesto nas, ili radimo zajedno sa njima na donošenju odluka, obično pretpostavljamo da je svaka odluka koju oni donesu ili savet koji daju zasnovani na dobrim dokazima (kao ulaz) i na racionalnim algoritmima (koji obrađuju ulazne podatke). Da bi odluka (ili savet) bila „dobra“ ili da bi zaključak bio tačan, unos mora biti tačan ili tačan i obrada mora biti ispravna. Ali kako da znamo da li jesu? Pouzdanost algoritma je funkcija njegovog unosa i njegove obrade. Da li dobija sve relevantne podatke? Da li je unos tačan ili možda postoji problem sa senzorima ili kako tumači unos? Da li je algoritam ispravan? Možemo li to razumeti? Možemo li objasniti ili opravdati njegove odluke? Da li je (namerno ili nenamerno) na neki način pristrasan, možda zbog načina na koji ga je njegov programer napisao ili – u slučaju programa za mašinsko učenje – kakav je bio njegov početni set za obuku?

Kako se može proceniti kompetentnost kompjuterskog odlučivanja? Jedan odgovor je: na isti način na koji se ocjenjuje ljudska kompetencija za donošenje odluka, naime, putem evidencije o donošenju odluka i njegovih opravdanja za svoje odluke.

Hajde da prvo ukratko razmotrimo rezultate rada računara. Razmislite još jednom o kompjuteru bez dokumenata koji je pronađen u pustinji. Pretpostavimo da otkrijemo da ona uspešno i pouzdano rešava određenu vrstu problema za nas. Čak i ako ne možemo da razumemo zašto i kako to radi, čini se da nema razloga da mu ne verujemo. Dakle, zašto bi opravdanja bila važna? Na kraju krajeva, ako računar stalno nadmašuje ljude u nekom zadatku donošenja odluka, zašto bi bilo važno kako to radi?

U stvari, Evropska unija je donela zakon koji korisnicima daje pravo na objašnjenje odluke računara koja se tiče njih ljudskih prava . Opravdanja, naravno, ne moraju biti ista kao ljudska opravdanja. Kao prvo, ljudska opravdanja mogu biti pogrešna ili nelogična. Ali šta ako su opravdanja nedostupna ili, možda još gore, obmanjujuća? Hajde da pogledamo ove dve mogućnosti.

7.1 Problem pristrasnosti

Može li postojati skrivena pristrasnost u načinu na koji su algoritmi razvijeni? Na primer, skup za obuku koji se koristi za kreiranje algoritma za mašinsko učenje je možda bio pristrasan (opet, možda nenamerno). Ovo ne mora biti zbog bilo kakve namere programera da obmane. Do kakvih problema mogu dovesti takve „slabosti“ ili pristrasnosti? Korisnici su otkrili da Google-ova aplikacija za fotografije, koja primenjuje automatske oznake na slike u digitalnim foto albumima, klasifikuje slike crnaca kao gorile. Google se izvinio; bilo je nenamerno. Nikonov softver za kameru pogrešno pročitati slike azijskih ljudi kao da trepću, itd Ovo je u osnovi problem podataka. Algoritmi uče tako što dobijaju određene slike, koje često biraju inženjeri, a sistem gradi model sveta na osnovu tih slika. Ako je sistem obučen na fotografijama ljudi koji su pretežno beli, biće mu teže da prepozna lica drugih boja. Možda se etička pitanja zaista tiču prirode različitih vrsta algoritama. „Uredni“ algoritmi su zasnovani na formalnoj logici i dobro razvijenim teorijama predmeta algoritma. “Scruffy” algoritmi nisu nužno zasnovani na bilo kojoj formalnoj teoriji.

„Heuristički“ algoritmi ne mora vam dati tačno rešenje za problem, ali bi trebalo da daju ono koje je dovoljno blizu ispravnog rešenja da bi

bilo korisno (tj. ono koje „zadovoljava“). Algoritmi za mašinsko učenje se obučavaju na skupu test slučajeva i „uče“ kako da rešavaju probleme na osnovu tih slučajeva i određene tehnike učenja koja se koristi

Ako je „uredan“ algoritam „tačan“ — sigurno, veliko „ako“ — onda izgleda da nema nikakvog moralnog razloga da se ne koristi (da ne bude „ispravno racionalan“). Ako je algoritam „otrcani“, onda bi neko mogao imati moralne dileme. Ako je algoritam heuristički onda nema više ili manje moralnog razloga da se koristi nego da se veruje ljudskom stručnjaku. Ako je algoritam razvijen mašinskim učenjem, onda će njegova pouzdanost zavisiti od njegovog skupa za obuku i metoda učenja.

7.2 Problem crne kutije

"Niko zaista ne zna kako najnapredniji algoritmi rade to što rade. To bi mogao biti problem"— Vil Najt (2017) [8]

"Algoritamska pravičnost: Ako vaš alat ne može da objasni svoje rezultate, ne bi trebalo da ga koristite."— Venkatasubramanian (2018) [9]

Postoje najmanje četiri izvora problema koji odluku računara mogu učiniti nepouzdanom:

1. kriterijume za donošenje odluka kodirane u algoritmu, bilo od strane njegovog programera (ili programera) ili od strane programa za mašinsko učenje koji je razvio te kriterijume iz test slučajeva,
2. sami ti test slučajevi,
3. sam kompjuterski program, i
4. podatke na kojima se zasniva data odluka.

Pretpostavimo, radi argumenta, da su ulazni podaci (4) što je moguće potpuniji i tačniji. Pretpostavimo takođe (iako je ovo mnogo veća pretpostavka) da je algoritam (3) formalno verifikovan. To ostavlja kriterijume za donošenje odluka i sve testne slučajeve kao primarni fokus pažnje.

U sadašnjoj fazi razvoja računara, dva načina na koja se ovi kriterijumi nalaze u algoritmu su, prvo, preko ljudskog programera i, drugo, kroz mašinsko učenje. Naravno, algoritam za mašinsko učenje dobija svoje test slučajeve od čoveka (ili iz baze podataka koju je generisao drugi program koji je napisao čovek), a tehniku mašinskog učenja dobija od svog programera. Ali kada je čovek van slike, a algoritam je prepušten sam sebi, da tako kažem, algoritmu se moramo obratiti za objašnjenje.

Shodno tome, jedno važno pitanje u vezi sa računarima koji donose odluke za nas (ili sa) jeste da li oni mogu ili treba da objasne svoje odluke. Za ovo pitanje su relevantne dve vrste algoritama. Jedna vrsta je simbolički ili logički algoritam koji ima takvu mogućnost objašnjenja. To može imati na jedan od dva načina: korisnik može da ispita trag algoritma, ili programer može da napiše program koji bi preveo taj trag u objašnjenje na prirodnom jeziku koje bi korisnik mogao da razume. Druga vrsta algoritma je onaj koji se zasniva na neuronskoj mreži ili na statističkom algoritmu mašinskog učenja. Takav algoritam možda neće moći da objasni svoje ponašanje, niti njegov programer ili korisnik mogu da razumeju kako i zašto se tako ponaša.

Da li treba da „zahtevaju veru“? Ili bi zakoni (kao što su oni u Evropskoj uniji) trebalo da zahtevaju transparentnost ili objašnjivost i da na taj način isključe algoritme mašinskog učenja iz „crne kutije“ ? Oslanjanje na uspešne, ali neobjašnjive odluke računara ne mora nužno značiti da odluke donosimo na osnovu vere. Na kraju krajeva, njeni uspesi bi sami po

sebi bili dokaz njegove verodostojnosti, baš kao što je korisnost aksioma u matematičkim izvodima dokaz u njegovu korist iako se – po definiciji – ne može dokazati.

8 Da li postoje odluke koje kompjuteri moraju da donose umesto nas?

Zapamtite Simonov problem ograničene racionalnosti: obično nemamo vremena ili sposobnosti da procenimo sve relevantne činjenice pre nego što treba da reagujemo. Šta je sa hitnim slučajevima ili drugim situacijama u kojima nema vremena za čoveka koji mora da reaguje da uključi preporuku računara u svoja razmatranja?

1. jula 2002. ruski avion se srušio na teretni avion iznad Nemačke, pri čemu su poginuli svi u avionu, uglavnom studenti. Rekorder leta ruskog aviona imao je automatski sistem za izbegavanje sudara koji je instruisao pilota da ide više (da preleti teretni mlaznjak). Ljudski kontrolor letenja rekao je ruskom pilotu da ide niže (da leti ispod teretnog mlaznjaka). Prema naučnom reporteru Džordžu Džonsonu, „piloti imaju tendenciju da slušaju kontrolora letenja jer veruju ljudskom biću i znaju da osoba želi da ih zaštiti" . Ali ljudski kontrolor vazdušnog saobraćaja bio je umoran i prezaposlen. A kompjuterski sistem za izbegavanje sudara nije „želeo" ništa; jednostavno je donosilo racionalne sudove. Pilot je sledio odluku čoveka, a ne kompjutersku, i dogodila se tragedija. Postoji interesantan kontrastni slučaj. U januaru 2009. godine, nakon nesreće sa pticama koje su se zaglavile u njegovim motorima, mlaznjak US Airways bezbedno je „sleteo" na reku Hadson u Njujorku, spasivši sve u avionu i od svog pilota napravio heroja. Ipak, Villiam Langeviesche (2009) [10] tvrdi da je avion, sa svojim kompjuterizovanim sistemom „fly by wire", bio pravi heroj. Drugim rečima, junaštvo pilota je bilo zbog njegove spremnosti da prihvati odluku kompjutera.

9 Da li postoje odluke koje kompjuteri ne bi trebalo da donose?

Pretpostavimo da imamo kompjuter za donošenje odluka koji objašnjava sve svoje odluke, nepristrasan je i ima odlične rezultate. Postoje li odluke koje čak ni takav kompjuter nikada ne bi trebalo da donese? Kompjuterski naučnik Džozef Vajzenbaum (1976) [11] je tvrdio da, čak i ako računar može da donosi odluke jednako dobro, ili čak bolje od čoveka, ne bi trebalo, posebno ako se njihovi razlozi razlikuju od naših. I Mur ističe da, moguće, računari ne bi trebalo da imaju moć da donose (određene) odluke, čak i ako imaju kompetenciju za to (barem kao, ako ne i bolje od ljudi). Ali, ako imaju nadležnost, zašto ne bi imali moć? Na primer, pretpostavimo da veoma sujeverna grupa pojedinaca donosi loše medicinske odluke u potpunosti zasnovane na njihovim sujeverima; zar ne bi trebalo da prednost ima „spoljna" medicina savremenog lekara? I da li činjenica da su računari imuni na ljudske bolesti znači da im nedostaje empatija da preporučuju tretmane ljudima? Mur sugeriše da, iako računar treba da donosi racionalne odluke umesto nas, računar ne bi trebalo da odlučuje koji bi naši osnovni ciljevi i vrednosti trebalo da budu. Računari bi trebalo da nam pomognu da postignemo te ciljeve ili zadovoljimo te vrednosti,

ali ne bi trebalo da se menjaju njih. Ali zašto ne? Računari ne mogu biti pravno ili moralno odgovorni za svoje odluke, jer nisu osobe. Barem ne još. Ali šta ako VI uspe? Batia Friedman i Peter H. Kahn, Jr. (1997) [12] tvrde da su ljudi — ali računari nisu — sposobni da budu moralni agenti i da bi, prema tome, računari trebalo da budu dizajnirani tako da:

1. ljudi nisu u „samo mehaničkim“ ulogama sa smanjenim osećajem za delovanje, i
2. kompjuteri se ne maskiraju kao agenti sa verovanjima, željama ili namerama.

Hajde da razmotrimo tačku 1: Fridman i Kan tvrde da kompjuteri treba da budu dizajnirani tako da ljudi shvate da su oni (ljudi) moralni agenti. Ali šta ako računar ima bolju evidenciju o donošenju odluka od ljudi? Fridman i Kan nude studiju slučaja APACHE, kompjuterskog sistema koji može donositi odluke o tome kada pacijentu uskratiti održavanje života. Prihvatljivo je ako se koristi kao sredstvo za pomoć ljudima koji donose odluke. Ali ljudski korisnici mogu doživeti „smanjenje osećaja moralnog delovanja“ kada ga koriste, verovatno zato što je uključen računar.

Ali zašto? Pretpostavimo da je APACHE zamenjen udžbenikom o tome kada treba uskratiti održavanje života, ili stručnjakom za ljude. Da li bi bilo koji od njih umanjio osećaj moralnog delovanja čoveka koji donosi odluke? U stvari, zar ljudi koji donose odluke ne bi bili neuredni ako nisu uspeli da konsultuju stručnjake ili spise stručnjaka? Pa zar i oni ne bi bili nemarski ako ne bi konsultovali stručni kompjuter?

Možda bi ljudi iskusili ovaj smanjeni osećaj moralnog delovanja iz sledećeg razloga: ako APACHE-ove odluke pokažu „dobar učinak“ i na njih se više oslanja, onda ljudi mogu početi da popuštaju njegovim odlukama. Ali zašto bi to bilo loše?

Vraćajući se na tačku 2, Fridman i Kan tvrde da kompjutere treba dizajnirati tako da ljudi shvate da računari nisu moralni agenti. Fridman i Kan tvrde da treba da budemo oprezni u vezi sa antropomorfnim korisničkim interfejsima, jer izgled verovanja, želja i namera ne implicira da ih oni zaista imaju. Ovo je klasična tema, ne samo u istoriji veštačke inteligencije, već i u književnosti i bioskopu. A ovo je u srcu Tjuringovog testa u VI.

Literatura

[1] Samuel Eilon. *What is a Decision?*, volume 16. INFORMS, 1969.

[2] Sendhil Mullainathan. Biased algorithms are easier to fix than biased people. *The New York Times*, 2019.

[3] James H Moor. Are there decisions computers should never make? In *Computer Ethics*, pages 395–407. Routledge, 2017.

[4] Gina Kolata. New studies question value of opening arteries. *New York Times, pages A*, 1:A21, 2004.

[5] Edwin Hutchins. *Cognition in the Wild*. MIT press, 1995.

[6] Andy Clark and David Chalmers. The extended mind. *analysis*, 58(1):7–19, 1998.

[7] Steven E Shladover. What 'self-driving'cars will really look like. *Scientific American*, pages 54–57, 2016.

[8] Will Knight. The dark secret at the heart of ai'11 april 2017, 2017.

[9] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018.

[10] William Langewiesche. *Fly by wire: the geese, the glide, the miracle on the Hudson*. Farrar, Straus and Giroux, 2009.

[11] Joseph Weizenbaum. Computer power and human reason: From judgment to calculation. 1976.

[12] Batya Friedman and Peter H Kahn Jr. People are responsible, computers are not. *Computers, ethics, and society*, 2, 1997.

[13] William J. Rapaport. *Philosophy of Computer Science*. University at Buffalo, The State University of New York, 2020.