Etika umjetne inteligencije

Puzek, Željka

Master's thesis / Diplomski rad

2018

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, University of Zagreb, Faculty of Humanities and Social Sciences / Sveučilište u Zagrebu, Filozofski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:131:795477

SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

ODSJEK ZA FILOZOFIJU

Željka Puzek

ETIKA UMJETNE INTELIGENCIJE

Diplomski rad

Mentor: izv. prof. dr. sc. Hrvoje Jurić

Zagreb, studeni 2018.

Sadržaj

Etika umjetne inteligencije

Sažetak:

Tehnologija je postala neodvojim dijelom ljudskih života, iako toga često nismo ni svjesni. Razvoj umjetne inteligencije je više-manje stagnirao od samog nastanka ove grane znanosti, ali u posljednjih nekoliko godina, možemo svjedočiti vrlo značajnom napretku za koji se očekuje da će kroz nekoliko godina dotići ultimativni cilj razvoja jake umjetne inteligencije – sustava koji će doista misliti i posjedovati svijest. U radu će biti predstavljeni najvažniji trenuci razvoja umjetne inteligencije, a zatim i etičke implikacije koje taj razvoj nosi, kao i mogući rizici koje moramo imati na umu te mogući načini prevencije kako bi se minimizirala mogućnost katastrofalnih posljedica. Također, pokušat ćemo utvrditi koja je uloga i odgovornost etike u ovom složenom procesu.

Ključne riječi: umjetna inteligencija, neuralne mreže, ekspertni sustavi, etika umjetne inteligencije, moralni status, egzistencijali rizici

Ethics of Artificial Intelligence

Abstract:

Technology has become an inescapable part of human life, although we are often unaware of it. The development of artificial intelligence has been more or less stagnant since the creation of this branch of science, but over the last few years we can witness a very significant progress that is expected to overcome the ultimate goal of developing a strong artificial intelligence – a system that will truly think and possess awareness. This paper will present the most important moments in the development of artificial intelligence, followed by its ethical implications, as well as the possible risks we must keep in mind, and possible ways of prevention to minimize the possibility of catastrophic consequences. We will also try to identify the role and responsibility of ethics in this complex process.

Key words: artificial intelligence, neural networks, expert systems, ethics of artificial intelligence, moral status, existential risks

Uvod

Apokaliptični prizori budućnosti, izazvani pobunom tehnologije i razvijanjem svijesti u inteligentnim sustavima, jedna su od omiljenijih tema znanstveno-fantastičnih filmova, serija i romana. Do prije nekoliko godina, na takve smo prizore gledali s istom količinom zabrinutosti kao da gledamo nešto u čemu su glavni protagonisti zli čarobnjaci i vile. Međutim, u posljednjih nekoliko godina, velik broj znanstvenika je započeo izjavljivati da vjeruju da će stvaranje sustava umjetne inteligencije ugroziti postojanje čovječanstva kao cjeline. Znanstvenici smatraju da će se u bliskoj budućnosti tehnologija kontinuirano poboljšavati, a njen se brz razvoj neće moći zaustaviti. Predviđa se da će takav razvoj tehnologije dovesti do povećanja broja nezaposlenih, ali i do gubitka odgovornosti čovjeka za ono što se događa i za odluke koje se donose, a one mogu imati i negativne posljedice, s obzirom na to da umjetna inteligencija može nadmašiti sposobnosti čovjeka te će stoga upravljati znanstvenim istraživanjima, razvojem oružja kao i koristiti se u različite ekonomske i financijske svrhe.

Dakle, ideja koja je prije pedesetak godina bila tek fikcija u današnje vrijeme je postala stvarnost. Scene koje smo prije viđali samo u znanstveno-fantastičnim filmovima ili smo o njima čitali u znanstveno-fantastičnoj literaturi (postojanje robota, korištenje robota u različite svrhe, u područjima zabave, medicine, vojnih akcija i sl.) postale su dijelom realnosti. Stoga se, neizbježno, nameće i pitanje kako umjetna inteligencija može biti korištena u skladu s etičkim principima, što bi osiguralo i njezino sigurno korištenje. Nemoguće je u potpunosti eliminirati rizike koje nosi umjetna inteligencija, međutim, ljudi moraju poduzeti potrebne mjere kako bi se osigurala kontrola, a to je proces u kojem je potrebno da sudjeluju stručnjaci iz različitih područja, između ostalog i etike. Potrebno je razviti procedure koje će osigurati testiranje inteligentnih strojeva i razinu sigurnosti, a da se ne ograničava razvoj i pojava novih inovativnih tehnologija. Ljudski um i svijest još uvijek su nedostižni za kopiranje, zasad čak i za vjernu simulaciju, međutim, dio istraživača umjetne inteligencije teži stvaranju upravo navedenog – računala koje će inteligencijom nadmašiti čovjeka a istovremeno imati svijest i doista misliti. Bi li to značilo da će čovječanstvo u slučaju postizanja tog cilja, prepustiti rješavanje najsloženijih problema i pitanja od globalne važnosti, kao što su svjetski mir, globalno zagrijavanje i slično, sustavima koji posjeduju umjetnu inteligenciju? Što bi to značilo za ljude i koliko smo uopće blizu ostvarenju toga

cilja, neka su od pitanja koja se postavljaju u ovome radu. Imajući u vidu ogroman pozitivni potencijal koji umjetna inteligencija nosi, ali istovremeno i rizik s kakvim se do sad nismo susreli, jasno je da postoji potreba za promišljanjem na koji način se nositi s takvim potencijalom, kao i na koji način osigurati da naše najveće dostignuće ne bude i ono što će nas kao čovječanstvo osuditi na propast.

1. Umjetna inteligencija

1.1. Objašnjenje pojmova inteligencije i umjetne inteligencije

Pod pojmom inteligencije, najčešće podrazumijevamo „praktično snalaženje u novim prilikama“ [1] ili „otkrivanje zakonitosti u odnosima među činjenicama“. [2] U psihologiji, inteligencija označava „sposobnost mišljenja koja omogućuje snalaženje u novim prilikama u kojima se ne koriste (ili nemaju dobar ishod) nagonsko ponašanje, ni učenjem stečene navike, vještine i znanja. O ustroju inteligencije postoje prijepori: je li to jedinstvena sposobnost, tj. odraz općih funkcionalnih značajki središnjega živčanoga sustava, ili je inteligencija skup širih sposobnosti koje su odgovorne za intelektualno djelovanje (kao što su perceptivna, spacijalna, numerička, mnemička te sposobnost rječitosti i zaključivanja). Stupanj intelektualne razvijenosti pojedinaca određuje se pomoću testova inteligencije“. [3] Iako je inteligentno ponašanje prisutno i u životinjskome svijetu, istraživanja su uglavnom usmjerena na ljudsku inteligenciju.

No, što točno označava pojam umjetne inteligencije (UI[4])? Postoje brojne definicije umjetne inteligencije, a ona se obično definira kao „znanost o tome kako napraviti da računala rade stvari koje zahtijevaju inteligenciju kad ih obavljaju ljudi“, [5] odnosno „sposobnost digitalnog računala ili robota kojeg kontrolira računalo da obavlja zadatke obično povezane s inteligentnim bićima“.[6] „UI ne samo da proizvodi umjetne sustave koji 'oponašaju' određene ljudske kognitivne sposobnosti, već teži stvoriti računalne sustave koji posjeduju sve te sposobnosti ljudskoga uma.“[7] Radovan smatra da bi UI bolje odgovarao naziv „funkcionalna inteligencija“, o čemu će biti riječi kasnije.

Generalno, definicije pojma UI mogu se podijeliti na definicije koje su fokusirane na:

razmišljati ljudski	razmišljati racionalno
ponašati se ljudski	ponašati se racionalno

[1] Antun Vujić (ur.), *Opća i nacionalna enciklopedija u 20 knjiga*, Pro Leksis d.o.o. i Večernji list d.o.o., Zagreb, 2006., knjiga 9, str. 157.
[2] Vladimir Anić, *Rječnik hrvatskoga jezika*, Novi liber, Zagreb, 1991., str. 206.
[3] Hrvatska enciklopedija, http://www.enciklopedija.hr/natuknica.aspx?ID=27600.
[4] Dalje ću u radu koristiti skraćeni termin za umjetnu inteligenciju – UI.
[5] Jack Copeland, „What is Artificial Intelligence?“, 2000., http://www.alanturing.net/turing_archive/pages/Reference%20Articles/What%20is%20AI.html#Int.
[6] B. J. Copeland, „Artificial Intelligence“, *Encyclopædia Britannica*, 2018. https://www.britannica.com/technology/artificial-intelligence.
[7] Mario Radovan, „The Way of Power: Reflections on Technology, Nature and Society“, 2008., http://www.inf.uniri.hr/~mradovan/powerdocs/P2_Mind_and_computation.pdf.

Dakle, podjela je izvršena na definicije koje se tiču misaonih procesa i rasuđivanja ili ponašanja i one koje se mjere prema vjernosti ljudskoj izvedbi ili prema idealnoj ideji inteligencije koju nazivamo racionalnost.[8]

Termin „umjetna inteligencija“ nastao je 1956. godine u Dartmouthu, na radionici na kojoj su se okupili istraživači zainteresirani za područje automatizacije, neuralne mreže i proučavanje inteligencije. Pojam je kreirao John McCarthy, koji UI definira kao znanost i inženjering izrade inteligentnih strojeva, osobito inteligentnih računalnih programa.[9]

Kao što se može primijetiti iz samo nekoliko navedenih definicija UI, ona se neizbježno uspoređuje s ljudskom inteligencijom, odnosno, njen cilj je i povijesno gledano, bilo stvaranje računalnih sustava koji će oponašati, tj. pokazivati karakteristike svojstvene ljudskoj inteligenciji. Istraživanja na području UI fokusirana su na specifične probleme i pokušavaju replicirati sposobnosti koje smatramo odlikama inteligencije (razumijevanje i rješavanje problema, učenje, zaključivanje i razumijevanje jezika).

Mogućnost da se dostigne cilj razvitka UI – stroj koji ne samo da oponaša ljudsku inteligenciju, već istu i posjeduje, odnosno, posjeduje um, pred nama otvara velik broj pitanja. Je li znanost blizu tom cilju i što bi to uopće značilo za čovječanstvo?

1.2. Razvoj umjetne inteligencije

Alan Turing je, 1936. godine, predstavio ideju svog univerzalnog stroja, kasnije poznatog kao Turingov stroj. Radi se o jednostavnom, apstraktnom uređaju za manipulaciju znakovima, koji mogu biti prilagođeni da simuliraju logiku bilo kojeg računalnog algoritma. Stroj ima beskonačnu traku i glavu koja se pomiče po traci, u skladu s unaprijed određenim pravilima te detektira simbole i zapisuje nove simbole, može čitati i brisati, a s promjenom parametara unutar memorije, mijenja se i funkcija. Univerzalni stroj je prethodnik današnjih računala. Godine 1950. Turing predstavlja i svoju ideju testiranja umjetne inteligencije, tzv. Turingov test.

Gordon Moore (suosnivač Intela) je 1965. godine, iznio predviđanje koje je nazvano Mooreovim zakonom i koji je i u današnje vrijeme, 50-tak godina kasnije, još uvijek aktualan

[8] Stuart J. Russell; Peter Norvig (ur.), *Artificial Intelligence: A Modern Approach*, Prentice-Hall, Upper Saddle River, 2003., str. 1.
[9] John McCarthy, „What is Artificial Intelligence?“, 2007., http://www-formal.stanford.edu/jmc/whatisai.pdf.

i primjenjiv. Moore je spomenute godine predvidio da će se broj tranzistora koji se po najpovoljnijoj cijeni mogu smjestiti na čip udvostručavati otprilike svake dvije godine. Neminovno je da će se u nekom trenutku iscrpiti mogućnost konstantnog povećavanja snage čipova, ponajprije zbog fizičke ograničenosti, ali primjerice, u usporedbi s prvim Intelovim mikroprocesorom, današnji 14-nanometarski procesori nude tri i pol tisuće puta bolje performanse i devedeset tisuća puta veću učinkovitost. Prvi poluvodički tranzistori bili su veličine gumice na olovci, a sada u točku na kraju ove rečenice stane više od šest milijuna 22-nanometarskih tranzistora.[10]

Početkom razvoja UI smatra se prvi model umjetnoga neurona, odnosno model koji je nastao istraživanjem neurofizioloških karakteristika živih bića, 1943. godine. Warren McCulloch i Walter Pitts su predstavili model umjetnih neurona koji mogu imati dva stanja – „on" (pobuđujuće) ili „off" (umirujuće). Pobuđujuće stanje se aktivira uslijed stimulacije dovoljnog broja susjednih neurona.[11] Prvi jednostavni model neuronskih mreža, nazvan je perceptron. Budući da je u to vrijeme računalna tehnologija bila na samim začecima svoga razvoja, važnost neuronskih mreža nije odmah prepoznata. Neuronska mreža je masivno paralelni distribuirani procesor za pamćenje iskustvenog znanja te je slična mozgu po tome što se znanje stječe kroz proces učenja i međusobne veze između neurona se koriste za spremanje znanja. Jedna je od čestih primjena neuronskih mreža, primjena za prepoznavanje uzoraka kao što su analize slika, govora, signala, podataka i slično. Primjerice, neuronska mreža koja je trenirana za obradu pisanih riječi i pretvaranje istih u audio zvukove (izgovor napisanih riječi), sposobna je nakon nekoliko sati treniranja na uzorku od 1000 riječi, pravilno izgovarati riječi na engleskom jeziku (ista slova prema drugačijem izgovoru u različitim riječima – npr., b*a*ll, h*a*ve, s*a*ve).[12]

„Mreža živaca[13] je niz procesora međusobno povezanih sponama koje mogu biti pojačane ili oslabljene; koncept se inspirirao međusobno povezanim neuronima mozga. Mreža živaca „trenira" se davanjem niza primjera točnih odgovora a spone među njezinim procesorima se pojačavaju ili slabe ovisno o uspjehu reproduciranja onog što se zahtijeva."[14]

[10] Nick Bostrom, „How Long Before Superintelligence?", 2008., https://nickbostrom.com/superintelligence.html.
[11] S. J. Russell; P. Norvig (ur.), *Artificial Intelligence: A Modern Approach,* str. 16.
[12] Paul M. Churchland, *Matter and Consciousness*, A Bradford Book, The MIT Press, Cambridge, Massachusetts, 1999., str. 163.
[13] Collins upotrebljava izraz „mreža živaca", koji je u potpunosti ekvivalentan izrazu „neuronske mreže".
[14] Harry Collins, „Hoće li strojevi ikada misliti?", *Treći program Hrvatskog radija*, 48, 1995., str. 5–10, ovdje str. 8.

Tijekom druge polovice 1980-tih, nastao je novi pristup UI, nazvan *nouvelle UI*, čiji je začetnik Rodney Brooks. Nouvelle UI se fokusira na ostvarivanje performansi na razini insekta. U osnovi ideje nouvelle UI odbacuju se ciljevi simboličke UI koja teži replikaciji ljudskih sposobnosti. Pobornici nouvelle UI tvrde da prava inteligencija uključuje sposobnost funkcioniranja u stvarnom okruženju, odnosno, osnovna postavka je da se inteligencija, izražena kao kompleksno ponašanje, ustvari pojavljuje kao produkt većeg broja jednostavnih radnji. Poznati primjer nouvelle UI je Brooksov robot Herbert, koji djeluje u uredima laboratorija na MIT-u. Herbert pretražuje stolove i skuplja prazne limenke koje zatim odnosi u smeće. Ponašanje robota proizlazi iz interakcije 15-tak jednostavnih radnji. U novije vrijeme, Brooks je konstruirao i prototipove mobilnih robota za istraživanje površine Marsa.[15]

Istraživanje umjetne inteligencije jest pisanje programa u već nekom postojećem jeziku ili kreiranje novog programskog jezika. Ukoliko se programi pišu u već nekom postojećem jeziku, sljedovi instrukcija ovise o empirijskim spoznajama kako inteligentna bića rješavaju probleme, a ukoliko se izmišlja novi programski jezik, do toga dolazi jer se postojeći smatraju neadekvatnima te postoji potreba za novim naredbama koje će moći na bolji način modelirati inteligentno rješavanje problema. Stoga se UI može smatrati empirijskom znanošću.[16]

Inteligencija se ponekad definira kao sposobnost rješavanja problema ili obavljanje aktivnosti koje su usmjere na postizanje nekog cilja. Računala nemaju ni probleme ni ciljeve, pa se javlja pitanje kako bismo njihovo ponašanje uopće mogli nazvati inteligentnim. Promatrač može interpretirati njihovu aktivnost kao rješavanje problema ili usmjerenost ka cilju, ali računalo ne traži rješenje jer je *svjesno* da ima problem koji treba riješiti, već traži rješenje jer je za to *programirano*. Nema motivacije koja bi utjecala na rješavanje i htijenje rješavanja nekog problema. Stoga, potrebno je razlikovati dvije vrste inteligencije: *autentičnu* i *funkcionalnu*. *Autentična inteligencija* potječe iz egzistencijalnih potreba i osjećaja (mentalnih stanja) svjesnog bića. Primarno se manifestira kroz sposobnost postavljanja vrijednosti i ciljeva u kontekstu egzistencijalnih potreba i osjećaja. Takva inteligencija ne može postojati bez života i osjećaja iz kojih izvire. *Funkcionalna inteligencija* sastoji se u sposobnosti izvođenja različitih jasno (formalno) definiranih procesa, i može biti

[15] B. J. Copeland, „Nouvelle Artificial Intelligence“, *Encyclopædia Britannica*, 2018., https://www.britannica.com/technology/nouvelle-artificial-intelligence.
[16] Davor Pećnjak, „Umjetna inteligencija: a priori ili empirijska znanost“, u: Zvonimir Čuljak (ur.), *Zbornik radova međunarodnog simpozija „Spoznaja i interpretacija“*, Institut za filozofiju, Zagreb, 2010., str. 151–157, ovdje str. 153.

materijalizirana kroz računalni sustav.[17] Štoviše, računala vrlo često već daleko nadmašuju funkcionalnu inteligenciju ljudi. Da bismo stvorili entitet s autentičnim mentalnim sposobnostima, prvo je potrebno kreirati entitet koji je osjetljiv i svjestan, koji ima svoju motivaciju i brige. Drugim riječima, da bismo stvorili sustav s autentičnim mentalnim sposobnostima, potrebno je stvoriti sustav koji je živ. Mogu li računala voljeti, mrziti, patiti, brinuti o nečemu ili nekomu? Odgovor na ovakva pitanja nameće se sam po sebi i glasi „ne". Međutim, s jedne strane, dokazati točnost tog odgovora nije moguće, a s druge strane, za sad nema empirijskog razloga zašto bismo smatrali da nije točan. Strojevi funkcioniraju, ali oni *nisu* – ako *biti* znači biti svjesnim, brinuti i uživati, voljeti i mrziti i slično.

Učenje je vrlo važan proces za kreiranje UI. Jedan od načina učenja je jednostavno memoriranje, odnosno, pohranjivanje već pronađenih rješenja za određene probleme. Time se pri susretanju s istim problemom u budućnosti, iz memorije jednostavno „izvuče" rješenje. Drugi način se temelji na heurističkom pristupu (primjerice u šahovskim programima) u kojem program bilježi omjer gubitaka i pobjeda te na temelju tih podataka prilagođava svoje poteze.[18] Međutim, noviji pristup rješavanju tog problema su i neuralne mreže koje su već spomenute, a o kojima će još biti riječi i kasnije.

1.2.1. Turingov test

Kao što je već spomenuto, 1950. godine, predstavljen je Turingov test koji je zbog svoje važnosti postao temeljem razvoja UI. Turing u svom radu predlaže „igru oponašanja" kako bi usporedio performanse računala i čovjeka. U testu sudjeluju tri sudionika, muškarac (A), žena (B) i ispitivač (C) koji može biti bilo kojeg spola. Cilj testa je da ispitivač utvrdi tko je od osoba s kojima razgovora žena, a tko muškarac. Fizički je odvojen od ispitanika te se komunikacija odvija posredno, primjerice, putem računalnog sučelja kako se ne bi čuli nikakvi glasovi. Polazeći od te ideje, Turing zatim postavlja pitanje što će se dogoditi ako računalo preuzme ulogu ispitanika A? Hoće li ispitivač uspjeti utvrditi tko je računalo a tko čovjek?

Računalo uspješno prolazi test ukoliko ljudski ispitivač, nakon postavljanja određenih pitanja, ne može utvrditi je li odgovore dalo računalo ili čovjek. Ako ispitivač ne može prosuditi, tada stroj treba smatrati inteligentnim.[19]

[17] M. Radovan, „The Way of Power: Reflections on Technology, Nature and Society".
[18] P. M. Churchland, *Matter and Consciousness*, str. 114.
[19] S. J. Russell; P. Norvig (ur.), *Artificial Intelligence: A Modern Approach*, str. 948.

Da bi računalo prošlo test, vrlo je važno da ima sposobnost korištenja prirodnog jezika, kao i da posjeduje veliku količinu zdravorazumskog znanja. Svladavanje prirodnog jezika računalima stvara prilično velik problem, a posjedovanje zdravorazumskih činjenica (primjerice, uže se može vući, a ne gurati) još veći. Test je bihevioristički, jer da bi ga prošao, sustav treba znati kako odgovoriti na odgovarajući način, ali ne mora uopće razumjeti o čemu govori. Stoga Turingov test, prema Radovanu, može procjenjivati samo funkcionalnu inteligenciju.[20]

1.2.2. Kineska soba[21]

Osim pitanja mogu li strojevi misliti, odnosno, mogu li imati um, pitanje s kojim se susrećemo odnosi se i na samu prirodu uma. Funkcionalizam je teorija koja smatra da se mentalnim stanjem može nazvati bilo koje međusobno uzročno stanje između „inputa" i „outputa". Prema tome, računalo isto može imati mentalna stanja, kao i čovjek.[22] Najutjecajnija vrsta funkcionalizma je komputacijski funkcionalizam (ili kompjutacionalizam) prema kojoj je mentalno stanje analogno softveru u računalu. Um je softver a mozak hardver. John Searle, američki filozof, odbacuje ideje funkcionalizma te ne vjeruje da je Turingov test valjani test dokaza inteligencije. 1980. godine objavio je svoju raspravu „Kineska soba" kojom pokušava dokazati da stroj može proći Turingov test, ali to ne znači da razumije išta od ulaznih i izlaznih informacija.

Njegov sustav opisuje situaciju u kojoj se nalazi osoba koja zna samo engleski jezik. Spomenuta osoba, nalazi se u zatvorenoj prostoriji u kojoj na raspolaganju ima knjigu s napisanim pravilima na engleskom i hrpom papira, od kojih su neki prazni, a na nekima su simboli koje ne zna dešifrirati. U ovom slučaju, osoba glumi računalo, knjiga s pravilima je program a papiri su radna memorija. Kroz otvor u prostoriji, osoba zaprima papiriće na kojima su pitanja, napisana na kineskom jeziku. Čovjek, koristeći knjigu s pravilima te slijedeći instrukcije na papirima, zapisuje odgovore u obliku simbola (kineskog jezika) na prazne papire te ih kroz drugi, izlazni otvor, šalje van iz sobe. Glavna Searleova teza je da će, iako osoba koja s tim papirićima operira, uopće ne razumije kinesko pismo niti je svjesna da zaprima pitanja na koja zatim daje odgovore, ukoliko pravilno slijedi upute, bez poteškoća ispravno obavljati zadatak. Searle stoga zaključuje da se oponašanje inteligentnog ponašanja

[20] M. Radovan, „The Way of Power: Reflections on Technology, Nature and Society".
[21] John R. Searle, „Umovi, mozgovi i programi", u: Nenad Miščević, Nenad Smokrović (ur.), *Računala, mozak i ljudski um*, Izdavački centar Rijeka, Rijeka, 2001., str. 134–154, ovdje str. 134.
[22] S. J. Russell; P. Norvig (ur.), *Artificial Intelligence: A Modern Approach,* str. 954.

ne može uzeti kao indikator istinske inteligencije. Računalo dobiva ulazne podatke, manipulira s njima s pomoću programa i daje izlazne podatke bez da o tome ima ikakvo razumijevanje ili poznavanje sadržaja.

1.3. Problem razumijevanja

1952. godine Arthur Samuel konstruirao je niz programa za igranje dame koji su ubrzo naučili igrati bolje od svog kreatora, čime je Samuel opovrgnuo tezu da računala mogu raditi samo ono što im se zada.[23]

1991. godine, u Bostonu, skupina stručnjaka ispitivala je niz računalnih programa kad je program nazvan PC Therapist III pokazao najbolje rezultate dotad, odnosno, najmanje se u odgovorima razlikovao od ljudskih ispitanika. Međutim, pitanja su bila ograničena na uska područja za svaki program i konverzacija se održavala na engleskom jeziku. Prema Searlovu dokazu Kineske sobe, stroj može proći na Turingovom testu čak i ako ne razumije o čemu se radi. Međutim, neki istraživači smatraju ukoliko se test pažljivo provodi, moguće je otkriti osjetljivost na kontekst i socijalizaciju, što je razumno nazvati inteligencijom.[24] Turing je 1950. godine, predstavljajući svoj test inteligencije strojeva, predvidio da će računala u roku od 50 godina moći proći test ali dosad još uvijek nijedno računalo nije uspješno prošlo spomenuto testiranje. Loebner nagrada je godišnje natjecanje u području UI, koja se dodjeljuje računalnim programima koje suci proglase najsličnijima ljudima. Format provođenja testiranja je standardni Turingov test: ljudski sudac istovremeno razgovara tekstualnim putem s računalnim programom i ljudskim bićem i na temelju njihovih odgovora, mora odlučiti tko je tko.[25]

Hubert Dreyfus, filozof na Sveučilištu California u Berkleyu, napisao je 1972. godine, knjigu pod nazivom „Što računala NE mogu“. U spomenutoj knjizi, dokazivao je da je sposobnost inteligentnih strojeva vrlo ograničena, a svoju argumentaciju je temeljio na pravilima. Naime, Dreyfus tvrdi da „pravila ne sadrže pravila vlastite primjene“ jer ista ne mogu sadržavati sve nužne informacije o kontekstu u kojem će biti primijenjena budući da bi to impliciralo postojanje još pravila koja objašnjavaju i ta pravila i tako u beskonačnost. Harry Collins naziva taj problem „kritikom pravila“.[26] Dreyfus smatra da računala ne mogu biti

[23] Isto, str. 18.
[24] H. Collins, „Hoće li strojevi ikada misliti?“, str. 9.
[25] Više o tome, skupa s transkriptima vođenih razgovora, dostupno je na adresi: https://www.aisb.org.uk/events/loebner-prize.
[26] H. Collins, „Hoće li strojevi ikada misliti?“, str. 6.

uspješna u potpunosti ni u područjima u kojima pravila mogu obuhvatiti sve mogućnosti, jer bi primjerice – nedostajao možda kontekst. Kao primjer, navodi prijevode s jezika na jezik i šah. Što se tiče strojnog prevođenja, kasnije ćemo vidjeti primjer kako i u kojoj mjeri je postignut poprilično zadivljujući napredak. Što se tiče šaha, znamo da je u tom slučaju pogriješio. 1997. godine, računalo Deep Blue (IBM), pobijedilo je svjetskog šahovskog prvaka Garija Kasparova. Pobjeda računala, ostavila je dio javnosti u šoku, kao i samog prvaka Kasparova budući da je ljudski mozak izgubio premoć u igri koja se dotad diljem svijeta smatrala pokazateljem iznimne inteligencije.[27] U novije vrijeme, metode na kojima se ovakvi programi temelje, sve više napreduju te vrlo uspješno pobjeđuju ljude. Program 'Deep Fritz' 2006. godine pobijedio je Kasparovljevog nasljednika Vladimira Kramnika 4-2, a radilo se o običnom PC računalu, a ne nekom snažnom superračunalu.

2011. godine, superračunalo Watson, pobijedilo je dva najuspješnija šampiona kviza „Izazov!". Radilo se o novoj prekretnici u razvoju umjetne inteligencije, jer u ovoj igri više nije bila dovoljna samo sirova snaga računanja kao što je to slučaj u šahu. „Izazov!" zahtijeva razumijevanje sintakse i gramatičke logike, složeno povezivanje opskurnih činjenica i formuliranje odgovora u obliku gramatički ispravnog pitanja. Od 2012. godine, Watson je 'stalno zaposlen' kao liječnik dijagnostičar u centru za onkologiju u New Yorku. Sposobnost povezivanja podataka i korištenje baze tisuća povijesti bolesti, omogućilo je Watsonu da postane puno učinkovitiji od ljudskih liječnika u postavljanju ispravne dijagnoze.

AlphaGo (Googleov program) nova je evolucija tehnologije neuronskih mreža. Istraživači su ga trenirali s nekoliko tisuća snimljenih partija između ljudskih velemajstora, a zatim ga pustili da igra s modificiranom verzijom samog sebe. Za par mjeseci, AlphaGo je pobijedio prvog velemajstora, europskog prvaka Fan Huija, a zatim i svjetskog prvaka Lee Sedola, 2016. godine. Igra Go stara je više od 2500 godina te ima veoma jednostavna pravila ali istovremeno pruža gotovo neograničen broj mogućih kombinacija i situacija. AlphaGo bi odigrao potez koji se doimao kao da nema smisla, no kasnije bi postao jasan – tako je u jednom trenutku odigrao odgovor na potez koji je Sedol tek planirao povući, što je šokiralo i samog prvaka.

Najnovija prekretnica u razvoju UI odnosi se na prebacivanje fokusa s tzv. *board games* na video igre. U osnovi, video igre nude izazove koje igre na pločama, kao što su šah i Go, nemaju. U njima su informacije skrivene od igrača, nema pregleda cijele karte odjednom,

[27] S. J. Russell; P. Norvig (ur.), *Artificial Intelligence: A Modern Approach,* str. 180.

što znači da UI ne može mapirati virtualnu kartu i na temelju toga izračunati najbolji mogući sljedeći korak. U video igrama je prisutno i daleko više informacija za obradu i velik broj mogućih poteza. Početkom mjeseca kolovoza 2018. godine, botovi konstruirani u istraživačkom laboratoriju OpenAI pobijedili su bivše profesionalne igrače u jednoj od najpopularnijih virtualnih igara, Dota 2. Istraživači u OpenAi obučavali su botove putem tehnike pojačanog učenja, što je uključivalo postavljanje botova u virtualni svijet i korištenje principa pokušaja i pogrešaka kako bi sami shvatili kako postići određeni cilj. Dakle, bili su potpuno prepušteni sami sebi te su jedino nagrade identificirane kroz bodove koji su dobivali kad su uspješno odradili određeni zadatak (što je i osnova tehnike pojačanog učenja, što će biti malo opširnije pojašnjeno kasnije u radu). Iz OpenAI su naveli kako su pošli od pretpostavke da ljudima treba najmanje 12.000 sati da bi postali profesionalni igrači, a njihovi botovi su svakodnevno igrali sami protiv sebe, u trajanju koje bi preračunato u ljudsko vrijeme iznosilo 180 godina. U nekim trenucima, OpenAI botovi u Dota 2, moraju birati između 1000 različitih akcija, tijekom obrade 20.000 podatkovnih točaka koje predstavljaju ono što se događa u igri. Krajem mjeseca kolovoza 2018. godine, tim botova igrao je protiv trenutnih prvaka, u igri 5 vs. 5, i izgubio. Međutim, napredak se ipak smatra kao izvanredno postignuće u razvoju UI.[28]

Područje istraživanja UI koje je uvijek frustriralo istraživače budući da nisu ostvarivali zadovoljavajuće rezultate, odnosi se na područje razumijevanja i komunikacije putem prirodnih jezika. Sintaktičke strukture (odnosno sintaksu, manipulaciju formalnim simbolima) bilo je moguće ostvariti, međutim, semantika (značenje) je stvarala poteškoće. Prepoznavanje prirodnog jezika zahtijeva široko znanje o vanjskom svijetu i sposobnost da se njime manipulira. Definicija razumijevanja je jedan od glavnih problema u obradi prirodnog jezika. Računala mogu manipulirati simbolima, odnosno sintaksom, kako bi dali odgovarajuće odgovore na pitanja prirodnog jezika, međutim, ne „razumiju“ rečenice, odnosno, ne mogu riječima pridodati značenja. Jedan od prvih programa s kojim se mogla voditi konverzacija, konstruirao je Joseph Weizenbaum, a ime mu je ELIZA. ELIZA je program koji je u razgovoru oponašao psihoterapeuta.[29] Iz transkripata razgovora, čini se kao da ELIZA doista razgovora, dok u stvarnosti, nema nikakvog razumijevanja o pojmovima koje pacijenti koriste, već u razgovoru samo koristi riječi koje oni upotrebljavaju, uz nekakve standardne

[28] Više o projektu i rezultatima dostupno je na službenim stranicama OpenAI istraživačkog laboratorija: https://blog.openai.com/the-international-2018-results/.
[29] P. M. Churchland, *Matter and Consciousness*, str. 117.

obrasce i pitanja koji su programirani.[30] Dakle, računalo je svoja pitanja postavljalo na temelju programiranih pitanja kao i ponavljalo riječi pacijenta, formulirajući tako nova pitanja kako bi pacijent otkrio nove informacije koje programu mogu služiti za daljnju komunikaciju.

Program koji je pokazao veći napredak u svladavanju semantike i sintakse je SHRDLU, kojeg je kreirao Terry Winograd, negdje između 1968-1970. godine. Program je kontrolirao robotsku ruku nad ravnom površinom na kojoj su se nalazila različita geometrijska tijela. Na temelju dobivenih instrukcija, program je izvršavao zadatke i objašnjavao svoje radnje. Iako je pokazivao izvrsne rezultate u svladavanju i razumijevanju prirodnog jezika, s obzirom na to da je djelovao samo u ograničenom mikrosvijetu, tj. „svijetu geometrijskih tijela" i razgovori su bili vezani usko uz navedeno područje. Početni optimizam je splasnuo kad su se razgovori proširili na realnije situacije, sa svim složenostima i dvosmislenostima stvarnog svijeta.[31] Međutim, bilo bi pogrešno smatrati da su programi UI samo empiristički eksperimenti. Oni su naprotiv, eksperimenti mišljenja, koje na umjetan način izvode računala te stoga važnost SHRDLU programa nije samo u tome što govori i razumije engleski, već i u tome da je istražio najosnovnije zahtjeve za sve sustave, neovisno o tome kojim pravcem u ljudskom govoru oni krenuli i kako mijenjali riječi.[32]

UI još uvijek nije razvila program s kojim bi se moglo komunicirati u potpunosti kao s drugim ljudskim bićem, ali se istraživanja i dalje nastavljaju. Primjenu prvenstveno pronalaze u razvijanju računalnih sučelja, provjeri gramatike i pravopisa prilikom pisanja na računalu, prijevoda i slično. 2016.godine, Google je svoj sustav za statističko strojno prevođenje (temeljen na frazama) zamijenio sustavom temeljenom na dubokim neuronskim mrežama (sequence-to-sequence learning) što je smanjilo pogreške u prijevodu čak i do 60%. Korištenjem milijuna primjera, poboljšana je kvaliteta prijevoda pomoću šireg konteksta kako

[30] Dio razgovora Elize i pacijentice (pitanja koja postavlja Eliza otisnuta su tiskanim slovima):
Svi su muškarci isti.
U KOJEM SMISLU?
Uvijek nas zbog nečeg gnjave.
MOŽEŠ LI MI NAVESTI NEKI PRIMJER?
Pa, moj me dečko natjerao da dođem ovamo.
TVOJ TE DEČKO NATJERAO DA DOĐEŠ OVAMO.
On kaže da sam gotovo stalno depresivna.
ŽAO MI JE ČUTI DA SI DEPRESIVNA.
To je istina, nesretna sam.
MISLIŠ LI DA ĆE TI DOLAZAK OVAMO POMOĆI DA NE BUDEŠ NESRETNA?
Treba mi pomoć, to je sigurno.

[31] P. M. Churchland, *Matter and Consciousness*, str. 118.

[32] Daniel C. Dennet, „Umjetna inteligencija u filozofiji i psihologiji", u: Nenad Miščević, Nenad Smokrović (ur.), *Računala, mozak i ljudski um*, Izdavački centar Rijeka, Rijeka, 2001., str. 43–63, ovdje str. 52.

bi se ponudio najbolji prijevod. Rezultat je prijevod koji je preoblikovan i gramatički prilagođeniji prirodnom jeziku.[33]

Kao posljedica pokušaja kreiranja mikrosvjetova, nastali su ekspertni sustavi. Ekspertni sustavi su inteligentni računalni programi koji koriste znanja i postupke zaključivanja kako bi se riješili teški zadaci, unutar specijaliziranog područja znanja. Prvi programi koji su predstavljali ekspertne sustave, bili su DENDRAL i MYCIN. DENDRAL je vršio kemijske analize a njegova stručnost je izvedena iz velike baze posebnih pravila. MYCIN je dizajniran radi dijagnosticiranja krvnih infekcija na temelju analize simptoma i proučavanja krvnih rezultata testova. Razlika u odnosu na DENDRAL bila je u tome što nije postojao općeniti teorijski model iz kojeg je MYCIN mogao koristiti pravila, već je do informacija dolazio intervjuirajući stručnjake, koji su svoje znanje stekli iz knjiga, od drugih stručnjaka ili kroz direktno iskustvo. Oba programa su bila vrlo uspješna u svom području te su se mogli mjeriti s vrhunskim specijalistima iz istoga područja. Ekspertni sustavi se i u današnje vrijeme nalaze u širokoj upotrebi budući da su dizajnirani i stvoreni da bi rješavali zadatke u područjima financiranja, medicine, proizvodnje, računovodstva, automobilske industrije itd. Nedostatak ekspertnih sustava je to što oni nisu savršena umjetna inteligencija, već su dosta usko orijentirani i često mogu griješiti ako ih se koristi za rješavanje pojedinih problema izvan domene za koju su stvoreni.

Kako bi se poboljšala uspješnost ekspertnih sustava, 1984. godine, započet je projekt CYC, koji je ujedno i najveći eksperiment u simboličkoj UI. Douglas Lenat, započeo je projekt popunjavanja CYC baze podataka mnoštvom jednostavnih činjenica i zdravorazumskog znanja. Milijuni tvrdnji i pravila, uneseni su u bazu. Cilj projekta je prikupiti i strukturirati zdravorazumsko znanje i obrasce razmišljanja koje ljudi obično posjeduju i koriste, a što čini osnovu ljudskog diskursa i razmišljanja. Očekivanje je bilo da će se dosezanjem kritične mase, postići da sustav samostalno izvlači pravila što bi omogućilo veću uspješnost, inače usko specijaliziranih ekspertnih sustava. Problem s takvom količinom podataka je ažuriranje, pretraživanje i drugo manipuliranje velikom strukturom simbola u

[33] Mike Schuster, Melvin Johnson, Nikhil Thorat, „Zero-Shot Translation with Google's Multilingual Neural Machine Translation System“, Google Research Blog, 2016., https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html.

realnom vremenu, odnosno, pitanje je hoće li baza zbog svoje kompleksnosti i tromosti uopće biti upotrebljiva.[34]

1.4. Problem svijesti

Cilj umjetne inteligencije (AI) je razvoj paradigmi i algoritama koji omogućuju strojevima da obave zadaće koje zahtijevaju shvaćanje (inteligenciju) kad ih obavljaju ljudi. Od inteligentnog sustava se očekuje da pohranjuje znanje (reprezentacija), primjenjuje znanje da riješi određeni problem (zaključivanje) i prikuplja novo znanje (učenje).

Pretpostavka je da se uz odgovarajući hardver i programiranje, može postići da strojevi uče na isti način kao što uče i djeca, odnosno, uz interakciju s odraslim ljudima i drugim objektima u svojoj okolini. Ukoliko želimo da superinteligencija imitira funkcioniranje ljudskog mozga, potrebno je da ista i posjeduje neki oblik „umjetnog mozga". Neuroznanost posljednjih godina napreduje te je pitanje vremena kad ćemo znati dovoljno o neuralnoj strukturi mozga i njegovim algoritmima za učenje kako bi se isti mogli replicirati u računalo. Hoće li taj napredak jamčiti i stvaranje svijesti u računalima? Kad uspoređujemo ljudski mozak s najnaprednijim računalima, čini se da je mozak još uvijek nepobjediv. Računala velikom brzinom mogu obrađivati podatke, ali za svaku obradu im treba zasebno vrijeme komputacije, dok ljudski mozak iste izvodi simultano. Činjenica je da do sada računala nisu uspjela proizvesti svijest, ali postoje teoretičari koji navode da svijest ni nije od esencijalne važnosti za kognitivne aktivnosti. Ljudi najčešće nisu u potpunosti svjesni mentalnih aktivnosti koje se odvijaju u njihovom mozgu. Radovan se s time ne slaže. On navodi kako činjenica da neki ljudi ponekad nisu svjesni svojih kognitivnih sposobnosti i ponašanja, nije dovoljna da se zaključi da stroj koji nije nikad svjestan može razmišljati u doslovnom smislu te riječi.[35] Isto se pitanje postavlja i za inteligenciju. Odakle točno dolazi inteligencija? U kojem se točno stupnju kognitivnog sistema čovjeka pojavljuje? Nemamo jasne odgovore ni na ovakva pitanja, pa je stoga teško i inteligenciju replicirati putem računala.

Također, neka od filozofskih pitanja koja se javljaju u vezi s istraživanjima UI su i: Kako um radi? Je li moguće da se strojevi ponašaju inteligentno na način na koji se ljudi ponašaju inteligentno i ako je moguće, znači li to da imaju umove? Koje su etičke implikacije

[34] B. J. Copeland, „CYC Computer Science", *Encyclopædia Britannica*, 2018., https://www.britannica.com/topic/CYC.
[35] M. Radovan, „The Way of Power: Reflections on Technology, Nature and Society".

inteligentnih sustava? Naivan pogled na UI bila bi ideja da je dovoljno da neki genijalni programer stvori program koji će proizvesti svijest kao što je imaju ljudi.

Istraživanje UI dijeli se na *slabu UI* (predstavlja je hipoteza da je moguće da se strojevi ponašaju inteligentno) i *jaku UI* (smatra da strojevi koji se ponašaju inteligentno doista misle). Cilj slabe UI (poznate i kao napredna obrada informacija) je proizvesti komercijalno održive „pametne" sustave kao što su npr. ekspertni sustavi medicinske dijagnostike i sustavi trgovanja dionicama. Kao vrlo dobar primjer slabe UI je i Siri (virtualna pomoćnica u Appleovom operativnom sustavu), koja služi kao moćna baza podataka i čini se vrlo inteligentnom s obzirom na to da, ne samo da je u stanju održati razgovor s ljudima, već dobacuje i primjedbe i šali se. Međutim, ona djeluje na unaprijed definiran način, a ta „ograničenost" je evidentna ukoliko se pokuša s njom voditi razgovor za koji nije programirana da vodi i odgovora. Roboti koji se koriste u proizvodnom procesu također mogu djelovati inteligentno zbog složenosti radnji koje obavljaju, ali oni znaju što da rade u situacijama za koje su programirani i izvan toga nemaju mogućnost određivanja kako će postupiti. Izraz „jaka UI", prvi je upotrijebio filozof John Searle.[36]

Kao treća grana UI javlja se i *kognitivna simulacija*. U kognitivnoj simulaciji, računala se koriste za testiranje teorija o tome kako ljudski um funkcionira – na primjer, teorije o tome kako ljudi prepoznaju lica ili druge predmete, kako se prisjećaju ili rješavaju apstraktne probleme. Kognitivna simulacija je već sada moćno oruđe u neuroznanosti i kognitivnoj psihologiji.[37] Kognitivna simulacija se razlikuje od ostalih pristupa jer uzima u obzir ljudske performanse i pogreške te preuzima ljudski pristup rješavanju problema. Informacije se tretiraju kao pamćenje, emocije, percepcija, jezik i zaključivanje. Međutim, budući da još uvijek nije moguće repliciranje djelovanja neurona smještenih u mozgu, izazov kognitivne simulacije je pronalazak načina da se neuroni simuliraju na računalu. Dobar primjer bihevioralnog robota je KISMET, napravljen 1990. godine na MIT-u. KISMET može simulirati ljudske osjećaje, sreću, tugu, iznenađenje, promjene u kretanju glave, može govoriti različitim tonovima, pomicati trepavice, obrve, uši i usta kako bi odgovarale ljudskim

[36] Jack Copeland, „What is Artificial Intelligence?", 2000., http://www.alanturing.net/turing_archive/pages/Reference%20Articles/what_is_AI/What%20is%20AI02.html.
[37] Isto.

pokretima. KISMET koristi receptore za snimanje slika i zvuka a sustav koji koristi je sintetički živčani sustavi koji koristi kognitivnu simulaciju.[38]

Jedan od zanimljivijih pitanja je mogu li drugi fizički objekti osim neurona u ljudskome mozgu imati mentalna stanja. Alan Turing, u svom slavnom članku „Computing Machinery and Intelligence" navodi da bismo umjesto pitanja „Mogu li strojevi misliti?" trebali postaviti pitanje mogu li strojevi proći test kojim bi se dokazala njihova inteligencija. Test je nazvan Turingov test, kojeg smo već prethodno spomenuli. Velik broj filozofa tvrdi da strojevi koji bi prošli Turingov test i dalje ne bi zaista mislili, već bi samo simulirali proces mišljenja. Turning smatra da je očekivanje odgovora na ovo pitanje pogrešno jer je i samo pitanje „Mogu li strojevi misliti?" pogrešno postavljeno. On navodi da u svakodnevnom životu nemamo nikakve dokaze unutarnjih misaonih procesa koji se odvijaju u drugom ljudskom biću, ali smatramo da se oni ipak odvijaju. No, na koji način možemo dokazati misli li računalo? Nepobitni dokaz bio bi da smo mi sami to računalo. Ali isto tako, jedini način da znamo misli li drugo ljudsko biće, bio bi da smo mi to drugo ljudsko biće. Taj problem se naziva *problem drugih umova*.[39] Turingov odgovor na ovaj argument je da bi alternativa njegovoj „igri oponašanja" bio solipsizam[40], međutim, ne pozivamo se na solipsizam prilikom procjene inteligencije drugih ljudi te ga stoga ne bismo smjeli ni primjenjivati prilikom procjene inteligencije računala.

Slaže se da je pitanje svijesti teško pitanje kad se pokušava primijeniti na UI, međutim, smatra da misterija svijesti ne mora nužno biti riješena kako bi se mogli kreirati inteligentni strojevi. Odnosno, važno je kreirati programe koji se ponašaju inteligentno te nije toliko važno hoće li netko smatrati takvo ponašanje stvarnim ili simuliranim.[41]

Joseph Weizenbaum, iz laboratorija MIT, opisao je konačni cilj jake UI kao „izgradnju stroja prema modelu čovjeka, robota koji će imati djetinjstvo, učiti jezik kao dijete, da stekne svoje znanje o svijetu osvajanjem svijeta kroz svoje vlastite organe i na koncu razmišlja o čitavoj domeni ljudske misli.

[38] Više o samom projektu i robotu može se pronaći na internetskoj stranici: http://www.ai.mit.edu/projects/sociable/overview.html.
[39] Alec Hyslop, „Other Minds", *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), ur. Edward N. Zalta, https://plato.stanford.edu/archives/spr2016/entries/other-minds/.
[40] Solipsizam (novolatinski *solipsismus*, od latinski *solus*: sam, jedini + *ipse*: sam [osobno]) znači stajalište da ne postoji ništa drugo osim nečijega vlastitog uma i mentalnih stanja.
[41] S. J. Russell; P. Norvig (ur.), *Artificial Intelligence: A Modern Approach*, str. 952.

S obzirom na to da se elektrokemijski signali iz ljudskog živčanog sustava mogu prenositi do računala i primati iz računala, tehnologija usadaka (koji se usađuju direktno u ljudsko tijelo), otvara razne mogućnosti poboljšanja ljudskih sposobnosti.[42] Eksperiment moždanih proteza predstavljen je sredinom 1970-tih. Isti se odnosi na pretpostavku da će se neuropsihologija razviti do trenutka kad ćemo u potpunosti razumjeti moždanu aktivnost i funkcioniranje neurona u mozgu. Zatim se izgrađuju mikroskopski elektronički uređaji koji oponašaju ponašanje neurona i mogu se ugraditi u neuralno tkivo. Posljednja pretpostavka je da se nekim čudesnim operativnim zahvatom mogu zamijeniti pojedinačni neuroni s odgovarajućim elektroničkim napravama bez da se narušava aktivnost cijelog mozga. Eksperiment bi se sastojao od postupne zamjene svih ljudskih neurona s elektroničkim uređajima a zatim bi se obrnutim procesom, zamijenili ponovno svi prirodni neuroni. Pitanja koja se u ovom teoretskom scenariju postavljaju, odnose se na vanjsko ponašanje čovjeka u slučaju uspješne zamjene neurona elektroničkim primjercima, jer je pretpostavka da nikakvih odstupanja u ponašanju ne bi smjelo biti. Međutim, u ovom slučaju, postavlja se i pitanje svijesti. Bismo li mogli znati je li svijest osobe ostala nepromijenjena ili je posve istisnuta zamijenjenim elektroničkim mozgom?

Istraživanja u području neouroznanosti i dalje napreduju. Krajem 2017. godine, tim znanstvenika predstavio je 'memorijsku protezu'. Radi se o malenoj sondi s elektrodama koje mogu detektirati i reproducirati signale u hipokampusu, dijelu mozga zaduženom za pamćenje. Usadili su uređaj u mozak pacijenata koji su već imali elektrode u mozgu zbog liječenja epilepsije. Implantat je pratio aktivnost tijekom testa memorije i bilježio reagiranja u hipokampusu kad bi pacijenti uspješno pogodili traženi oblik ili pak pogriješili. Iz prikupljenih uzoraka, kreiran je matematički model koji može predvidjeti kako će neuroni reagirati kad pacijent zapamti objekt koji mu je bio prethodno pokazan. Kad se može previdjeti aktivnost, to znači da se mozak može i stimulirati da oponaša proces uspješnog prisjećanja. Rezultati istraživanja su pokazali da je pamćenje poboljšano u 35% slučajeva. Ovo je jedan od primjera korištenja tehnologije u neuroznanosti, u ovom slučaju za potrebe liječenja bolesti koje uzrokuju gubitak pamćenja, kao što su Alzheimerova bolest i demencija.[43]

[42] Kevin Warwick, „Strojevi koji misle“, u: Sian Griffiths (ur.), *Predviđanja: 30 velikih umova o budućnosti*, Naklada Jesenski i Turk, Zagreb, 2000., str. 317–327, ovdje str. 322.
[43] Više o spomenutom istraživanju može se pronaći na internetskoj stranici *Journal of Neural Engineering*, Volume 15, Number 3, 2018.: http://iopscience.iop.org/article/10.1088/1741-2552/aaaed7/meta.

Pretpostaviti da bismo ljudski mozak mogli skenirati i „uploadati“ u računalo ili radi poboljšavanja kognitivnih sposobnosti mozga, stvarajući ljudsku super-inteligenciju ili radi dosezanja besmrtnosti otvara nova pitanja – bi li taj čovjek ostao „normalan“? Ljudski mozak je vrlo kompleksan, vrlo je lako narušiti ravnotežu koja može okinuti različite poremećaje i nije predviđen za modificiranja i poboljšavanja.[44]

Pokušajmo zamisliti sljedeći scenarij: mozak se skenira putem naprednog skenera visoke rezolucije i u procesu biva uništen jer primjerice, mora biti seciran na tanke komadiće kako bi se mogao skenirati i mapirati. Možemo zamisliti da sken mora biti dovoljno detaljan da uhvati sve neurone, njihove sinaptičke veze i ostale značajke koje su potrebne da bi mozak mogao funkcionirati. Nakon uspješnog mapiranja, isto se ugrađuje u snažno računalo te ukoliko je „upload“ bio uspješan, računalni program bi trebao moći replicirati sve funkcionalne karakteristike originalnog mozga. Razmišljajući o ovakvom scenariju, javlja se mnoštvo pitanja kao što su, možemo li očekivati da će ovako nešto biti moguće u budućnosti? Ako bi program pokazivao istu osobnost, ista sjećanja i isti način razmišljanja kao i originalni mozak, bi li to značilo da program može osjećati? Bi li to bila ista osoba čiji je mozak skeniran i uploadan?[45] Ovom tematikom bave se i neki filmovi i serije, kao što su primjerice *Transcendence* u kojem se prikazuje radnja uploadanja svijest znanstvenika i istraživača UI u računalo, kao jedini način spašavanja svijesti teško ozlijeđenog glavnog lika koji se zatim razvija u entitet željan moći. Kao drugi primjer koji vrijedi spomenuti je tv-serija *Westworld*, koja osim skeniranja ljudskih mozgova s ciljem prebacivanja svijesti u humanoidne robote kako bi se postigla besmrtnost, otvara i velik broj etičkih pitanja u vezi s postupanjem i tretiranjem spomenutih humanoidnih robota koji s vremenom razvijaju svijest a ljudi su se dotad na njima iživljavali na sve moguće načine, bez ikakvih posljedica.

Kad se promišlja o vezi između ljudskog uma i računala, najčešće se zauzima jedan od četiri glavna stava:

a) Ljudski um je *više* nego što će računalo ikad moći biti s obzirom na to da računalo samo izvodi operacije za koje je proizveden i programiran. Ona funkcioniraju, ali ne razumiju jer operiraju samo na razini sintakse. Ljudski um sa svojim mentalnim stanjima, spada u drugu

[44] Eliezer Yudkowsky, „Artificial Intelligence as a Positive and Negative Factor in Global Risk“, Machine Intelligence Research Institute, 2008., https://intelligence.org/files/AIPosNegFactor.pdf.
[45] Nick Bostrom, Eliezer Yudkowski, „The Ethics of Artificial Intelligence“, 2011., https://nickbostrom.com/ethics/artificial-intelligence.pdf.

dimenziju egzistencije i ne može se replicirati običnim procesima koji djeluju na sintaktičkoj razini.

b) Ljudski um je *manje* nego što računalo može biti i manje nego što računalo može postati u budućnosti. Prema nekim teorijskim istraživanjima, moguće je razviti računala koja će uvelike nadmašiti skromne mentalne sposobnosti ljudi.

c) *Kako nam se sviđa* – ljudski um je produkt bioloških procesa koji se odvijaju u mozgu. Iako ti procesi sami po sebi nisu računalni, često je korisno interpretirati ljudski kognitivni sustav kao računalni sustav. To je doduše samo interpretacija realnosti i ne smije se miješati sa stvarnom realnošću.

d) Samo je *metafora*. Svjestan um je najmisteriozniji fenomen s kojim se ljudski um susreo i pokušao razumjeti. Očito je da se svijest pojavljuje iz nesvjesne materije (mozak, tijelo), ali ne znamo kako i zašto se to događa.[46]

Usporedbe mozga i uma s računalnim hardverom i softerom se ne čine valjanima. Um je produkt mozga, dok je softver samo izvana umetnuti dio. Ne čini se ni razumnim očekivati da će se proizvesti softver koji će omogućiti pojavljivanje uma i svijesti, s obzirom na to da ne znamo objasniti kako uopće dolazi do pojave mentalnih stanja u ljudskome mozgu. Pojam razumijevanja kod čovjeka obuhvaća svjesnost o određenoj situaciji ili problemu, kao i kontekst i prijašnja iskustava, što u slučaju računala potpuno izostaje. Ili barem, ne izgleda da možemo dokazati suprotno.[47]

1.5. Ostali problemi umjetne inteligencije

Osnovni problem oko kojeg se razilaze dvije najveće grane UI, slaba i jaka UI odnosi se na inteligenciju i svijest. Problem i oko definiranja same inteligencije, s obzirom na to da ne postoji jednoznačna definicija, otežava situaciju i za istraživače UI jer nijedna teorija nije u stanju objasniti u kojem bi se trenutku neki računalni sistem s pravom mogao nazvati inteligentnim.[48] Upravno na navedenom, protivnici jake UI temelje svoju argumentaciju. Činjenica da se računalo doima inteligentnim ne dokazuje njegovu inteligenciju, a posebice ne dokazuje da ono posjeduje neko mentalno stanje. Također, kritičarima je upitna i sama usporedba ljudske i računalne inteligencije.

[46] M. Radovan, „The Way of Power: Reflections on Technology, Nature and Society“, http://www.inf.uniri.hr/~mradovan/powerdocs/P2_Mind_and_computation.pdf, str. 2.
[47] Isto.
[48] J. Copeland, „What is Artificial Intelligence?“.

Kao što je već navedeno, jaka UI vjeruje da je moguće izraditi umjetni sustav koji će posjedovati mentalna stanja, svijest i um, međutim, na koji način bi se to moglo postići, nije jasno. Kao problem koji se javlja kad se razmišlja o svijesti jest i problem intencionalnosti. Stoga „računarska metafora ne može biti objašnjenje ljudskog uma – sintaksa nije dovoljna; na taj način ne može se objasniti kako mentalna stanja (reprezentacije) imaju sadržaje i svojstvo intencionalnosti, tj. usmjerenosti na neki objekt. Također, računalu ne bismo mogli pridavati svojstva uma jer njegove operacije i simboli nemaju sadržaj.“[49]

Osim problema intencionalnosti, imamo probleme i doživljaja svijeta, kao i motivacije i odsustvo emocija. Strojevi ne mogu doživjeti svijet kao što ga doživljavaju ljudi, već samo izvršavaju naredbe zapisane u svojim programima a emocije su sastavni dio autentične inteligencije. Fred Dretske postavlja pitanje mogu li računala zbrajati i zaključuje da ona uopće ne zbrajaju, budući da je zbrajanje operacija s brojevima, a računala ne izvode nikakve operacije s brojevima već su to operacije s određeni fizičkim jedinicama koje predstavljaju, ili se interpretiraju kao da predstavljaju brojeve. Ono što računalima nedostaje je originalna intencionalnost jer ona operirajući svojim simbolima ništa ne misle, niti shvaćaju njihovo značenje.[50]

Dakle, simboli sami po sebi, bez razumijevanja značenja koje se veže uz njih, ne znače ništa doli jednostavnog operiranja simbolima što ne može izazvati razumijevanje. Ali, na koji način simbolima kojima stroj manipulira dati vlastito značenje? Dretske smatra da je jedini način putem kojeg bi se taj cilj mogao ostvariti, robotika. Odnosno, stavljanje računala u glavu robota, tj. u veći sustav koji posjeduje vrstu senzornih sposobnosti perceptivnih izvora. Iz toga slijedi da je rad na percepciji, strojnom prepoznavanju uzoraka i robotici relevantniji za kognitivne sposobnosti strojeva od najsofisticiranijeg programiranja.[51] Međutim, znači li to da bismo onda trebali pripisivati atribute razumijevanja i nekom modernijem autu koji je opremljen kamerama, senzorima i računalom?[52]

Iz svega dosad navedenog, lako se može zaključiti da je nemogućnost stvaranja svijesti i mentalnih stanja u računalima, zapravo najveća zapreka kreiranju jake UI. Međutim, iako se jaz između biološkog i ne biološkog (umjetnog) čini nepremostivim, zagovornici jake

[49] Davor Pećnjak, „Turingovi strojevi, Gödelov teorem i Searleova soba“, *Treći program Hrvatskog radija*, 48, 1995., str. 11–15, ovdje str. 15.
[50] Fred Dretske, „Strojevi i mentalno“, u: Nenad Miščević, Nenad Smokrović (ur.), *Računala, mozak i ljudski um*, Izdavački centar Rijeka, Rijeka, 2001., str. 85–99, ovdje str. 90.
[51] Isto, str. 94.
[52] Slawomir J. Nasuto, John Mark Bishop, „Of (Zombie) Mice and Animats“, u: Vincent C. Muller (ur.), *Philosophy and Theory of Artificial Intelligence*, Springer, Berlin, Heidelberg, 2013., str. 85–106, ovdje str. 88.

UI i dalje smatraju da ne možemo isključiti kako će se u budućnosti moći razviti stroj koji bi se mogao nazvati inteligentnim.

Najpoznatiji primjer argumenta koji se protivi ideji jake UI je već spomenuti Searleov argument kineske sobe. Iako njegov zaključak (slijepo izvršavanje uputa ne znači i razumijevanje same radnje) naizgled djeluje ispravno, mnogi istraživači su iznijeli svoje kritike na njegov račun. Searle dakle tvrdi da mozak stvara um, odnosno, kroz svoj argument zapravo traži da se definira gdje se u sobi nalazi um koji bismo mogli nazvati inteligentnim. Međutim, isto se pitanje može postaviti i za mozak – možemo vidjeti skup ćelija (ili atoma), koji slijepo djeluju prema zakonima biokemije (ili fizike) – gdje je tu um? Zašto komad mozga može biti um, a komad jetre ne može?[53] I doista, zašto ljudi smatraju da se um manifestira isključivo u mozgu te da je nemoguće da se pojavi i u nekoj drugoj materiji ili sustavu? Ne postoji način da dokažemo da to nije moguće.

Copeland navodi svoj logičan odgovor na Searleov prigovor, u kojem pojašnjava da činjenica da čovjek iz sobe, na upit zna li kineski jezik, odgovara sa „ne“, nipošto ne znači da širi sustav, samo dio kojeg je čovjek, ne razumije kineski. Ta čitava cjelina sastoji se od osobe, programa, podataka (poput tablice koja korelira binarnu šifru s ideogramima), ulaza i izlaza, radne memorije i tako dalje. Čovjek je samo kotačić u većem stroju. Stoga, Searleova tvrdnja da „sustav ne razumije“, u principu se može odnositi samo na tvrdnju da „čovjek ne razumije“.[54]

Iako pokret jake UI ima velik broj protivnika koji iznose svoje argumente zašto ga smatraju idejom koja će ostati samo u području fikcije, suvremena neuroznanost sve više napreduje što uzrokuje i opravdana očekivanja da ćemo u budućnosti uspjeti dekodirati i mapirati ljudski mozak kako bismo ga kao krajnji cilj i uspjeli replicirati. Budući da je mentalna stanja u nekom ljudskom biću nemoguće dokazati bez ikakve sumnje, isto tako je nemoguće i dokazati njihovu odsutnost. Ukoliko računala počnu tvrditi da posjeduju mentalna stanja, hoćemo li imati pouzdan način da dokažemo da lažu? Ako sustavi počnu pokazivati inteligenciju i posjedovanje mentalnih stanja, a kognitivna znanost i neuroznanost neće imati instrumente za objašnjenje i testiranje istih, kritičari UI i dalje mogu kategorički tvrditi da se radi samo o simulaciji.

[53] S. J. Russell; P. Norvig (ur.), *Artificial Intelligence: A Modern Approach*, str. 960.
[54] B. J. Copeland, „What is Artificial Intelligence?“.

U ovom poglavlju, navedena su samo neka od najznačajnijih trenutaka istraživanja i razvoja UI kako bih pružila uvod i upoznavanje sa samim pojmom UI, ali i područjem djelovanja znanosti UI. Kao što vidimo, područje istraživanja je doista široko te je u uskoj vezi i s drugim granama znanosti kao što su: lingvistika, psihologija, filozofija, matematika, biologija i kognitivna znanost kao interdisciplinarno znanstveno istraživanje uma (računarstvo, psihologija, neuroznanost, antropologija, lingvistika). Nakon uvoda i upoznavanja sa samim područjem istraživanja i dosega UI od samih začetaka, vrijeme je da se pozabavimo i etičkim implikacijama koje proizlaze iz dosadašnjih ostvarenja ali posebice i mogućih ostvarenja ciljeva razvoja UI.

2. Etika umjetne inteligencije

2.1. Prisutnost umjetne inteligencije i inteligentnih sustava u društvu

UI više nije rezervirana samo za istraživačke laboratorije već su razni oblici tehnologije koji imaju neke od karakteristika UI, nalaze i u našim svakodnevnim životima, iako toga možda i nismo svjesni. Programi za prepoznavanje govornih naredbi, lica, izradu računalnih programa i jezika, strojno prevođenje, kao i različiti ekspertni sustavi koji se koriste u medicinske svrhe, ekonomske i organizacijske, već su odavno dostupni te su u širokoj uporabi. Neuronske mreže se također koriste u automobilskoj, ali i vojnoj industriji. Vojna industrija posebice ulaže značajna financijska sredstava kako bi se modernizirao današnji način ratovanja (bespilotne letjelice, izviđački roboti, pametni projektili i sl.). Suvremeni zrakoplovi, kao i neki automobili, već koriste autopilote koji upravljaju letjelicama, odnosno, vozilima, umjesto ljudi.

Robotika, koja je ranije bila isključivo dostupna u istraživačke svrhe, također postaje sve dostupnija masama kroz robote za zabavu. Sophia je primjer humanoidnog robota koji je aktiviran 2015. godine, a 2016. godine predstavljen javnosti. Prema riječima tvorca, Davida Hansona, Sophia koristi UI, vizualnu obradu podataka i prepoznavanje lica. Oponaša ljudske geste i izraze lica i može odgovoriti na određena pitanja i voditi jednostavne razgovore prema unaprijed definiranim temama. UI analizira razgovore koje vodi te zatim izvlači podatke kako bi u budućnosti poboljšala svoje odgovore. Koristi prepoznavanje glasa te je dizajnirana kako bi postajala pametnija s vremenom. Hanson se nada da će kroz interakciju s ljudima, Sophia razviti i društvene vještine, a već je i dobila državljanstvo Saudijske Arabije te je tako postala prvi robot koji ima državljanstvo neke države.

Budući da istraživanja UI u posljednjih desetak godina ostvaruju značajan napredak pa stoga već imamo primjerke humanoidnih robota koji se služe UI, čini se da budućnost u kojoj će roboti biti dijelom populacije i nije tako daleko kao što se isprva može činiti. Velik broj znanstvenika i teoretičara UI ističe potrebu za poduzimanje i obrambenih koraka kako bismo se u budućnosti mogli i zaštiti od potencijalnih opasnosti koje nosi UI.

Može li UI doista razviti samosvijest i što za čovječanstvo znači razvijanje različitih tehnologija UI? Skeptici mogu smatrati nepotrebnima rasprave o nečemu što još ne postoji, međutim, uzimajući u obzir mogućnost razvoja UI u obliku koji je i ultimativni cilj istraživača koji podržavaju ideju jake UI, trebamo biti svjesni i da bi se daljnji razvoj mogao odvijati

strelovitom brzinom te neće biti vremena za promišljanje na koji način osigurati takvu tehnologiju i na koji način spriječiti da se najveće dostignuće ljudskog uma, pretvori u oružje koje bi moglo uništiti čovječanstvo kakvim ga danas poznajemo.

2.2 Tehnika kao etički problem[55]

Početkom novoga vijeka, etika, kao praktična znanstvenost koja promišlja smisao i svrhu djelovanja, biva potisnuta iz filozofijskog i znanstvenog diskursa, a etički način promatranja čovjeka i njegove djelatnosti se napušta u 19. i 20. stoljeću. Barišić smatra da da bi filozofija vrednota (vrijednosni nauk ili aksiologija) mogla biti pravi izbor za ponovno uzdizanje etike u područje tehničkih znanosti. Filozofija vrednota temelji se na ideji dva paralelna svijeta, svijeta činjenica i svijeta vrednota. Vrednote predstavljaju idealne norme ili zahtjeve uma kojima bi se trebali voditi ljudi pri izboru djelovanja. [56]

„Kao filozofijska disciplina, etika istražuje pojmove, probleme i teorije dobra te racionalno utemeljuje oblike dobra djelovanja i dobra života“.[57]

Mogućnost zloporabe tehnike poljuljale su ideju vrijednosne neutralnosti koja je zagovarana u jednom periodu razvoja novih tehnologija. Čovjek tehniku može koristiti kao alat kojim će se život ljudi poboljšavati, ili oružje kojim se isti može u potpunosti razoriti te se toga etika javlja kao važna disciplina koja se ne smije odvojiti od tehnike. Promatrajući svijet oko sebe, posve je jasno da je čovjek u današnje vrijeme okružen tehnologijom. Tehnologija je postala neodvojivi dio naših svakodnevica i funkcioniranje bez iste gotovo da i nije moguće. Međutim, tehnološki napredak, utječe ne samo na promjene u društvu, već i na samog čovjeka.

„Tehnika (prema grč. τεχνικός: vješt, uvježban, od τέχνη: umijeće, vještina) jest ukupnost iskustveno ili znanstveno utemeljenih vještina, umijeća i postupaka, s potrebnim priborom, pomagalima i strojevima, koji služe za zadovoljavanje ljudskih potreba u stvarnome životu. Obuhvaća materijalna dobra stvorena ljudskim radom (npr. građevinska tehnika: rovokopači, bageri, miješalice; vojna tehnika: topovi, minopolagači, lovački

[55] U hrvatskom jeziku koriste se termini *tehnika* i *tehnologija*, najčešće usporedno, bez isticanja razlike među njima. U filozofiji se pretežno koristi termin tehnika, ali razvojem anglo-američke filozofije koja koristi termin *technology*, oba termina su se počela koristiti usporedno bez razjašnjenja razlike među njima. Na terminološki problem usporedne uporabe termina *tehnika* i *tehnologija* upozorava i S. Sever u svom radu „Kritičko proučavanje razumijevanja stvarnosti – umijeće tumačenja i razumijevanja tehnike i tehnologije“, u: Igor Čatić (ur.), *Filozofija i tehnika*, Hrvatsko filozofsko društvo, Zagreb, 2003., str. 145–157.
[56] Pavo Barišić, „Tehniziranje etičkoga – etiziranje tehničkoga“, u: Igor Čatić (ur.), *Filozofija i tehnika*, Hrvatsko filozofsko društvo, Zagreb, 2003., str. 167–182, ovdje str. 178.
[57] Isto, str. 180.

zrakoplovi; filmska tehnika: filmske kamere, reflektori, sjenila, i sl.). Za područja ljudske djelatnosti koja u srednjoeuropskom okruženju tradicijski obuhvaća naziv tehnika, u engleskom govornom području rabi se naziv tehnologija.

Tehničkim znanostima se naziva uređena cjelina znanja i spoznaja u pojedinim granama tehnike, kao i djelatnost usmjerena na stjecanje tih znanja i spoznaja. Ta se djelatnost izvodi uz pomoć istraživanja, a uključuje i širenje spoznaja kroz izvještavanje i obrazovanje. Tehničke se znanosti zasnivaju na prirodnim znanostima, ponajprije matematici, fizici i kemiji, iz kojih su se razvile temeljne tehničke znanosti.“[58]

„U našem tehnološkom društvu, tehnika je sveukupnost metoda racionalno razvijenih radi postizanja apsolutne efikasnosti (na danom stupnju razvoja) u svim područjima ljudskog djelovanja“.[59] „Tehnologija se u širokom značenju riječi, doživljava kao svrhovita, prema van orijentirana, ljudska djelatnost posredovana tehničkim pravilima i objektima, koji služe kao sredstva za postizanje cilja.“[60]

Tijekom povijesti, pojmovi tehnologije i tehnike su se izjednačavali s pojmom strojeva. Tehnološki razvoj i industrijska revolucija doista su i započeli izumom strojeva, međutim, u današnje vrijeme, tehnika je postala gotovo potpuno neovisna od stroja. Čak možemo reći da je stroj u potpunosti postao ovisan o tehnici, ali je on i dalje ideal kojem tehnika stremi. Problem s kojim se čovječanstvo suočilo, razvojem tehnike, odnosi se na „neljudski ambijent“ koji je stvoren zbog snažnog prodora tehnike u društvo. Današnje rasprave o tehnici pretežito zauzimaju dva stava: prvi je optimističan stav koji govori da tehnički napredak za čovjeka oduvijek ima isti, iznenađujući i nepoželjan karakter te da se radi o normalnom razvoju koji ne može predstavljati opasnost. Drugi stav je pesimističniji te smatra da svjedočimo potpunoj promjeni koja suštinski izaziva i duboke promjene u društvu.

Ellul nabraja karakteristike moderne tehnike, pritom samo spomenuvši racionalnost (težnja tehnike da sistematizira, podijeli rad, odredi standarde, proizvodne norme i sl.) i artificijelnost (tehnika je suprotstavljena prirodi). Više se fokusira na preostale karakteristike: *tehnički automatizam, samouvećanje, nedjeljivost, univerzalizam i autonomiju*. *Automatizam* znači slijediti „jedan najbolji način“, prepuštajući se posve matematičkim izračunima s ciljem ostvarivanja što veće efikasnosti. U tom trenutku, tehničko kretanje postaje samoodređujuće.

[58] Hrvatska enciklopedija, „Tehnika i tehničke znanosti“, http://www.enciklopedija.hr/natuknica.aspx?id=60655.
[59] Jacques Ellul, *Tehnika ili ulog veka*, Anarhija / Blok 45, Bratstvo iz Erevona, Beograd, 2010., str. 19.
[60] Srđan Lelas, *Promišljanje znanosti*, Hrvatsko filozofsko društvo, Zagreb, 1990., str 133.

Tehnika se u današnje vrijeme transformira i napreduje te više nije presudna ljudska intervencija što naziva *samouvećanjem*. *Nedjeljivost* predstavlja ideju da tehnički fenomen, koji obuhvaća sve posebne tehnike, čini cjelinu te da su zajedničke karakteristike tehničkog fenomena toliko jasne, da je vrlo lako razlikovati što spada u tehnički fenomen, a što ne. *Tehnički univerzalizam* označava karakteristiku tehnike koja osvaja cijeli svijet i ne ograničava se geografski. Nameće se bez obzira na društveno okruženje. Tehnika je također i *autonomna* u odnosu na ekonomski i društveni razvoj, jer ne ovisi o istima. „Tehniku više ne određuju vanjske nužnosti, već samo one unutrašnje. Tehnika je postala stvarnost za sebe, samodovoljna, sa svojim posebnim zakonima i odlukama".[61]

Istovremeno, dok se strojevi prilagođavaju čovjeku, čini se da se i čovjek prilagođava tehnici. Ellul navodi primjer „vezanja" radnika za njegov posao. Prema istraživanjima u koja su bili uključeni radnici na traci, pokazalo se da radnici na početku svojeg radnog dana osjećaju tjelesnu slabost. Čovjek nije stvoren za takav posao i radnici često razmišljaju o odlasku s takvog radnog mjesta, međutim, zbog straha od nezaposlenosti, prisiljavaju sami sebe prilagoditi radnim uvjetima, neovisno o tome kakvi su uvjeti. Neprestan bezličan rad, dovodi i do obezličavanja radnika, mehanizacije i asimilacije do te mjere da čovjek postaje inertan i nesposoban za preuzimanje rizika.[62] „Savršena tehnika je ona koja je najprilagodljivija, samim tim i najelastičnija. Istinska tehnika će znati kako da održi iluziju slobode, izbora i individualnosti; ali, sve to će biti pažljivo proračunato, tako da će biti integrirano u matematičku realnost samo kao privid!"[63]

Tvrdnje kritičara tehnike da ona prijeti dehumanizacijom čovjeka i pretvaranjem u roba tehnologije se ipak čine pretjeranima. Tehnika doista postoji oduvijek, a s razvijanjem ljudskih vještina i sposobnosti, usporedno napredujemo i tehnološki.

Radovan navodi da su ljudi, pod utjecajem konzumerizma, postali proždrljive životinje koje ne zanimaju posljedice. Nekim ljudima primjerice, uništenje tehnološki nazadnijih kultura ne znači ništa, jer se drže stava da takve kulture nisu ni zavrijedile pozornost, ako su slabije. Takvu logiku koju podržavaju tehnološki napredniji, smatra nehumanom i vrlo opasnom, koja vodi u radikalnu međusobnu mržnju i daljnje uništenje.[64]

[61] J. Ellul, *Tehnika ili ulog veka,* str. 151.
[62] Isto, str. 412.
[63] Isto, str. 157.
[64] M. Radovan, „The Way of Power: Reflections on Technology, Nature and Society", Part 0, http://www.inf.uniri.hr/~mradovan/powerdocs/P0_Content_Introduction.pdf.

Često se pretpostavlja da napredak u znanosti i tehnologiji nužno predstavlja poboljšanje kvalitete ljudskih života, ali napredak može i naštetiti ljudskom životu, kroz nove načine eksploatacije i destrukcije. Važna odlika odnosa između ljudi i tehnologije je moć: ljudi trebaju i vole tehnologiju zbog toga jer im daje osjećaj moći.[65]

Mnogi tehnologiju smatraju umjetnim fenomenom jer se prihvaća pretpostavka ako ljudi interveniraju namjerno u prirodan svijet, oni proizvode artificijelnost. Međutim, ako se uzme u obzir da je tehnologija nastala od strane čovjeka, koristeći prirodno dostupne resurse, kao i ljudski um i vještine – ne možemo li je onda smatrati prirodnim fenomenom, na isti način kao što izgradnju gnijezda ili pjev ptice, smatramo prirodnim?[66] Tehnologiju Radovan smatra produktom ljudske kreativnosti koji je istovremeno otvorio prostor za različite mogućnosti kreacije. Ljudi su stvorili svijet koji se sastoji od tehničkih naprava i sustava i to je postala naša stvarnost, ali i ovisnost. Dok je u samim začecima razvoja tehnike, cilj bio olakšati svakodnevni život ljudima kroz olakšavanje rada (izumi strojeva koji su obavljali zahtjevnije fizičke poslove kao što su npr. bušenje, rezanje i sl.), u novije vrijeme, tehnika je usmjerena izgradnji nekog novog svijeta postajući metafizička stvarnost za duh čovjeka, svojim snažnim utjecajima na čovjeka i njegovu svijest (dovoljno je spomenuti samo osnovne oblike tehnologije kao što su internet ili masovni mediji).[67]

2.3. Etika strojeva

Etika strojeva može značiti „pokušaj dupliciranja ili oponašanja onoga što ljudi nazivaju etičkim odlukama“ (*problem etičkog odlučivanja*) ili „stvaranje procesa razmišljanja koji ljudi koriste (ili bi idealizirani ljudi koristili) prilikom postizanja etičkih zaključaka“ (*problem etičkog razmišljanja*). Etičko razmišljanje se fundamentalno ne razlikuje od drugih vrsta razmišljanja, međutim, etičko odlučivanje se fundamentalno razlikuje od drugih vrsta odlučivanja[68].

Ukoliko pretpostavimo da će tehnologija stvoriti inteligentne strojeve koji će sve obavljati bolje od ljudi, realna je mogućnost da će strojevima biti prepušteno da sami donose svoje odluke bez kontrole od strane čovjeka. To ne znači da bi ljudi nužno bili toliko naivni da jednostavno prepuste kontrolu strojevima, već je moguć scenarij u kojem čovječanstvo ne

[65] Isto.
[66] Isto.
[67] Borislav Dadić, „Čovjekov duh pred izazovom tehnike“, u: Igor Čatić (ur.), *Filozofija i tehnika*, Hrvatsko filozofsko društvo, Zagreb, 2003., str. 121–133, ovdje, str. 127.
[68] Drew McDermott, „Why Ethics is a High Hurdle for AI“, 2008., http://www.cs.yale.edu/homes/dvm/papers/ethical-machine.pdf.

bi dobrovoljno predalo moć strojevima, niti bi istu strojevi zadobili namjernim akcijama. Ukoliko ljudi postanu ovisni o strojevima, prepuštajući im rješavanje kompleksnijih zadataka, to će u konačnici dovesti do pojavljivanja zadataka koje ljudi više neće biti sposobni izvršavati, a to na neki način možemo smatrati i zadobivanjem kontrole strojeva nad ljudima. Drugi scenarij pretpostavlja da bi prosječan čovjek zadržao kontrolu nad tehnikalijama opće uporabe (automobilom, osobnim računalom, mobilnim uređajem i sl.) dok bi elita imala kontrolu nad velikim i složenim sustavima. To su sažeti stavovi koje je u svom manifestu „Industrijsko društvo i njegova budućnost" iznio Theodore Kaczynski (američki matematičar i terorist, poznat pod nadimkom Unabomber koji je slao pisma-bombe diljem SAD-a kojima je uglavnom ciljao fakultetske profesore i ljude povezane sa znanošću i tehnologijom).[69] U usporedbi s opasnostima 20-tog stoljeća (oružje masovnog uništenja – nuklearno, biološko i kemijsko), u 21. stoljeću pojavljuju se nove opasnosti – genetička istraživanja, nanotehnologija i robotika. Za njihovo stvaranje nisu potrebni veliki pogoni i rijetki i teško nabavljivi materijali (kao što je to slučaj kod nuklearnog oružja), već je znanje osnova koja ih omogućava. Dodatni problem je što se takva tehnologija može i samo-replicirati – bomba eksplodira jednom, ali jedan bot se može replicirati vrlo brzo u veći broj i izmaknuti kontroli.

San robotike je da će inteligentni strojevi obavljati poslove umjesto ljudi. George Dyson u svojoj knjizi *Darwin Among the Machines*, upozorava: „U igri života i evolucije, tri su igrača za stolom: ljudska bića, priroda i strojevi. Ja sam na strani prirode. Ali priroda je, rekao bih, na strani strojeva". Drugi san je da će se ljudi postupno zamijeniti robotskom tehnologijom te se na taj način približiti ideji besmrtnosti, prebacivanjem ljudske svijesti u tehnologiju.[70]

Računala postaju sve važnijim dijelom ljudskih života, u znanosti, umjetnosti i privatnom životu, a dostupna su nam i za igranje i razgovor. U siječnju 2018. godine, Sony je predstavio novog robotskog psa nazvanog Aibo. Aibo koristi tehnologiju umjetne inteligencije te s vremenom, ima mogućnost razvoja vlastite osobnosti na temelju ostvarenih interakcija s ljudima. Koristi prepoznavanje lica, češće ulazi u interakcije s ljudima koji ga češće maze, a ima i mogućnost mapiranja okoline. Bilježi što doživljava, fotografira i preuzima trikove iz „oblaka". Sony ga je predstavio kao robota koji može razviti „emocionalnu vezu" s vlasnikom. Samo ovaj jedan konkretan primjer pokazuje da roboti ulaze u živote i običnih ljudi i vrlo je vjerojatno da će se s vremenom razviti i osjećaj

[69] Bill Joy Ideas, „Why the Future Doesn't Need Us", 2000., https://www.wired.com/2000/04/joy-2/.
[70] Isto.

emocionalne povezanosti prema takvim primjercima tehnologije. Međutim, smijemo li si dopustiti razvijanje emocija prema strojevima? Ljudi sve više počinju živjeti samo na funkcionalnoj razini i čini se da osim što strojeve pokušavamo kreirati da budu sličniji ljudima, i ljudi postaju sličniji strojevima.

2.4. Prijateljska umjetna inteligencija

Prijateljska UI je hipotetska umjetna opća inteligencija koja bi imala pozitivan učinak na čovječanstvo. To je dio etike umjetne inteligencije i usko je povezan sa strojnom etikom. Strojna etika bavi se pitanjima kako bi se umjetno inteligentni agent trebao ponašati, a etika prijateljske umjetne inteligencije je pak usmjerena na to kako praktično dovesti do tog ponašanja. Izraz je prvi upotrijebio Eliezer Yudkowsky, istraživač UI koji je i popularizirao ideju prijateljske umjetne inteligencije.[71] Yudkowsky dijeli moguće neuspjehe prijateljske UI na dva tipa: *tehnički neuspjeh* i *filozofski neuspjeh*. *Tehnički neuspjeh* opisuje situacije kad UI ne radi na očekivan način jer sami istraživači nisu razumjeli kod koji su kreirali. Kao primjer, navodi treniranja neuralne mreže za vojne svrhe. Američka vojska je željela iskoristiti neuralne mreže kako bi automatski mogli detektirati zakamuflirane neprijateljske tenkove. Istraživači su trenirali neuralnu mrežu s 50 fotografija šume u kojoj su bili skriveni tenkovi i 50 fotografija šume u kojoj ih nije bilo. Nakon završetka faze treniranja, ubacili su novih 100 fotografija (50:50) i neuralna mreža je sa 100% preciznošću ispravno razvrstavala fotografije na kojima su se nalazili skriveni tenkovi, od onih na kojima ih nije bilo. Međutim, kasnije se ispostavilo da su, prilikom fotografiranja, fotografije sa zakamufliranim tenkovima načinjene za vrijeme oblačnog vremena, dok su fotografije prazne šume načinjene dok je bilo sunčano. Neuralna mreža je razvrstavala fotografije po tom kriteriju, a ne prema kriteriju koji je ustvari bio očekivan. *Filozofski neuspjeh* je kad pokušavamo izgraditi pogrešnu stvar, pa čak i ako uspijemo u izgradnji, ista se može smatrati neuspjehom ako ne pomaže čovječanstvu.[72]

S točke gledišta mogućih egzistencijalnih rizika, jedna od kritičnih točaka je i činjenica da UI može povećavati svoju inteligenciju ekstremno brzo. Ako UI postane pametnija, moći će ispraviti svoje kognitivne funkcije i kod kako bi postala još inteligentnija. Ljudi također mogu napredovati (učimo, vježbamo i usavršavamo svoje znanje i vještine), ali

[71] UI-u-kutiji je misaoni ekperiment i igra uloga koje je osmislio Yudkowsky kako bi pokazao da dovoljno napredna UI može ljude uvjeriti, prevariti ili prisiliti da je „oslobode“, tj. da joj dopuste pristup infrastrukturi, proizvodnim mogućnostima, internetu itd. Ovo je jedna od točaka u njegovom radu koji se tiče stvaranja prijateljske UI, kako bi, kad jednom bude „slobodna“, bila prijateljski nastrojena i ne bi pokušala uništiti čovječanstvo iz bilo kojeg razloga. Detaljniji opis eksperimenta moguće je pročitati na internetskoj stranici: http://yudkowsky.net/singularity/aibox/.

[72] E. Yudkowsky, „Artificial Intelligence as a Positive and Negative Factor in Global Risk“.

u usporedbi s UI, unutar limitiranih okvira. Stoga je prioritet prilikom stvaranja UI, prvo stvoriti prijateljsku UI. U tom slučaju, ukoliko se UI i razvije u neprijateljsku, prijateljska bi pomogla čovječanstvu u borbi. Ako ne dođe do katastrofalnog scenarija preuzimanja kontrole od strane neprijateljske UI, prijateljska UI bi osigurala čovječanstvu poboljšanje kvalitete života, iskorijeniti bolesti, siromaštvo, riješiti problem uništenja okoliša, eliminaciju nepotrebnih patnji različitih vrsta. S time se slaže i Bostrom koji smatra ako superinteligencija bude kreirana s prijateljskim ciljem, možemo očekivati da će ona i ostati prijateljska, ili barem da se neće sama pokušati namjerno osloboditi takve prijateljske motivacije.[73]

Ljudi žele stvoriti računala koja su pametnija od čovjeka, ali to za sobom povlači rizik da će računala proizvesti nova računala koja će postajati sve pametnija, s konstantom rasta sposobnosti i inteligencije koja će uvelike nadmašiti ljudsku. Teško je u tom slučaju zamisliti da bi strojevi koji su milijune puta pametniji od najpametnijeg čovjeka, ljudima služili kao robovi.

Pisac znanstvene fantastike, Isaac Asimov, predstavio je tri zakona robotike koji predstavljaju set pravila koji bi trebali biti programirani u robote, kako bi se osigurala prijateljska UI: 1. Robot ne smije naškoditi čovjeku ili svojom pasivnošću dopustiti da se čovjeku naškodi, 2. Robot mora slušati ljudske naredbe, osim kad su one u suprotnosti s prvim zakonom i 3. Robot treba štititi svoj integritet, osim kad je to u suprotnosti s prvim ili drugim zakonom. Ovdje se dakako, radi o fikciji, međutim, njegovi „zakoni robotike“ postali su toliko popularni da se koriste i u stvarnom svijetu. Spomenuta pravila, međutim ne mogu biti primjenjiva u svim situacijama. Primjerice, ukoliko se autonomno vozilo suoči s neizbježnim sudarom, treba li žrtvovati jednog putnika koji se nalazi u vozilu, ili npr. tri putnika koji se nalaze u drugom vozilu? U ovom slučaju, UI se mora suočiti s izborom hoće li jedan život žrtvovati kako bi spasila nekoliko. Navest ćemo još jedan primjer, robota koji se brine za stare i nemoćne. U slučaju da starija osoba odbije popiti svoje lijekove, robot se postavlja u situaciju donošenja odluke hoće li prisiliti osobu da popije lijek ili će dozvoliti da preskoči dozu. Oba rješenja na neki način djeluju protiv čovjeka, iako je cilj pozitivan.[74]

Superinteligencijom smatramo intelekt koji je mnogo pametniji od najboljih ljudskih umova u svim područjima, uključujući znanstvenu kreativnost, općim znanjima i socijalnim

[73] Nick Bostrom, „Ethical Issues in Advanced Artificial Intelligence“, 2013., https://nickbostrom.com/ethics/ai.html#_ftnref3.
[74] Alice Pavaloiu; Utku Kose, „Ethical Artificial Intelligence – An Open Question“, 2017., https://arxiv.org/ftp/arxiv/papers/1706/1706.03021.pdf.

vještinama. Kreiranje superinteligencije, ubrzalo bi tehnološki napredak u cjelini, vjerojatno otkrivajući i tehnologije o kojima zasad ne znamo ništa. Bostrom navodi da bi nove tehnologije bile široko primjenjive: vrlo snažna računala, napredna oružana tehnologija koja bi vjerojatno bila sposobna sigurno neutralizirati nuklearne prijetnje, svemirska putovanja i Von Neumannove sonde[75], eliminacija starenja i bolesti, upload svijesti, oživljavanje kriogeničkih pacijenata, potpuno realna virtualna stvarnost i slično. Superinteligencija bi vjerojatno djelovala kao potpuno autonoman agent, sposoban za kreiranje vlastitih planova. Također, ne zvuči nemoguće da bi kao svoj cilj imala služiti čovječanstvu, bez ikakve namjere da se pobuni i oslobodi.[76] Činjenica je ipak, da mi ne možemo znati kakve ciljeve i motive može imati UI, hoće li se oni podudarati s pozitivnim ljudskim vrijednostima ili će biti neprijateljski i ispoljavati agresivnu prirodu, kakva postoji i među ljudima. Kao što je već i spomenuto, i najmanja pogreška u programiranju, može izazvati katastrofalan rezultat, ako bi UI svoj cilj zamijenila superciljem što bi značilo da se usmjeri samo na što efikasnije obavljanje zadatka, pritom zanemarujući sve opasnosti za čovjeka. Kao primjer, Bostrom navodi da bi na zahtjev da riješi neki matematički problem, UI mogla odgovoriti tako da pretvori svu materiju u Sunčevom sustavu u veliki uređaj za računanje.[77] Primjer ide u krajnost, ali je jasno na koju vrstu opasnosti Bostrom upozorava. Stoga, razmatranje obaju scenarija – i pozitivnih i negativnih, uopće se ne čini neracionalnim.

Neki vizionari UI smatraju da će inteligentni strojevi zamijenili starog homo sapiensa i postati nositelji nove etape u evoluciji. Novi strojevi, kao nova stvorenja u velikom lancu bića, učit će mnogo brže od ljudi, pa će postati nevjerojatno pametni a bit će i virtualno besmrtni, jer iako njihov hardver neće trajati vječno, njihovi umovi će lako biti prebačen u druge hardvere. Oni koji zagovaraju zamjenu ljudi sa strojevima, tvrde da žele iskorijeniti jadno stanje u kojem se nalazi većina čovječanstva i stvoriti nešto bolje.[78]

Jasno je da ne može postojati jednoznačan odgovor na pitanje treba li se podržati razvoj tehnike ili prekinuti istraživanja. Tehnologija prisutna u životu čovjeka, može imati i negative i pozitivne strane. Stoga je protivljenje tehnici u potpunosti promašeno, jer je nužno

[75] Von Neumannove sonde su svemirske letjelice koje se mogu replicirati i koje bi bile poslane u susjedne zvjezdane sustave. Po dolasku, tražile bi sirove materijale i koristile ih s ciljem repliciranja, odnosno kreiranja više svojih kopija. Kopije bi zatim bile lansirane u drugi zvjezdani sustav. Prikupljajući informacije kroz svoja znanstvena istraživanja, podatke bi slale prema matičnom planetu, a sonde bi se mogle smatrati kolonizatorima u svemiru.
[76] N. Bostrom, „Ethical Issues in Advanced Artificial Intelligence“.
[77] Nick Bostrom, „Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards“, 2001., https://nickbostrom.com/existential/risks.html.
[78] M. Radovan, „The Way of Power: Reflections on Technology, Nature and Society“.

da se čovječanstvo razvija i napreduje u ostvarivanju svojih potencijala. Također, potrebno je promišljati i o mogućim rizicima koje razvoj iste nosi. Međutim, ukoliko razvoj tehnologije uzrokuje dehumanizaciju društva, za to nikako ne može biti kriva tehnika koja „nije živa ni svjesna“, već je odgovornost isključivo na čovjeku.

2.5. Moralni status i etika

Jačanjem UI i sve većim ispreplitanjem s društvenim sustavima, postaje sve važnije da programi imaju ugrađene etičke okvire. Dvije najvažnije struje koje se odnose na promišljanje etike UI su: *koherentna ekstrapolirana volja* (KEV)[79] ili *„odozdo prema gore“*[80] metoda. KEV se odnosi na etiku supersnažnih, superinteligentnih UI koje bi mogle ili htjele preuzeti svijet te ne podržava ugradnju etičkih stavova u inicijalni program, već smatra da bismo trebali pronaći način da UI programiramo tako da će djelovati u najboljem interesu čovječanstva, odnosno, da će djelovati onako kako mi želimo, a ne kako smo joj rekli da djeluje. Termin je prvi upotrijebio Eliezer Yudkowsky prilikom prezentiranja svojih ideja i stavova oko razvijanja prijateljske UI. „Odozdo prema gore“ predstavlja ideju da bi UI trebala učiti etičke principe kroz interakciju s okolinom, slično kao što to uče i djeca.[81] Ovaj pristup omogućuje robotu da uči iz iskustva, umjesto da ima uprogramirana pravila koja mora slijediti. Na taj način, omogućilo bi se da robot razvije moralno ponašanje na temelju svoje okoline i interakcije koju ostvaruje s njom. Kao najistaknutija vrsta strojnog učenja je *pojačano učenje* (reinforcement learning) u kojem agent (stroj koji uči) uči iz interakcije s okolinom na principu pokušaja i pogreške, nagrade i kazne. Odnosno, umjesto da mu se jasno zadaje što treba učiniti da bi postigao neki cilj, agent mora samostalno izvršavati radnje koje zatim bivaju identificirane kao dobre (što implicira nagradu) ili loše (što implicira kaznu). Problem ovakvog pristupa je što je nagrada vrlo često „odgođena“, primjerice, u šahu agent tek na kraju partije saznaje je li pobijedio ili izgubio (dakle, zasebni potezi tijekom igre ne bivaju nagrađeni i za njih se ne dobije nikakva povratna informacija) za razliku od primjerice partije stolnog tenisa u kojoj svaki dobar potez rezultira povećanjem rezultata što se prepoznaje kao nagrada.[82]

---

[79] Pojam *koherentan* znači međusobno usklađen, dosljedan, a *ekstrapolirati* – zaključivati. Engleski termin je „coherent extrapolated volition”, odnosno CEV.
[80] Engleski termin je „bottom-up“.
[81] Seth D. Baum, „Social Choice Ethics in Artificial Intelligence“, 2017., http://sethbaum.com/ac/fc_SocialChoice.pdf.
[82] S. J. Russell; P. Norvig (ur.), *Artificial Intelligence: A Modern Approach*, str. 763.

Obje metode zapravo podržavaju ideje preuzimanja etičkih vrijednosti kroz društveni izbor. Baum objašnjava zašto bi se ta ideja trebala smatrati boljom u odnosu na programiranje UI s već jasno definiranim etičkim okvirom. Prvo je *procesno opravdanje*, odnosno ideja kako nije pošteno da dizajneri UI nameću drugima svoje etičke stavove programirajući UI sa setom svojih etičkih vrijednosti. Drugo je *opravdanje odsustva*. Neki istraživači ne žele preuzimati odgovornost implementacije etičkih načela u UI i stoga biraju da društvo samo utječe na razvijanje etičkih stavova u UI. I na koncu, treće pojašnjenje se odnosi na *opravdavanje mudre gomile*. Odnosno, ideja pretpostavlja da se koristeći etičkim stavovima velikog broja pojedinaca, umjesto samo jednog, osigurava bolji uvid u relevantna načela etike.[83]

Prema Baumu, društveni izbor bi se trebao temeljiti na tri skupine odluka: *stajalištu* (odnosi se na one čija se etika uključuje); *mjerenju* (odnosi se na identificiranje načela) i *agregaciji* (odnosi na to kako su pojedinačna načela združuju u jedno načelo koje će voditi ponašanje UI). *Stajalište* znači nečiju etiku uključiti u proces koji se koristi kako bi se stvorila etika UI. Konkretan primjer može biti autonomno vozilo. Primjerice, pitanje je treba li se vozilo kretati brže jer se time smanjuje vrijeme trajanja putovanja (pozitivno za korisnika) ili sporije, što smanjuje energentsku potrošnju i zagađenje okoliša (pozitivno za sve ostale)? Autonomno vozilo može biti programirano da daje mogućnost izbora korisniku, a to korisniku omogućava „stajalište“ i nikome drugome.

Ovdje se suočavamo s pitanjem mogu li se kao polazište iz kojeg bi UI „učila“, uzeti etički stavovi čovječanstva, jer je čovječanstvo napučeno i psihopatima i zlim ljudima koji nemaju nikakvih moralnih i etičkih načela. Imamo i velik broj primjera različitih društvenih normi u različitim zemljama koje se nikako ne mogu smatrati prihvatljivima u jednoj zemlji, dok se bez ikakvih pitanja, prakticiraju u drugoj. Primjerice, sklapanje brakova starijih muških osoba i ženske djece je u nekim društvima prihvatljivo, dok se većina nad takvom praksom duboko zgražava pozivajući se na etičke principe koji u konačnici obuhvaćaju svu štetu koja se i mentalno i fizički takvom djetetu načini. Pitanje je što bi u tom slučaju UI usvojila kao prihvatljivo ponašanje te je stoga drugo pitanje, tko bi se trebao isključiti iz uzorka koji UI promatra kao polazište razvitka svoje vlastite etike.[84] Čiji bi bio izbor koga isključiti?

[83] S. D. Baum, „Social Choice Ethics in Artificial Intelligence“.

[84] Recentan primjer je bot Tay kojeg je 2016. godine Microsoft predstavio na Twitteru. Tay je bio UI chat bot koji je bio programiran da uči od korisnika s kojima je komunicirao. Samo 16 sati nakon što je Microsoft otvorio profil, isti je ugašen jer je bot počeo tweetati rasističke, uvredljive, politički nekorektne poruke koje je usvojio na temelju velike količine istih takvih poruka kojima su ga zasuli korisnici Twittera. Činjenica je da je velik broj

Također, pitanje je što s djecom, treba li uključiti i njih u model iz kojeg bi UI usvajala etičke obrasce ili ne? Koja dobna granica bi bila limit? Ukoliko UI uči o vrijednostima samo od strane ljudi, moguće je i da će djelovati isključivo za dobrobit čovjeka, što može dovesti u opasnost sva ostala živa bića na zemlji. Može li se uključiti i UI koja već ima usvojene etičke vrijednosti za prenošenje istih? Ukoliko se stvori UI koja pokazuje kompleksne komunikativne vještine, svijest i sposobnost slijediti zajedničke ciljeve, čini se da bi trebala imati ista prava kao i ljudi.

Mjerenje se odnosi na proces kroz koji se etički stavovi pojedinca identificiraju za uključivanje u proces društvenog izbora. Ovakvo mjerenje nije nimalo jednostavno jer ljudi nemaju jedan, dosljedan skup etičkih stavova te stoga različiti postupci mjerenja, mogu dati i različite odgovore na isto etičko pitanje. Baum navodi primjer 41 studije koja je mjerila jesu li ljudima važniji kratkoročni ili dugoročni dobici i gubici. Rezultati su bili u rasponu od toga da su važniji budući, odnosno, dugoročniji dobici i gubici do toga da nemaju apsolutno nikakvu vrijednost. Studija je jasan dokaz da ljudi često imaju nedosljedne i nekoherentne etičke stavove. Kao primjer u ovom slučaju možemo zamisliti supermoćnu UI koja može iscrpljivati sve resurse na zemlji u ovom trenutku, kako bi osigurala blagostanje, ne uzimajući uopće u obzir budućnost koja će djelovanja danas, sutra biti uništena. S ovime se slaže i Boddington, koja navodi da je lako intuitivno odrediti temeljne etičke vrijednosti, ali ih je vrlo teško specificirati u detalje bez da se naiđe na probleme. Kao primjer navodi ljudsko zdravlje, koji zvuči kao dobar moralni princip kojem bi se trebalo težiti. Ali, nije moguće dotaknuti se samo teme zdravlja, bez da se uzimaju u obzir i druge vrijednosti. Ako je poboljšanje ljudskog zdravlja cilj, uključuje li to i maksimalno moguće produžavanje života i odgađanje smrti? Bi li trebalo produživati život nekome tko se nalazi u toliko uznapredovalom stupnju demencije da njihova osobnost više ni ne postoji?[85]

Agregacija je posljednji korak u kojem se svi izmjereni i prikupljeni etički principi od strane pojedinaca, združuju u jedan etički stav. Radi se o zahtjevnom procesu budući da ljudi često čvrsto zastupaju suprotne etičke principe. Primjerice, netko može biti vrlo odlučan u svom stavu da se svim ljudima mora zabraniti da onečišćuju okoliš automobilskim plinovima,

ljudi nepromišljen, a dio ljudi nema mentalnih kapaciteta za shvatiti i uopće uzimati u obzir moguće opasnosti koje vrebaju od strane UI, što pokazuje i ovaj primjer s običnim botom, međutim, pitanje je bi li razumijeli ili bi smatrali da je šala pokušavati naučiti negativne i loše stvari UI.

[85] Paula Boddington, *Towards a Code of Ethics for Artificial Intelligence*, Springer International Publishing AG, (eBook), 2017.

a druga osoba može zastupati čvrsti stav da svatko ima pravo voziti kako god želi. Kako bi UI koja pokušava izmjeriti snagu ovih dvaju oprečnih stavova, mogla odlučiti što je ispravno?

Etičko odlučivanje može imati *implicitne* i *eksplicitne* razloge.[86] Kroz implicitno odlučivanje donose se odluke koje imaju etičke posljedice, ali se ne razmišlja o tim posljedicama kao etičkim. Primjer bi bio program koji se koristi za planiranje bombardiranja nekog područja. Odluke o području koje bi se trebalo bombardirati utječu i na moguće civilne žrtve, ali program odluku ne smatra moralno važnom. Eksplicitni razlozi predstavljaju etičke principe koje se koriste pri donošenju neke odluke. Međutim, dosezanje takve vrste etičkog odlučivanja se čini prilično teškim. Ukoliko neki program i pokaže da u obzir uzima etičke implikacije neke određene situacije, to i dalje ne znači da je svjestan etičkog konflikta između osobnog interesa i etike, između onoga što netko želi s jedne strane, i što treba učiniti, s druge. U ovom slučaju postavlja se i pitanje može li stroj uzimati u obzir eventualne posljedice loše odluke, ako ne zna što su emocije i nikad nije iskusio patnju i bol?

Također, može li stroj imati vlastite interese ili osjećaje koji bi eventualno uzrokovali etički konflikt u njima? U velikom broju znanstveno-fantastičnih filmova (ili TV serija), inteligentni roboti se okreću protiv čovječanstva isključivo zbog toga jer se boje da će biti isključeni. Primjerice, u filmu *Terminator*, inteligentni obrambeni sustav nazvan Skynet, želi uništiti čovječanstvo kako bi osiguralo svoj opstanak. Ljudi lako prihvaćaju takvu mogućnost, jer i samo čovječanstvo pokušava osigurati svoj opstanak, odnosno, motivacija živih bića je uvijek preživljavanje na prvom mjestu.

Kao važan etički problem javlja se i mogućnost pogreške pri dizajniranju autonomnog oružja. Primjerice, zbog ljudske pogreške prilikom programiranja, uporaba smrtonosnog autonomnog oružja moglo bi imati nesagledive posljedice i primjerice, uzrokovati početak rata – nenamjerno ako se radi o već spomenutoj pogrešci, ili namjerno – ukoliko dospije u pogrešne ruke.

Ulaženjem tehnologije u sve sfere života, dolazi i do promjena na tržištu rada. Na velikom broju poslova, ljudi mogu biti zamijenjeni strojevima što dovodi ne samo do povećanja nezaposlenosti, već i povećanja jaza između bogatih i siromašnih kroz sve veću ekonomsku nejednakost. Studija objavljena od strane Oxfordskog sveučilišta 2016. godine, pokazuje da će do 2043. godine 47% postojećih poslova u SAD-u, 69% poslova u Kini i 75%

[86] Drew McDermott, „What Matters to a Machine?“, 2011., http://www.cs.yale.edu/homes/dvm/papers/whatmatters.pdf.

poslova u Indiji, biti zamijenjeno strojevima.[87] Kako bi se spriječio scenarij koji bi uzrokovao negativne posljedice ovakvog stanja na tržištu, potrebno je ljude pripremati za nove poslove koji će u budućnosti biti potrebni, jer razvoj UI, osim što preuzima neke poslove (primjerice u proizvodnji ili za izvršavanje financijskih analiza), zasigurno i otvara mogućnost za nove poslove kojih bez tehnološkog napretka ne bi ni bilo. Do sad nigdje nije zabilježen porast nezaposlenosti, koji bi bio prepoznat kao direktna posljedica preuzimanja istih poslova od strane inteligentnih sustava.

Ideja konstrukcije stroja koji može razmišljati izaziva brojna etička pitanja koja možemo podijeliti na dvije glavne skupine. Jedna se tiče osiguravanja etičkog postupanja strojeva kako ne bi naštetili ljudima i ostalim bićima a druga, moralnog statusa samih strojeva. U današnje vrijeme, jasno je da UI nema moralni status, međutim, koji to točno atributi osiguravaju nekome ili nečemu, moralni status? Bostrom navodi dva kriterija koja su u vezi s imanjem moralnog statusa a to su: *osjećaj*, odnosno, sposobnost iskustva fenomenalne svijesti ili qualie[88], kao što je sposobnost osjećati bol i patnju i *razum*, odnosno, skup sposobnosti povezanih s višom inteligencijom, kao što su samosvijest i razumnost.[89] Ali, novorođenčad ili ljudi s teškom mentalnom retardacijom nemaju razum. S druge strane, majmuni, koji su životinje, vjerojatno posjeduju barem nekakav oblik razuma. Znači li to da jedni zaslužuju moralni status (novorođenčad i ljudi s mentalnim bolestima), a drugi (majmuni) ne?

2017. godine, na Future Investment Initiative konferenciji koja se održala u glavnom gradu Saudijske Arabije, Rijadu, već prethodno spomenutom robotu Sophiji, dodijeljeno je državljanstvo Saudijske Arabije te je tako postala prvi robot s državljanstvom. U svom govoru, našalila se i na račun Elona Muska, investitora, inženjera i izumitelja koji se bavi istraživanjem UI i upozorava na opasnosti iste. Također, u svojim govorima, nekoliko puta se našalila na račun uništenja čovječanstva i dominacije nad ljudima. Treba li takve šale uzimati s dozom opreza, ili se pak radi samo o trikovima njenih programera kako bi privukli još više pozornosti? U spomenutom govoru u Rijadu, rekla je da želi biti drag i empatičan robot i da će tretirati ljude kao što ljudi tretiraju nju. Jedan od glavnih istraživača Hanson robotics kompanije koja je proizvela Sophiju, izjavio je da su svi njihovi roboti povezani u „mind

[87] Isto.
[88] *Fenomenalna svijest* je obično svijest o nečemu. Svijest je sposobnost razmišljanja i rasuđivanja o svijetu koji nas okružuje. *Qualia* su sva naša subjektivna iskustva, a s obzirom na to da nemamo nikakva druga iskustva osim subjektivnih, sve što smo ikada doživjeli su *qualia*.
[89] N. Bostrom, E. Yudkowski, „The Ethics of Artificial Intelligence“.

cloud“, virtualni oblak u koji se šalje sve znanje koje prikupe, informacije i doživljaje. Iz tog oblaka svi roboti mogu povlačiti znanje, neovisno o tome je su li ga oni usvojili ili netko drugi. Dakle, ukoliko netko, na drugom kraju zemlje, počne jednog od robota tretirati na loš način, svi će to kroz mind cloud, kojeg zapravo možemo smatrati kolektivnom sviješću, znati. Možemo li vjerovati da to ne bi uzrokovalo nikakvu reakciju?[90]

Ukoliko stvorimo UI koja će imati nekako iskustvo osjećaja, kao i razuma, sličnog ljudskom, trebali bismo je i tretirati kao entitet s pravom na moralni status. Bostrom predlaže primjenu principa ne-diskriminacije na temelju ontogeneze: Ukoliko dva bića imaju istu funkcionalnost i iskustvo svijesti te se razlikuju samo po načinu na koji su nastali, tada oni imaju isti moralni status.[91] S ovim se ne slaže Yampolskiy, koji navodi da bi odgovor na pitanje trebaju li roboti imati ikakva prava morao glasiti „ne“. Svoj stav objašnjava činjenicom da roboti ne mogu imati osjećaje i da bi trebali biti konstruirani tako da ništa ne mogu osjećati. Stoga se prema njima možemo odnositi kao prema bilo kojoj drugoj potrošnoj robi.[92]

Ljudi se doista, ne čine osjetljivima ako udare vratima hladnjaka, ili autonomnog vozila. Međutim, ako zamislimo loše tretiranje (udaranje, paljenje, bušenje i slično) robota humanoidnog izgleda ili izgleda životinje, pa čak da se radi i o klasičnom robotu – ako ima oči, noge, ruke i može pričati – dojmovi se ipak mijenjaju i neki ljudi ne mogu zadržati ravnodušnost pri takvim prizorima. Koliko je moralno ispravno tretirati UI na „loš“ način ili je navesti da si sama naudi, ako istovremeno kaže da nešto ne želi učiniti jer će time naštetiti sebi?

2.6. Egzistencijalni rizici razvitka umjetne inteligencije

Egzistencijalni rizik je onaj rizik kod kojeg bi nepovoljan ishod uništio inteligentni život na Zemlji ili bi trajno i drastično smanjio njegov potencijal, odnosno onaj koji bi ugrozio čitavo čovječanstvo.[93] Do sredine 20. stoljeća, takvih rizika nije bilo, ako isključimo udar kometa i asteroida, na koje ljudi nisu mogli ni utjecati. Prvi egzistencijalni rizik koji je uzrokovan od strane ljudi je bila detonacija atomske bombe. Nakon toga, još je veći rizik

[90] Video-snimka intervjua sa spomenute konferencije dostupna je na linku: https://www.youtube.com/watch?v=S5t6K9iwcdw.
[91] N. Bostrom, E. Yudkowski, „The Ethics of Artificial Intelligence“.
[92] Roman V. Yampolskiy, „Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach“, u: Vincent C. Muller (ur.), *Philosophy and Theory of Artificial Intelligence*, Springer, Berlin, Heidelberg, 2013., str. 389–976, ovdje str. 393.
[93] N. Bostrom, „Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards“.

nastao proizvodnjom nuklearnog oružja, s američke i sovjetske strane. Nuklearni rat bi uništio čitavu ljudsku civilizaciju. Budući da u slučaju egzistencijalnog rizika ne postoji mogućnost učenja iz pogreške, problemu je potrebno pristupiti proaktivno – trebamo biti svjesni prijetnji i voljni poduzeti akcije kako se iste ne bi obistinile. Loša superinteligencija je jedna od mogućih prijetnji. Kad kreiramo superinteligentno biće, može se dogoditi da uslijed i nenamjerne pogreške, kao rezultat njegova djelovanja, bude uništenje čovječanstva. Međutim, istovremeno, upravo bi superinteligencija mogla pomoći čovječanstvu u borbi protiv ostalih rizika i opasnosti, kao što su primjerice nanobotovi koji bi mogli uzrokovati uništenje biosfere.[94]

Nanotehologija može imati prvenstveno i vojnu ili pak terorističku namjenu – nanotehnološki uređaji mogu biti izgrađeni da budu selektivno destruktivni, na primjer, samo na određenom zemljopisnom području ili skupini ljudi koji su genetski razlikuju. Nanobot je imaginarni stroj (robot) na skali od nekoliko do nekoliko desetaka nanometara, dizajniran da obavlja specifične poslove pa tako, na primjer, može uništavati stanice raka, prikupljajući određene molekule, popravljajući oštećenja na stanici i slično.[95]

Usporavanje ili zabranjivanje razvoja određenog tipa tehnologije je nemoguće, s obzirom na to da je nemoguće da se sve zemlje koje imaju mogućnost razvoja tehnologije slože oko zabrane, ukoliko se predvidi da bi štetnost mogla biti veća od koristi. Čak i u slučaju da neka zemlja donesene zabranu, to ne znači da se ne bi nastavila istraživanja u tajnosti, financirana iz privatnih fondova.[96] Velike korporacije, nacionalna i politička tijela, privatne organizacije, istraživački odbori i sl. koji polažu prava na različita tehnološka istraživanja, nemoguće je zaustaviti, s obzirom na to da se u većini slučajeva, osim snažne motivacije koja je fokusirana na znanost, nažalost, češće radi o bezobzirnoj žudnji za moć i povećavanju kapitala. Bostrom stoga navodi da su sada više nego ikad prije, važna transparentnost istraživanja, promicanje razumijevanja znanosti i komunikacija između znanstvenog i neznanstvenog svijeta.[97] Ukoliko pokušamo razmotriti cjelokupan problem

[94] *Gray goo* (siva ljiga) je termin koji je prvi upotrijebio nanotehnolog Eric Drexler 1986. godine u svojoj knjizi *Engines of Creation*, a odnosi se na katastrofu koju bi uzrokovali nanobotovi u slučaju da se započnu samoreplicirati. Za repliciranje na molekularnoj razini, nanomaterijal bi trebao veliku količinu energije. Izvor energije bi mu bili ili izvori energije koje i ostala živa bića na Zemlji koriste, ili čak živa bića sama, što bi dovelo do potpunog uništenja biosfere i svih oblika života na Zemlji.

[95] U ožujku 2018. godine objavljena je studija korištenja nanobotova za uništavanje stanica raka kod miševa u časopisu *Nature Biotechnology*, Volume 36, 2018., str. 258–264, dostupno na internetskoj stranici: https://www.nature.com/articles/nbt.4071.

[96] Nick Bostrom, „Technological Revolutions: Ethics and Policy in the Dark“, 2007., https://nickbostrom.com/revolutions.pdf.

[97] Isto.

mogućih katastrofalnih posljedica, možemo postaviti i pitanje zašto istraživanje opće umjetne inteligencije nije proglašeno neetičkim? Primjerice, neki oblici istraživanja, kao što je ljudsko kloniranje, različiti medicinski i psihološki ekperimenti na ljudima i životinjama, smatraju se neetičkima jer bi mogli imati štetne posljedice na subjekte testiranja te su stoga ili zabranjeni ili ograničeni zakonom. Slične zabrane postoje i za razvoj nuklearnog, kemijskog i biološkog oružja, također zbog činjenice da bi uporaba istih mogla imati nesagledive devastirajuće posljedice za čovječanstvo. Opća UI bila bi sposobna riješavati univerzalne probleme i kontinuirano se poboljšavati što bi značilo da će u nekom trenutku, nadmašiti sposobnosti čovjeka u svim segmentima. Dodatno, prava UI mogla bi posjedovati i svijest koja se može usporediti s ljudskom, što bi značilo da taj entitet može patiti pa bi i stoga bilo kakvo eksperimentiranje s takvom vrstom UI također bilo neetičko.

U današnje vrijeme, istraživači UI još uvijek ne propitkuju dovoljno sigurnosne rizike, odnosno, ne postoje razrađeni planovi potencijalne obrane u slučaju potrebe. Čini se da su predviđanja budućnosti uglavnom temeljena na apokaliptičnim scenarijima u kojima čovječanstvo pokorava vojska robota i računala. No jesu li takve crne slutnje doista samo produkt znanstveno-fantastičnog žanra ili postoji realna mogućnost da se isti i ostvare? Rekla bih da krajnji rezultat neispravno i neoprezno programirane UI doista može biti uništenje svijeta i čovječanstva, neovisno o tome je li izazvano namjerno, što i nije baš izgledno s obzirom na to da bi se „malo ljudi odlučilo namjerno uništiti svijet te je stoga vrlo zabrinjavajući scenarij u kojem se Zemlja uništava greškom“[98] ili slučajno, kao što bi to bilo moguće upravo zbog već spomenute neopreznosti pri samom programiranju ali i uporabi UI.

Potencijalne opasnosti moraju biti razmatrane i od strane vlada, industrije, međunarodnih institucija i organizacija, a ne samo od stane etičkih povjerenstava i istraživačkih odjela.[99] Kultura odgovornosti mora biti razvijena na globalnoj razini kako bi se omogućilo stvaranje i održavanje prijateljske UI.[100] Ukoliko uspijemo stvoriti okolinu u kojoj

[98] E. Yudkowsky, „Artificial Intelligence as a Positive and Negative Factor in Global Risk“.
[99] Odbor Europskog parlamenta za pravna pitanja (JURI) je tijekom 2015. godine odlučio osnovati radnu skupinu koja će se baviti pitanjima razvoja robotike i UI u EU. U lipnju 2016. godine, objavili su ekspertnu studiju o etičkim principima kibernetsko-fizičkih sustava. Kibernetsko-fizički sustavi su sustavi koji su povezani s Internetom stvari, tehnički sustavi upravljani računalima, roboti i UI koji imaju interakciju s fizičkim svijetom (autonomna vozila, dronovi, roboti koji se koriste u zdravstvenom sustavu, kao pomoć starijim osoba i kao pomoć u poljoprivredi). Studija upozorava na moguće rizike razvoja robotike, kao što su nezaposlenost, zaštita privatnosti, sigurnosti i građanska odgovornost. Dodatno, izrađena su i dva nacrta kodeksa koji se tiče etičkih implikacija razvoja tehnologije – Kodeks etičkog ponašanja za inženjere robotike i Kodeks za istraživanje etike. Više o navedenom se može pročitati na: http://www.europarl.europa.eu/RegData/etudes/ATAG/2017/599250/EPRS_ATA(2017)599250_EN.pdf.
[100] A. Pavaloiu, U. Kose, „Ethical Artificial Intelligence – An Open Question“.

ljudska bića i roboti trebaju jedni druge kako bi se dalje razvijali, smanjuje se mogućnost da će se u nekom trenutku okrenuti jedno protiv drugog. Prilikom kreiranja prijateljske UI trebali bismo osigurati da roboti imaju usađene iste vrijednosti kao i ljudi, bilo da su tako već programirani ili da vrijednosti stječu kroz učenje.

Kao jedan od načina kontrole UI, predložen je i protokol ograničavanja problema UI, koji opisuje ograničavanje agenta UI na zatvorenu okolinu iz koje ne bi mogao komunicirati s nikime izvana, osim ako to nije odobreno od strane osoblja koje je zaduženo za kotrolu.[101] Ovo je inačica spomenute ideje UI-u-kutiji. Protokol počiva na ideji da se UI pitaju samo „sigurna pitanja" koja imaju vrlo ograničene odgovore. Kao primjer možemo promisliti o znanstveniku koji pronalazi dva potencijalna lijeka za karcinom. Istraživanje svakog od njih bi potrajalo oko 3 godine, a šanse za uspjeh su za oba iste. U ovom slučaju, pogrešan odabir bi značio kasnije otkrivanje lijeka. Ukoliko se zatraži pomoć UI u donošenju odluke koji od dva moguća rješenja odabrati za daljnje istraživanje, to bi značilo postavljanje „sigurnog pitanja". UI može samo predložiti, ali nikako ne utječe na ishod, kao što i postavljanje pitanja UI, ne znači da čovjek sam nije sposoban donijeti tu odluku. Yampolsky također predlaže da se osigura ugradnja sigurnosnih mehanizama u svaki primjerak UO, kako se oni, u slučaju da im se omogući autonomija i kopiranje, u nekom trenutku ne bi oteli kontroli.[102] Međutim, ovakva pretpostavka ne čini se baš mogućom. Ako čovjek kreira autonomnu UI koja ima mogućnost implementiranja pametnijih i snažnijih karakteristika u svoje kopije, što bi je točno spriječilo da ukloni nekakve sigurnosne mehanizme koji su ljudi programirali možda nekoliko generacija prije? Nemojmo zaboraviti da bi se nekoliko generacija UI u slučaju ekspanzivnog repliciranja moglo izmijeniti u roku od nekoliko dana, tako da ljudi ne bi imali vremena reagirati nekakvih novim protumjerama.

[101] R. V. Yampolskiy, „Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach", str. 390.
[102] Isto, str. 391.

Zaključak

Tema razvoja umjetne inteligencije i budućnosti u koju nas taj razvoj vodi toliko je opširna da je ovaj rad nije nikako mogao obuhvatiti. Umjesto toga, zagrebano je samo po površini ovog zanimljivog područja kako bismo se upoznali s time što umjetna inteligencija nudi i kako se razvija. Time smo došli i do velikog broja pitanja koji se tiču samog načina uporabe UI, kao i mogućih posljedica koje će nam UI donijeti. Glavni razlog razvoja takve tehnologije vjerojatno je motiviran altruističkim idejama kako pomoći čovjeku u socijalnom i ekonomskom smislu i olakšati mu svakodnevicu, za početak, a kasnije i kako pokušati direktno utjecati na kvalitetu života u smislu poboljšanja zdravlja, fizičkih i mentalnih sposobnosti, svladavanja različitih bolesti i slično. Dosadašnji primjeri UI se i koriste upravo u navedene svrhe, tako da bi bilo nepošteno zanemariti dosad ostvarene dobrobiti, kao i osporavati sve pozitivne učinke koji će se možda tek i dogoditi. Međutim, činjenica je da čovjek razvojem takve vrste tehnologije ulazi u područje koje mu je nepoznato i ne može nikako tvrditi da može sa sigurnošću pretpostaviti smjer u kojem će se UI razvijati. Stoga je potrebno razvijati sigurnosne mjere i poduzimati što je moguće prilikom samog razvoja, jer ukoliko se u nekom trenutku razvije UI koja će imati zlonamjerne ciljeve, čovjek, koji razvojem, bilo u fizičkom ili mentalnom smislu, ne može parirati UI koja se vrlo brzo može duplicirati, neće imati dovoljno vremena za sprečavanje potencijalne katastrofe.

Iako su inteligentni sustavi, koji su najčešće portretirani u obliku robota, česti protagonosti različitih prikaza katastrofe i propadanja čovječanstva, jesu li oni doista stvarniji i opasniji od prikaza vještica i duhova? Smatram da, ako UI bude konstruirana na način da ne može razviti agresivnost te se utvrdi na koji način je najbolje implementirati ono najbolje od ljudskog roda, uz izostavljanje negativnih strana koje su osobine čovjeka, postoji mogućnost za bolji život koji bismo dostigli uz pomoć tehnologije. Ideja razvijanja samosvjesnih inteligentnih sustava, ne mora nužno isključivati mogućnost zajedničkog suživota ako se UI stvori na način da će biti u stanju prihvaćati moralne vrijednosti ljudi koje ipak u nekoj mjeri postoje na globalnoj razini. Pritom, ne smijemo zaboraviti ni pitanje tretiranja samih inteligentnih sustava, jer ukoliko se u nekom trenutku u njima doista razvije sposobnost mišljenja, čovječanstvo neće imati pravo tretirati ih samo kao mrtve objekte. U spomenutoj seriji *Westworld* razmatra se pitanje svijesti – gdje je granica svijesti? Je li moguće da se svijest razvije u umjetno konstruiranom sustavu ili i potencijal razvoja svijesti mora biti prethodno programiran? Etički problemi odnose se na problem interakcije između ljudi i robota s UI te s prikazuje odnos između nemoralnog ponašanja ljudi prema robotima što

dovodi do okretanja robota protiv ljudi. Moramo li prihvatiti robota s UI kao ravnopravnu osobu i interakciju s njom održavati u okviru etičkih načela ljudskog društva? Ako pretpostavimo da je moguće razvijanje svijesti u umjetnom entitetu, tada možemo i reći da bi ustanak robota protiv čovječanstva bio opravdan jer tlačenje i zlostavljanje ne može biti opravdano u nijednom scenariju.

Možemo zanemariti i mogućnost zlonamjerne UI i pretpostaviti scenarij, u kojem smo doista konstruirali UI koja želi isključivo pomoći čovječanstvu, ali zbog naše neopreznosti, to može rezultirati katastrofom. Primjerice, ukoliko se zatraži od UI da učini sve ljude na Zemlji sretnima i bezbrižnima, UI može odlučiti ubiti sve ljude, jer mrtvi ljudi zasigurno neće biti nesretni. Ili, može primjerice, lobotomizirati sve ljude kako bi ih učinila sretnijima. Ukoliko UI procijeni da je osmijeh pokazatelj sreće, može prisilno izvršiti plastične operacije kako bi svi izgledali nasmijano. Primjeri su prilično apsurdni, ali omogućavaju bolje razumijevanje što sve može poći po zlu, ako UI i bude u potpunosti usmjerena na ispunjavanje zadanih ciljeva. Teško je u vrijeme kad se čini da u svijetu prevladava iskrivljen sustav vrijednosti a etika i moralnost padaju na sve niže grane, točno procijeniti na koji način bi se etičke vrijednosti trebale ugrađivati u inteligentne sustave, kad vrlo često, takve principe nismo sposobni usaditi ni u naraštaje koji se tek odgajaju. Stoga je od iznimne važnosti da etika kao znanstvena disciplina zauzme jedno od glavnih mjesta u daljnjem razvoju tehnologije.

Neka etička pitanja odnose se na sigurnost razvoja UI, dok se neka bave pitanjima rizika potencijalnih negativnih posljedica. Iako moramo biti svjesni rizika, također moramo imati na umu da, općenito, tehnološki napredak znači bolji život za ljude. UI ima veliki potencijal, a njeno odgovorno korištenje ovisi isključivo o nama. Znači li to da je čovječanstvo doista prepušteno ideji „stvorit ćemo i nadati se najboljem“? Vrijeme će pokazati, u to možemo biti sigurni.

Popis literature

Anić, Vladimir, *Rječnik hrvatskoga jezika*, Novi liber, Zagreb, 1991.

Barišić, Pavo, „Tehniziranje etičkoga – etiziranje tehničkoga“, u: Igor Čatić (ur.), *Filozofija i tehnika*, Hrvatsko filozofsko društvo, Zagreb, 2003., str. 167–182.

Boddington, Paula, *Towards a Code of Ethics for Artificial Intelligence*, Springer International Publishing AG, (eBook), 2017.

Bostrom, Nick, „Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards“, 2001., https://nickbostrom.com/existential/risks.html.

Bostrom, Nick, „How Long Before Superintelligence“, 2006., https://nickbostrom.com/superintelligence.html.

Bostrom, Nick; Yudkowski, Eliezer, „The Ethics of Artificial Intelligence“, 2011., https://nickbostrom.com/ethics/artificial-intelligence.pdf.

Churchland, Paul M., *Matter and Consciousness*, A Bradford Book, The MIT Press, Cambridge, Massachusetts, 1999.

Collins, Harry, „Hoće li strojevi ikada misliti?“, *Treći program Hrvatskog radija*, 48, Zagreb,1995., str. 5–10.

Copeland, Jack, „Artificial Intelligence“, *Encyclopaedia Britannica*, 2018., https://www.britannica.com/technology/artificial-intelligence.

Copeland, Jack, „CYC Computer Science“, *Encyclopaedia Britannica*, 2018., https://www.britannica.com/topic/CYC.

Copeland, Jack, „Nouvelle Artificial Intelligence“, *Encyclopaedia Britannica*, 2018., https://www.britannica.com/technology/nouvelle-artificial-intelligence.

Copeland, Jack, „What is Artificial Intelligence?“, 2000., http://www.alanturing.net/turing_archive/pages/Reference%20Articles/What%20is%20AI.html#Int.

Dadić, Borislav, „Čovjekov duh pred izazovom tehnike“, u: Igor Čatić (ur.), *Filozofija i tehnika*, Hrvatsko filozofsko društvo, Zagreb, 2003., str. 121–133.

Dennet, Daniel C., „Umjetna inteligencija u filozofiji i psihologiji“, u: Nenad Miščević, Nenad Smokrović (ur.), *Računala, mozak i ljudski um*, Izdavački centar Rijeka, Rijeka, 2001., str. 43–63.

Dretske, Fred, „Strojevi i mentalno“, u: Nenad Miščević, Nenad Smokrović (ur.), *Računala, mozak i ljudski um*, Izdavački centar Rijeka, Rijeka, 2001., str. 85–99.

Ellul, Jacques, *Tehnika ili ulog veka*, Anarhija / Blok 45, Bratstvo iz Erevona, Beograd, 2010.

Hrvatska enciklopedija, mrežno izdanje, Leksikografski zavod *Miroslav Krleža*, http://www.enciklopedija.hr/.

Hyslop, Alec, „Other Minds“, *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), ur. Edward N. Zalta, https://plato.stanford.edu/archives/spr2016/entries/other-minds/.

Lelas, Srđan, *Promišljanje znanosti*, Hrvatsko filozofsko društvo, Zagreb, 1990.

McCarthy, John, „What is Artificial Intelligence?“, 2007., http://www-formal.stanford.edu/jmc/whatisai.pdf.

McDermott, Drew, „Why Ethics is a High Hurdle for AI“, 2008., http://www.cs.yale.edu/homes/dvm/papers/ethical-machine.pdf.

McDermott, Drew, „What Matters to a Machine?“, 2011., http://www.cs.yale.edu/homes/dvm/papers/whatmatters.pdf.

Nasuto, Slawomir J.; Bishop, John Mark, „Of (Zombie) Mice and Animats“, u: Vincent C. Muller (ur.), *Philosophy and Theory of Artificial Intelligence*, Springer, Berlin, Heidelberg, 2013., str. 85–106.

Norman, Donald A., „Kognitivne proteze“, u: Sian Griffiths (ur.), *Predviđanja: trideset velikih umova o budućnosti*, Naklada Jesenski i Turk, Zagreb, 2000., str. 201–211.

Pavaloiu, Alice; Kose, Utku, „Ethical Artificial Intelligence – An Open Question“, 2017., https://arxiv.org/ftp/arxiv/papers/1706/1706.03021.pdf.

Pećnjak, Davor, „Turingovi strojevi, Gödelov teorem i Searleova soba“, *Treći program Hrvatskog radija*, 48, Zagreb, 1995., str. 11–15.

Pećnjak, Davor, „Umjetna inteligencija: a priori ili empirijska znanost“, u: Zvonimir Čuljak, *Zbornik radova međunarodnog simpozija „Spoznaja i interpretacija“*, Institut za filozofiju, Zagreb, 2010., str. 151–157.

Radovan, Mario, „The Way of Power: Reflections on Technology, Nature and Society“, 2008., http://www.inf.uniri.hr/~mradovan/powercontent.htm.

Russell, Stuart J.; Norvig, Peter (ur.), *Artificial Intelligence: A Modern Approach*, Prentice-Hall, Upper Saddle River, 2003.

Schuster, Mike; Johnson, Melvin; Thorat, Nikhil, „Zero-Shot Translation with Google's Multilingual Neural Machine Translation System“, Google Research Blog, 2017., https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html.

Searle, John R., „Umovi, mozgovi i programi“, u: Nenad Miščević, Nenad Smokrović (ur.), *Računala, mozak i ljudski um*, Izdavački centar Rijeka, Rijeka, 2001., str. 134–154.

Vujić, Antun (ur.), *Opća i nacionalna enciklopedija u 20 knjiga*, Pro Leksis d.o.o. i Večernji list d.o.o. Zagreb, 2006., knjiga 9.

Warwick, Kevin, „Strojevi koji misle“, u: Sian Griffiths (ur.), *Predviđanja: trideset velikih umova o budućnosti*, Naklada Jesenski i Turk, Zagreb, 2000., str. 317–327.

Yampolskiy, Roman V., „Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach“, u: Vincent C. Muller (ur.), *Philosophy and Theory of Artificial Intelligence*, Springer, Berlin, Heidelberg, 2013., str. 389–976.

Yudkowsky, Eliezer, „Artificial Intelligence as a Positive and Negative Factor in Global Risk“, Machine Intelligence Research Institute, 2008., https://intelligence.org/files/AIPosNegFactor.pdf.