THEORIA 2
(2022) : 65 : str. 61-76

https://doi.org/10.2298/THEO2202061S
Originalni naučni rad
Original Scientific Paper

Mirjana Sokić

MORALNI STATUS VEŠTAČKE INTELIGENCIJE

APSTRAKT: Glavni cilj u ovom radu je analiza problema moralnog statusa veštačke inteligencije. Rad započinjem razjašnjenjem pojma moralnog statusa, kao i dihotomiju subjekt/objekt moralnog delanja, koja ima važnu ulogu u okviru mnogih dilema koje spadaju u primenjenu etiku. Ovo razjašnjenje je potrebno kako bi se, u poglavljima koja slede, došlo do jasnije slike o ključnim pitanjima na koja ću pokušati da odgovorim u radu; konkretnije, na pitanje (a) da li prema inteligentnom veštačkom sistemu možemo da postupamo na moralno (ne)ispravan način, kao i (b) da li inteligentni veštački sistem može i sam da postupa na način koji je moguće procenjivati u moralnim terminima.

KLJUČNE REČI: veštačka inteligencija, etički bihejviorizam, moralni delatnik, moralni status.

1. Uvod

Zamislite sledeći scenario.[1] Šetajući ulicom, Endru je naišao na grupu huligana koji su, bez ikakvog razloga, počeli da ga udaraju, šutiraju i da mu upućuju najstrašnije verbalne uvrede. Nakon nekoliko jakih udaraca, Endru pada na zemlju i doživljava oštećenja ekstremiteta i glave. Endru shvata da mu niko od prisutnih posmatrača neće pritrčati u pomoć i počinje da se brani, pri čemu huligane ozbiljno povređuje. Verujem da ćemo se svi složiti da je opisani postupak grupe huligana bio za moralnu osudu; takođe, gotovo svi ćemo se složiti da je Endru imao (moralno) pravo da se zaštiti od napasnika.[2] Činjenice o Endruovoj rasi, nacionalnosti, veroispovesti, stepenu

1 Ovaj rad je nastao u okviru naučnoistraživačkog projekta *Čovek i društvo u vreme krize*, koji finansira Filozofski fakultet Univerziteta u Beogradu.

2 Sa stavom prema kojem Endru ima moralno pravo da se zaštiti od fizičkih napada verovatno se ne bi složili jedino zastupnici radikalne verzije pacifizma. Za više informacija o inkoherentnosti radikalnog pacifizma, vidi u Narverson 1965.

inteligencije, ili seksualnim preferencijama danas se ne uzimaju kao relevantan faktor prilikom odlučivanja o ispravnosti napada na njega.[3]

Međutim, šta ako je Endru inteligentni humanoidni robot? Da li su huligani postupili na moralno neispravan način prema njemu? Da li su huligani uopšte mogli da naprave moralni prestup prema artefaktu, veštačkoj tvorevini? Da li Endru ima moralno pravo da se zaštiti od fizičkih napada iako će, prilikom tog procesa, ljudska bića biti povređena? Ukoliko bismo usvojili shvatanje prema kojem bi, na primer, etika robota bila koncipirana prema tzv. *asimovljevim zakonima robotike* – nazvanih po čuvenom piscu naučnofantastičnih romana, Isaku Asimovu (Isaac Asimov), koji ih je i formulisao – ispostavlja se da huligani iz našeg primera uopšte ne čine *moralni* prestup oštećujući Endrua; njihovo postupanje je pre nalik na oštećenje *privatne* svojine (u slučaju da je Endru vlasništvo nekog privatnog lica) ili oštećenja *javne* svojine (u slučaju da je Endru državno vlasništvo). U oba slučaja, njihov postupak ne bi bio neispravan u *moralnom* smislu, već bi predstavljao samo *pravni* prekršaj, a to nije ono što nas interesuje u sferi morala. Isto tako, asimovljevi zakoni robotike jasno zabranjuju da robot zaštiti svoje postojanje na način koji bi ugrozio ili povredio ljudska bića.[4] Prema tome, ispostavlja se da, za razliku od pomenutih činjenica od rasi, nacionalnosti i sl., činjenica da Endru predstavlja jedan veštački sistem, a ne biološki organizam, igra značajnu ulogu u našim etičkim procenama. Moj glavni cilj u ovom radu je da ispitam ispravnost i opravdanost usvajanja ove činjenice.[5]

2. Moralni status

Verujem da priznavanje moralnog statusa odraslom i, u mentalnom i fizičkom pogledu, potpuno zdravom čoveku ne izaziva nikakve kontroverze, ovakav čovek se standardno smatra paradigmatičnim primerom bića koji ima moralni status. Međutim, navođenje ovako neproblematičnog primera teško može da nam ukaže na ključne karakteristike koje jedno biće mora da ima kako bi mu se mogao pripisati moralni status. Prema jednoj grupi autora, moralni status robota proizlazi isključivo iz njihovog unutrašnjeg, mentalnog ustrojstva , tj. proizlazi iz činjenice da oni poseduju određene mentalne atribute, sposobnosti, sadržaje, kognitivnu strukturu itd. Drugim rečima, nije

3 Na žalost, uprkos nastojanjima etičara, ovo još uvek nije univerzalno prihvaćen stav.

4 Za više informacija o asimovljevim zakonima robotike, vidi u Asimov 1990; za više informacija o prigovorima ovim zakonima, vidi u Anderson 2008.

5 Koliko je reč o aktuelnom etičkom pitanju svedoči činjenica da je u poslednjih nekoliko godina u značajnom porastu broj autora koji brane tezu o moralnom statusu inteligentnih, autonomnih robota; vidi npr., Gunkel 2014; 2018a, 2018b; Levy 2009; Neely 2014; Schwitzgebel & Garza 2015; Danaher 2020.

dovoljno da robot samo izuzetno dobro „oponaša“, na primer, ljudsko ponašanje, postupke i odluke; ono što se traži jeste da to ponašanje, postupci i odluke imaju svoje poreklo u adekvatnoj vrsti mentalnog izvora (ili realizatora), kako bismo na osnovu njih mogli da pripišemo moralni status veštačkih entitetima (vidi, Nyholm & Frank 2017: 223).[6] Predlažem da ovakvo shvatanje nazovemo *internalizmom u pogledu moralnog statusa* robota. Internalistička pozicija među filozofima još uvek predstavlja standardno gledište o prirodi moralnog statusa, ali to ne znači da je danas usvajaju svi.

U nedavno objavljenom tekstu, Džon Danaher (John Danaher) odbacuje internalističko gledište i oslanja se na poziciju koju zove *etičkim bihejviorizmom* kako bi odbranio tezu prema kojoj autonomnim robotima možemo da pripišemo moralni status. Etički bihejviorizam je pozicija prema kojoj roboti mogu da imaju moralni status ukoliko su u *performativnom* ili *izvršnom* pogledu istovetni entitetima kojima se smisleno može pripisati moralni status (vidi, Danaher 2020: 2023). Konkretnije, ukoliko se neki robot *dosledno ponaša* na način koji je (u bihejvioralnom pogledu) istovetan ponašanju odraslog i mentalno potpuno zdravog čoveka (čiji moralni status se ne dovodi u pitanje), onda tom robotu možemo da pripišemo moralni status. Danaherova pozicija bi se mogla nazvati *eksternalizmom u pogledu moralnog statusa*. Glavne prednosti eksternalističke pozicije sastoje se u tome što nam omogućavaju da prilikom donošenja odluke o moralnom statusu veštačkih entiteta možemo da izostavimo problematičn i još uvek nerešena pitanja o prirodi svesti, mentalnih stanja i sl. – ukratko, ovu poziciju možemo da predstavimo u formi: „Nije bitno da li su roboti svesni ili ne, bitno je da li se ponašaju potpuno isto kao mi“.[7]

Međutim, pitanje o prirodi izvora moralnog statusa nije jedino pitanje koje treba razjasniti. Naime, ukoliko je Džon slagao, očigledno je da je Džon slagao *nekoga*; ukoliko je Džon prevario, on je prevario *nekoga*; ukratko, ukoliko je Džon izvršio bilo koji moralno (ne)ispravan čin, taj čin je morao da bude učinjen *nekome*. Jedna od značajnih distinkcija u etičkim raspravama tiče se činjenice da svaki vid moralnog postupanja nužno uključuje dve stranke:

(a) nekoga *ko vrši* postupak koji se može procenjivati u moralnim terminima,

(b) nekoga *nad kime se vrši* postupak koji se može procenjivati u moralnim terminima.

Treba imati u vidu da je navedena distinkcija prevashodno pojmovne ili logičke prirode, kao što primećuje Tomas Mekfirson (Thomas McPherson) u sledećem pasusu:

---

6 Niz argumenata koji su oslonjeni na navedenu, internalističku vrstu rezonovanja protiv pripisivanja određenih prava robotima nastao je kao reakcija na odluku iz 2017. godine, da humanoidni robot „Sofija“ – koju je konstruisala kompanija *Hanson Robotics* – dobije državljanstvo Saudijske Arabije. Za više informacija o ovim argumentima i prigovorima, vidi u Vincent 2017.

7 U ovom radu ću se uzdržati od zauzimanja konačnog stava u pogledu izbora između ove dve pozicije.

> Pojam moralnog delatnika generalno uključuje i pojam nekog *pacijenta*.[8] Ako neko izvrši postupak mučenja, neko drugi mora biti mučen; ako neko da obećanje, on to obećanje mora dati nekome itd. Pojam moralnog delatnika nema smisla ukoliko ga potpuno izolujemo od pojma *pacijenta*. Ovo je logička ili pojmovna, a ne moralna, poenta. (McPherson 1984: 173, moj kurziv)

Kao što možemo da vidimo iz ovog pasusa, svaki moralni postupak nužno podrazumeva dve klase. Prva klasa je klasa moralnih delatnika (eng. *moral agent*), tj. klasa svih onih bića (i entiteta)[9] čiji postupci se mogu procenjivati u moralnim terminima. Drugim rečima, radi se o klasi entiteta čiji postupci mogu biti (moralno) ispravni, neispravni, dopustivi, zabranjeni i sl., dok same ove entitete možemo okarakterisati kao (moralno) odgovorne, zaslužne, krive, dobre itd. Nazovimo klasu u koju spadaju ovi entiteti klasom „subjekata moralnog delanja" (u daljem tekstu SMD). Druga klasa u ovoj dihotomiji je klasa entiteta prema kojima je moguće postupiti na način koji se može procenjivati u moralnim terminima (eng. *moral patient*).[10] Nazovimo klasu ovih entiteta klasom „objekata moralnog delanja" (u daljem tekstu OMD).

Iako razlika između ove dve klase, bar na prvi pogled, ne izgleda problematično, to ipak nije slučaj.[11] Glavni problem se javlja kada pokušamo da precizno odredimo ekstenzije ovih klasa.[12] Na primer, čini se da prema stolu na kojem upravo pišem ne

---

8 U kurziv sam stavila reč „pacijent", koja na ovom mestu predstavlja bukvalan prevod engleske reči „*patient*". Ipak, ova reč se u daljem toku rada neće prevoditi bukvalno; za više informacija o prevodu ovog termina, vidi fusnota 9.

9 Iako se, prema standardnom shvatanju, moralnim delatnicima smatraju samo bića (tj. biološki organizmi), glavni cilj ovog rada sastoji se u ispitavanju da li i *veštački* entiteti (poput inteligentnih robota, kompjutera i sl.) mogu da spadaju u ovu klasu. Imajući to u vidu, u nastavku ću govoriti na neutralan način o *entitetima*, kako bih izbegla biološku predrasudu koja još na samom startu diskusije izostavlja veštačke sisteme iz sfere morala.

10 Na žalost, pokazuje se da je engleski izraz *moral patient* praktično nemoguće direktno prevesti na naš jezik – bukvalni prevod „moralni pacijent" svakako ne predstavlja adekvatno rešenje. Imajući u vidu glavne karakteristike koje u okviru ove dihotomije igraju centralnu ulogu – tj. da je reč o onim bićima (i entitetima) prema kojima se može postupiti na način koji je podložan proceni u moralnim terminima – odlučila sam se da ovaj izraz prevedem kao „objekt moralnog delanja". Kako bi se očuvala terminološka veza u okviru same dihotomije, klasu moralnih delatnika sam nazvala „klasom subjekata moralnog delanja".

11 Lučijano Floridi (Luciano Floridi) i Džej Sanders (J. Sanders) u zajedničkom tekstu ističu značaj ove dihotomije u kontekstu filozofskih rasprava koje se odnose na moralni status robota, mašina i veštačke inteligencije uopšte (Floridi & Sanders 2004: 350).

12 Još jedno veoma značajno pitanje je da li klase OMD i SMD predstavljaju *koekstenzivne* klase; tj. da li su svi pripadnici jedne klase ujedno ujedno pripadnici druge klase i *vice versa*? Prema gledištu koje je najzastupljenije među filozofima, svi entiteti koji spadaju u klasu OMD jesu i članovi klase SMD i *vice versa*. Ipak, ovo gledište je suočeno sa zamerkom prema kojoj postoje mnoga bića za koja bismo intuitivno rekli da spadaju u klasu OMD (npr. životinje, ljudi u trajnom vegetativnom stanju, fetusi itd.), iako svakako nisu moralni delatnici.

mogu da postupim na moralno (ne)ispravan način; ovaj sto ne mogu da oštetim u bilo kom moralno relevantnom smislu. U skladu s tim, ovaj sto – kao i ostali artefakti iz našeg svakodnevnog života – ne spada u klasu OMD. Međutim, šta je sa životinjama? Šta je sa fetusima (ili jajnim ćelijama, embrionima i sl.)? Da li svi ili bar neki od navedenih entiteta spadaju u ovu klasu? Sva ova pitanja ispostavljaju potrebu za specifikacijom nekog svojstva ili karakteristike na osnovu koje se određuje ekstenzija ove dve klase. Razmotrimo ukratko neke od najčešćih pozicija.

Standardna pozicija. Gledište koje još uvek predstavlja najrasprostranjeniji stav o tome ko ili šta spada u sferu morala dolazi do nas – bar kako primećuje Piter Singer (Peter Singer) – iz judeo-hrišćanske tradicije, prema kojoj samo ljudi (tj. članovi vrste *Homo sapiens*) mogu biti objekti moralnog delanja (vidi, Singer 1993: 88-89).[13] Ipak, ovo gledište – uprkos činjenici da je skoro dva milenijuma predstavljalo gotovo univerzalno prihvaćeni standard za određenje klase OMD – suočeno je sa jakim prigovorom. Naime, uopšte se ne vidi zašto bi neko biće bilo isključeno iz sfere moralnih razmatranja na osnovu puke činjenice da pripada određenoj rasi ili vrsti. Singer smatra da je ovakvo određenje klase OMD u osnovi svih rasističkih, nacionalističkih, vrstističkih i, generalno, diskriminativnih moralnih shvatanja (Singer 1993: 88).

Utilitaristička pozicija. U radovima Džeremi Bentama (Jeremy Bentham) pronalazimo kriterijum koji je dosta naklonjeniji neljudskim bićima, a u okviru kojeg ključnu ulogu igra sposobnost za patnju (tj. za osećanje bola, patnje, neprijatnosti i sl.). Osnovne postavke ovog kriterijuma na najprecizniji način su izložene u čuvenom pasusu, koji navodim u celosti:

> Možda će doći dan kada će ostatak životinjskih stvorenja steći ona prava koja im nikad ne bi mogla biti uskraćena osim od tiranske ruke. Francuzi su već otkrili da tamna koža nije razlog da ljudsko biće bude prepušteno hirovitosti svog mučitelja. Moguće je da će doći dan kada će se shvatiti da broj nogu, dlakavost kože, ili završetak karlične kosi, ne predstavljaju dovoljne razloge za prepuštanje jednog osećajnog bića istoj takvoj sudbini. Šta bi još moglo da zacrta tu nepremostivu liniju? Da li je to sposobnost za rasuđivanje, ili možda sposobnost za govor? Ali odrasli konj ili pas je neuporedivo razumnija životinja i mnogo prikladnija za konverzaciju od novorođenčeta starog jedan dan, ili nedelju dana, ili čak mesec dana. Ali čak i da pretpostavimo da je drugačije, čemu bi to služilo? Pitanje ne glasi 'Da li oni mogu da rasuđuju?' niti 'Da li oni mogu da govore?', već 'Da li mogu da pate?' (Bentham 2007, XVII, IV, n. 1.)

Ukratko, Bentam smatra da neko biće spada u klasu OMD ukoliko je sposobno da oseća bol, patnju ili neprijatnost, dok činjenice o vrsti, rasi, nacionalnosti, inteligenciji,

13 Singer ovu poziciju na više mesta naziva *doktrinom o svetosti ljudskog života* (Singer 1993: 83-85, 117, 150).

sposobnosti za govor i sl., ne predstavljaju relevantnu osnovu za uključenje ili isključenje nekog entiteta iz ove klase. U skladu s tim, ukoliko svojim postupkom entitetu *x* mogu da izazovem bol ili zadovoljstvo, onda *x* spada u klasu entiteta prema kojima mogu da postupim na moralno (ne)ispravan način.

Pozicija dubinske ekologije. Mnogi filozofi smatraju da opisani utilitaristički kriterijum ne ispostavlja ni nužne ni dovoljne uslove za određenje pojma moralnog pacijenta. Ne ispostavlja nužne uslove, jer mogu učiniti (moralno) neispravan postupak prema (i) *neživom* objektu koji nema sposobnost za bol i zadovoljstvo (npr., huliganski čin kojim se uništava ili oštećuje neko umetničko delo), kao i (ii) prema *živom* biću koje nema sposobnost za bol i zadovoljstvo (npr., čin kojim se seče stablo staro dve hiljade godina).[14] Ne ispostavlja dovoljne uslove, jer izazivanje bola i/ili zadovoljstva (*per se*) ne mora da ima moralne kvalifikacije. U osnovi ovog zaključivanja leži uverenje da zadovoljstvo i bol živih bića ne povlači moralni značaj bez dodatnih kvalifikacija o prirodi tih bića.[15] Čitava ideja, kao i sam naziv, dubinske ekologije (*deep ecology*) proizašla je iz kratkog teksta norveškog filozofa Arne Nesa (Arne Naess 1973). Najkraće rečeno, nastojanja zastupnika dubinske ekologije jeste proširenje sfere moralne procene i na *nežive* objekte; prema ovoj poziciji, među objekte moralnog delanja spadale bi ne samo sve pojedinačne žive jedinke različitih vrsta, već i ekosistemi, biološke vrste,[16] pa čak i biosfera u celosti.

Lokovska pozicija. Džon Lok (John Locke) je smatrao da je ličnost ili osoba (*person*) „misleće razumno biće koje ima um i moć refleksije i koje može da smatra sebe sobom, jednom istom mislećom stvari u raznim vremenima i mestima; a ono to čini pomoću one svesti koja je neodvojiva od mišljenja – koja je čak, čini se, bitan deo mišljenja; jer kada neko opaža, on svakako mora i opažati da opaža“ (Lok 1962: 358). Ovaj lokovski uslov ličnosti ne mora biti *nužan*, ali svakako može biti *dovoljan* za članstvo u klasi OMD.[17] Ukoliko usvojimo takvo gledište, ispostavlja se da ne postoje nikakve principijelne prepreke da u ovu klasu uvedemo i inteligentne veštačke osobe

14 U oba slučaja, moguće je odgovoriti da se neispravnost navedenih postupaka sastoji u izazivanju nezadovoljstva kod *svesnih* bića. Prema tome, svi navedeni primeri se lako mogu podvesti pod bentamovski kriterijum.

15 Najčešće korišćeni primeri su životinje, pripadnici drugih rasa, osobe sa mentalnim poremećajima i sl.

16 Ovde treba naglasiti da se pod „biološkim vrstama“ ne misli na skup ili konglomerat živih jedinki, već pre na apstraktni pojam čija moralna vrednost i status ne predstavlja puki zbir moralne vrednosti jedinki koje ga čine.

17 Da pojasnim, ukoliko entitet *x* ne zadovoljava lokovski uslov za ličnost, *x* još uvek može (na osnovu svojih drugih karakteristika, kao što je, na primer, *x*-ova sposobnost za patnju) da bude uključen u klasu OMD; međutim, ukoliko *x* zadovoljava lokovski uslov za ličnost, prema gledištu koje zastupam, *x* bi svakako bilo uključeno u ovu klasu. To je ono što podrazumevam kada kažem da ovaj uslov ne mora biti nužan, ali je dovoljan.

(bili oni roboti, kompjuteri, mašine ili nešto slično). Ova pozicija je nazvana „lokovskom“ zato što se oslanja na Lokovo određenje nužnih i dovoljnih uslova koje neki entitet mora da zadovolji kako bi bio osoba ili ličnost, pri čemu ne želim da sugerišem da je sam Lok izveo i dalji zaključak da samo i isključivo osobe mogu da budu uključene u klasu OMD, iako činjenica da je termin „osoba“ smatrao „forenzičkim“ izrazom (Lok 1962: 372) – tj. izrazom koji ima *pravne i etičke* konotacije – svakako omogućava interpretacije koje idu u tom pravcu.[18]

Navedena četiri kriterijuma možemo da sumiramo na sledeći način: prema prvom, klasu OMD čine samo ljudi, prema drugom sva bića koja mogu da osećaju, prema trećem u ovu klasu spadaju čak i apstraktni entiteti poput ekosistema i biosfere, dok prema četvrtom u ovu klasu spadaju oni entiteti koji zadovoljavaju lokovski kriterijum ličnosti. U narednom poglavlju ću pokušati da pokažem da nam kombinacija utilitarističkog i lokovskog kriterijuma pruža osnove za uključivanje veštačkih inteligentnih entiteta u klasu OMD.

3. Da li inteligentne robote i mašine možemo da uvrstimo u klasu OMD?

Verovatno tipičan odgovor na pitanje sadržano u naslovu ovog poglavlja nalazimo u stavu Džoane Brajson (Joanna Bryson), koja smatra da metafora o gospodaru i robu pruža adekvatan i ispravan opis odnosa između ljudi i robota (Bryson 2008: 65). Međutim, u mnogim slučajevima ova metafora ima dosta neprimeren karakter. Drugim rečima, ideja o braku sa robotom – koju predviđa autor Dejvid Livaj (David Levy 2008) – ili o seksualnom opštenju sa robotom (koja je danas aktualizovana i postaje sve popularnija) ne slaže se sa opisom tog odnosa po modelu metafore gospodara i roba. Stoga pitanje da li roboti predstavljaju entitete prema kojima možemo da postupamo na način koji se može procenjivati u moralnim terminima, a kojim ćemo se baviti u toku ovog poglavlja, ne samo što predstavlja pitanje koje ima realan značaj, već je i centar živih debata u oblasti etike veštačke inteligencije i robotike (vidi, Müller 2020).

Na prvi pogled, utilitarističko gledište nam ispostavlja prihvatljiv kriterijum za uključenje nekog bića u klasu OMD – ne verujem da bi iko osporio shvatanje prema kojem, ukoliko neko biće (ili neki entitet) *x* oseća patnju, bol ili neprijatnost prilikom nekog postupka, upravo ta patnja (bol, neprijatnost) ispostavlja (moralne) *prima facie* razloge da ovaj postupak *x*-u ne činimo; drugim rečima, činjenica da *x* ima sposobnost za patnju i interese svakako ga uključuje u klasu OMD. U skladu s tim kriterijumom, ukoliko budući tehnološki razvoj omogući nekom veštačkom entitetu da ima toliko razvijenu kognitivnu arhitekturu na osnovu koje bi nedvosmisleno imao sposobnost

18 Za više o lokovskom kriterijumu ličnosti, kao i shvatanju forenzičnosti termina „osoba“, vidi u Milojević 2018.

za doživljavanje neprijatnosti i zadovoljstva, taj entitet bi trebalo uključiti u klasu objekata moralnog delanja. Vratimo se na naš primer sa početka teksta, ukoliko bi robot Endru mogao da oseti patnju (ili neprijatnost) prilikom napada huligana, to bi svakako predstavljalo osnov za *moralnu* osudu napada na njega. Mislim da je ovaj zaključak najbolje obrazložiti pomoću primera iz novije kinematografije.

U naučnofantastičnom trileru *Ex Machina* (2014), Kejleb (Domhnall Gleeson) – mladi i talentovani programer – dobija jedinstvenu priliku da sprovede Tjuringov test na humanoidnom inteligentnom robotu Evi (Alicia Vikander). Tokom testa, Kejleb saznaje od Eve da njen tvorac Nejtan (Oscar Isaac) namerava da je deaktivira i izbriše sva njena sećanja, što bi za nju predstavljalo proces prilikom kojeg bi njena sadašnja ličnost bila u potpunosti „uništena". Po pretpostavci koja se eksplicitno navodi u filmu, Eva je imala sposobnost ne samo za kognitivna, već i za afektivna i konativna stanja; u skladu s tim, za Evu se može reći da *nije želela* da bude deaktivirana te da joj je misao o deaktivaciji pričinjavala *neprijatnost*, *patnju* i sl. Prema utilitarističkom kriterijumu nema sumnje da bi Evina trajna deaktivacija – kao i bludničenje koje Nejtan u filmu sprovodi nad svojim robotima – predstavljali oblike moralno neispravnog postupanja. Međutim, Eva možda ne predstavlja dovoljno problematičan slučaj i to iz dva razloga. Prvo, Eva je humanoidni inteligentni robot, pri čemu upravo njena humanoidna konstitucija, lik i gestikulacije prilikom komuniciranja mogu da budu osnova za brojne predrasude u korist njenog uključenja u klasu OMD. Drugo, po pretpostavci, Eva poseduje mentalna stanja iz sve tri standardne psihološke grupe – kognitivna, afektivna, konativna – što zapravo znači da se jedina razlika između nje i bioloških bića sastoji u tome što njen mentalni život realizuju silikonski čipovi, umesto neurona. Navedena razlika u pogledu realizatora mentalnih stanja, međutim, ne bi smela da predstavlja relevantan osnov za njeno isključenje iz klase OMD.

Zapitajmo se sada, šta se dešava kada sa slučaja koji je (uslovno rečeno) neproblematičan (kao što je slučaj sa Evom) pređemo na slučajeve veštačke inteligencije koji se razlikuju u pogledu toga što (a) nije reč o humanoidnim robotima, i (b) što ne mogu (ili je upitno da li mogu) da imaju afektivna i konativna stanja? I ovde će nam biti od pomoći primer iz kinematografije; naime, razmotrimo slučaj mašine HAL-9000 iz čuvenog Kubrikovog (Stanley Kubrick) filma „2001: Odiseja u svemiru" [*2001: A Space Odyssey*]. Ovaj primer, zapravo, povlači mnogo zanimljivijih i kontroverznijih pitanja za razmatranje. Prvo, veliko je pitanje da li je HAL-9000 imao *bona fide* afektivna i konativna stanja (što u slučaju Eve nije upitno), ili je samo veoma uspešno *simulirao* ova stanja. Drugo, pitanje je da li je uspešna simulacija ovih stanja dovoljna za pripisivanje ovih stanja jednom veštačkom sistemu ili entitetu.[19] Treće, pitanje je da li je skup kognitivnih stanja (kao što su verovanja, koja HAL-u svakako možemo

19 Na ovo pitanje se često daje odričan odgovor, ali to svakako ne treba učiniti bez dobrog filozofskog obrazloženja i razvijene argumentacije.

da pripišemo) dovoljan za izvođenje preferencija i interesa.[20] Nijedno od navedenih pitanja nije moguće detaljno razmotriti u radu ovako ograničenih razmera. Međutim, u nastavku ću pretpostaviti najčešće zastupanu poziciju, prema kojoj se na sva ova pitanja može dati određan odgovor. Da li u okvirima takve pozicije ima osnova za tvrđenje da HAL spada u klasu OMD? Razmotrimo scenu iz filma u kojoj je HAL deaktiviran. HAL svakako *veruje* da će biti deaktiviran, on *shvata* pojam svog nepostojanja koje će uslediti nakon deaktivacije;[21] HAL može da *zamisli* i da *razume* svoje postojanje u prošlom i budućem vremenu i ima sećanja iz ranijih perioda u kojima je postojao.

Svakako, donošenje nedvosmislene odluke o tome da li ćemo HAL-a (kojeg smo opisali na takav način da pored čisto kognitivnih stanja, ne poseduje nikakve interese ni preferencije) uvrstiti u klasu OMD je složeno i iskreno sumnjam da ću biti u stanju da uspešno odbranim i obrazložim bilo koju od opcija. U skladu s tim, jedino što mi preostaje jeste da navedem svoje shvatanje, bez ikakve ambiciozne pretenzije da time iznosim konkluzivne argumente ili razloge u pogledu ovog pitanja. Na osnovu razmatranja koja sam posvetila problematici vezanoj za moralni status veštačke inteligencije, sklona sam da usvojim shvatanje prema kojem jedino što može da uvrsti mašinu poput HAL-a (koja poseduje isključivo kognitivna stanja i nema nikakve preferencije) u klasu OMD jeste neki vid *poštovanja* prema činjenici da je reč o *racionalnom* i *inteligentnom* entitetu, koji shvata pojam svog vlastitog postojanja u toku vremena (i pojam svog nepostojanja i deaktivacije), kao i da se radi o entitetu čija mentalna stanja konstituišu *jedinstvenu individuu* ili *ličnost* (shvaćenu u lokovskim terminima kao specifično jedinstvo svesti). Dakle, ograničiću se na tvrđenje da upravo navedeni skup faktora i razmatranja omogućava da ovakav veštački entitet uključimo u klasu OMD. Ovo tvrđenje je u skladu sa lokovskim kriterijumom. Sumirajmo sada glavne zaključke koji su izloženi u toku ovog poglavlja. Utilitaristički kriterijum omogućava uvođenje u klasu OMD onih veštačkih entiteta koji imaju sposobnost za doživljavanje patnje, neprijatnosti i bola, dok bi lokovski kriterijum omogućio uvođenje u ovu klasu čak i onih veštačkih entiteta koji ovu sposobnost nemaju, ukoliko zadovoljavaju lokovski kriterijum ličnosti.

---

20 Kao i u slučaju prethodnog pitanja, najčešće se na ovo pitanje daje odričan odgovor.

21 Na osnovu premisa iznesenih u filmu HAL očigledno *nastoji* da spreči svoju deaktivaciju; štaviše, HAL ubija članove posade kako bi sprečio ovakav ishod, što sugeriše da on *preferira* svoje postojanje u odnosu na svoje nepostojanje. Ipak, ja ću se u ovom radu ograničiti i pripisati HAL-u samo skup čisto kognitivnih stanja, iako film jasno sugeriše da on poseduje i određena druga stanja koja, strogo rečeno, ne možemo da okarakterišemo na taj način.

4. Da li inteligentne robote i mašine možemo da smatramo moralnim delatnicima?

U tekstu „Kada je robot moralni delatnik" Džon Salins (John Sullins, 2006) brani tezu prema kojoj, pod određenim okolnostima, tzv. autonomni roboti mogu da budu tretirani kao moralni delatnici. Pre nego što se upustimo u analizu razloga za i protiv uključenja robota u klasu subjekata moralnog delanja (ili popularnije, u klasu moralnih delatnika), potrebno je pojasniti u kom smislu robot uopšte može da bude delatnik. Ukoliko je robot naprosto „oruđe", onda se pitanje o moralnom statusu njegovih postupaka automatski prebacuje na njegove korisnike i/ili programere (dizajnere). U slučaju nekih robota (kao što su danas popularni „teleroboti" – tj. mašine koje su daljinski upravljane i koje vrše minimalno autonomne radnje), ovo je svakako tačno. Ipak, ovakav tip robota teško da ostavlja prostor za debatu; gotovo niko ne dovodi u pitanje da u slučaju radnji koje ovakvi roboti izvršavaju odgovornost i krivicu snose oni koji njima daljinski upravljaju. Ovo naročito važi u slučaju telerobota koji se koriste u ratovima – kao što su npr. naoružani dronovi. Naime, imajući u vidu da ljudi u ovim slučajevima donose sve glavne odluke koje mašine izvršavaju, sasvim je jasno da upravo ti ljudi ispostavljaju moralno rasuđivanje koje može da bude predmet kritike ili pohvale. Roboti koji predstavljaju mnogo zanimljiviji problem jesu tzv. „autonomni roboti". Naravno, pojam autonomije je veoma složen filozofski problem te je kompletnu analizu ovog pojma nemoguće pružiti u okvirima rada ovako ograničenog obima. U skladu s tim, slediću upotrebu ovog termina koju koristi Salins, ograničiću se na upotrebu termina „autonomni robot" koji je danas standardan u robotici – reč je o robotima koji bar neke od značajnih odluka o svojim radnjama donose na osnovu svog programa (Arkin 2009; Lin et al. 2008).[22] Ukoliko pojam robota shvatimo na ovaj način, postoje četiri moguća gledišta u pogledu pitanja da li roboti mogu da budu moralni delatnici. Razmotrimo ova gledišta po redu.

Pozicija 1: Roboti nisu moralni delatnici, ali bi mogli da postanu u budućnosti. Drugim rečima, još uvek nismo dostigli nivo tehničkog razvoja koji bi nam omogućio

22 Treba naglasiti da stipulativno određenje pojma „autonomije" koje je ovde navedeno nije neproblematično; naime, u debatama koje se odnose na strukturu programa na osnovu kojeg bi tzv. „samovozeći automobili" u realnim okolnostima „odlučivali" u teškim slučajevima o tome šta je bolji ishod – npr., da li u situaciji u kojoj je sudar neminovan auto treba da skrene i ubije jednu osobu ili da ostane u istoj traci i ubije više ljudi – moguće je reći da, iako ova vozila zaista predstavljaju autonomne robote u gore navedenom smislu reči (budući da njihove „odluke" zavise od njihovog programa) to ne oslobađa od moralne odgovornost programere koji su taj program napisali. Ono što opisani problem čini još težim jeste to što se etičari čak i danas jasno dele u dva tabora u pogledu toga šta čini zadovoljavajuće rešenje ovog problema, dok se rešenje (kako izgleda) još uvek ne nazire. Ukratko, sve dok se ne dođe do nekog konkretnog rešenja ove dileme, oba ishoda za koji bi se programer odlučio ne mogu da budu podložna opravdanoj i racionalnoj kritici.

konstruisanje robota koji zadovoljavaju uslove za moralne delatnike, ali to je samo tehnička prepreka. Danijel Denet (Daniel Dennett) zastupa ovu poziciju u svom eseju „Kada HAL ubije, koga treba kriviti?“ (1998). Denet u tekstu razmatra slučaj robota HAL-a, o kojem je bilo reči u prethodnom poglavlju i tvrdi da se HAL može opravdano smatrati ubicom, imajući u vidu da je, kako on smatra, mašina svakako imala *mens rea* (tj. zločinačku nameru). Denet u tekstu ističe da je jedan od ključnih uslova koje robot mora da zadovolji kako bi mogao da bude moralni delatnik (te da bude entitet koji s pravom možemo da kaznimo, osudimo, nagradimo i sl.) jeste „intencionalnost višeg reda“. Drugim rečima, robot bi morao da ima verovanja o svojim verovanjima, želje o svojim željama, verovanja o svojim mislima, nadanjima itd. Iako se mašine koje danas konstruišemo ne zadovoljavaju ove uslove, Denet ne vidi nikakav načelni razlog koji bi onemogućio da se to stanje u budućnosti promeni.

Pozicija 2: Roboti nisu moralni delatnici i ova vrsta veštački stvorenih entiteta nikad neće biti moralni delatnici. Selmer Bringsjord (Bringsjord 2007) zastupa ovu poziciju. On smatra da roboti nikada neće biti autonomni, s obzirom na to da nikada neće moći da urade nešto u pogledu čega nisu unapred bili programirani da urade. Primer koji Bringsjord koristi je robot PERI; ovaj robot je programiran da donese odluku u pogledu toga da ili (a) ispusti kuglu na pod, ili (b) da zadrži kuglu u svojoj mehaničkoj šaci. Za koji god ishod da se PERI odluči, Bringsjord smatra da će postupak biti *determinisan* njegovim programom. Shodno tome, PERI (kao ni bilo koji robot kojeg možemo da konstruišemo sada ili u budućnosti) ne može da ima slobodnu volju u dovoljno jakom smislu koji je neophodan za moralnog delatnika. Kako Džon Salins primećuje, pozicija koju Bringsjord zastupa suočava se sa ozbiljnim prigovorom. Naime, ljudi su svakako proizvod najrazličitijih faktora (okoline, kulture, društvenog statusa, obrazovanja, hemije u mozgu itd.). Imajući to u vidu, možemo da zaključimo da ni ljudi nisu (i ne mogu biti) moralni delatnici, budući da ni ljudska verovanja, želje, namere, ciljevi nisu (strogo rečeno) autonomni u dovoljno jakom smislu. U najmanju ruku, ovaj argument bi trebalo osporiti kako bi se zastupao isključiv zaključak koji Bringsjord usvaja. Prema tome, jak pojam slobodne volje koji Bringsjord zahteva u slučaju robota predstavlja zahtev za koji se lako može ispostaviti da ga čak ni ljudska bića ne zadovoljavaju.

Pozicija 3: Roboti zaista jesu moralni delatnici, ali ljudi nisu. Na prvi pogled ovo izgleda kao totalno bizarno stanovište i nisam sigurna da možemo pretpostaviti da bi ga bilo ko ozbiljno zastupao. Ali, ovaj prvi utisak je pogrešan, budući da u tekstu objavljenom 2006. Džozef Emil Nejdou (Joseph Emile Nadeau) brani upravo ovu poziciju. Kako Nejdou smatra, neki postupak je *slobodan* ako i samo ako proizlazi (tj. sledi) iz razloga koje je delatnik u potpunosti promislio. Posledica toga je da samo delatnik koji operiše prema striktno logičkim teoremama može da vrši radnje koje su slobodne u upravo opisanom smislu. Kako on ističe, roboti su eksplicitno programirani

da budu delatnici koji zadovoljavaju ovaj uslov, dok ljudi, s druge strane, nisu slobodni delatnici u ovom smislu reči. Ukratko, u slučaju ljudi, postupak uvek uključuje i neki iracionalni element, budući da nikad *ne sledi* (u striktno logičkom smislu) iz razloga koje je delatnik uzeo u obzir. Pored toga, postupci ljudskih bića su problematični u tom pogledu što nikada ne proizlaze iz *svih* razmatranja koja su nam dostupna: mi se po pravilu ograničavamo na samo jedan broj razloga koji su nam dostupni prilikom donošenja odluke o tome kako treba da postupamo. U slučaju autonomnih robota i mašina, kako smatra Nejdou, postupci su lišeni oba navedena nedostatka, a upravo to je ono što se zahteva od jednog moralnog delatnika u punom smislu reči.

Pozicija 4: Rasprave o moralnom statusu robota i ljudi mogu imati više ishoda, a sve zavisi od načina na koji shvatamo pojmove autonomije, intencionalnosti i odgovornosti. Lučijano Floridi (Luciano Floridi) i Džej Sanders (J. Sanders) ističu da tradicionalno određenje pojma moralnog delatnika u filozofskoj literaturi zavisi prvenstveno od pojmova autonomije, slobodne volje, intencionalnosti, odgovornosti i sl.[23] Ovo su, međutim, filozofski pojmovi koji nam još uvek ne omogućavaju jednoznačno određenje ekstenzije pojma moralnog delatnika. Naime, u zavisnosti od nivoa apstrakcije u određenju navedenih pojmova, moguće je braniti (a) poziciju prema kojoj da roboti *jesu* (ili bar mogu da budu) moralni delatnici, kao i poziciju na potpuno suprotnoj strani spektra prema kojoj (b) ljudi *nisu* (i ne mogu da budu) moralni delatnici (Floridi & Sanders 2004). Ukratko, odgovor na pitanje „Da li su roboti moralni delatnici?“ koji bismo mogli da izvedemo na osnovu ove pozicije jeste da sve zavisi od nivoa apstrakcije u pogledu pojmova autonomije, intencionalnosti i odgovornosti koji smo spremni da usvojimo. U sledećem odeljku ću razmotriti koji nivo apstrakcije ovih pojmova je neophodno usvojiti kako bismo robote uključili u klasu moralnih delatnika.

5. Roboti kao moralni delatnici

Džon Salins brani tezu prema kojoj do odgovora na pitanje „Da li roboti mogu da budu moralni delatnici?“ možemo doći ukoliko pružimo odgovor na sledeća tri, blisko povezana, pitanja:

(i) Da li robotu možemo pripisati autonomiju?

(ii) Da li je ponašanje robota intencionalno?

(iii) Da li robot ima odgovornost i obaveze prema drugim moralnim delatnicima?

23 U pogledu ove tvrdnje Floridi i Sanders su svakako u pravu, jer su u sva tri gledišta koja smo do sada razmotrili, pojmovi slobode volje i intencionalnosti igrali ključnu ulogu.

Oslanjajući se na poziciju Floridija i Sandersa (2004), Salins određuje nivo apstrakcije na način koji mu omogućava da na sva tri pitanja pruži potvrdan odgovor:

(i) *Autonomija*: Salins određuje pojam autonomije na način koji implicira da mašina nije pod direktnom kontrolom bilo kog drugog delatnika. On smatra da ukoliko robot zadovoljava opisani nivo autonomije, onda se može smatrati samostalnim delatnikom u relevantnom smislu reči. Ipak, on ističe da ovaj stepen autonomije nije dovoljan da neki entitet (ili biće) bude moralni delatnik, budući da bakterije, životinje, pa čak i ekosistemi ili kompjuterski virusi mogu lako da zadovolje ovaj uslov. U skladu s tim, Salins smatra da je ovaj uslov *nužan*, iako *ne i dovoljan* uslov za moralnog delatnika – drugim rečima, moralni delatnik mora da zadovolji ovaj uslov, ali neće svaki entitet koji ovaj uslov zadovolji ujedno biti moralni delatnik.

(ii) *Intencionalnost*: robot mora da ima sposobnost da postupa na intencionalan način. Kako Salins ističe, ukoliko je ponašanje robota dovoljno kompleksno da nas „navodi" da te postupke opisujemo koristeći se terminima folk psihologije (poput „*x* namerava", „*x* želi", „*x* veruje"...), možemo da zaključimo da robot zadovoljava uslov intencionalnosti.[24]

(iii) *Odgovornost*: Salins smatra da robot jeste moralni delatnik ukoliko postupa na takav način da možemo da pripišemo smisao njegovim postupcima samo ukoliko pretpostavimo da on ima odgovornost i obaveze prema nekom drugom moralnom delatniku. Drugim rečima, ukoliko robot vrši neku ulogu sa kojom su povezane određene socijalne obaveze i ukoliko je ponašanje tog robota takvo da je jedini način da ga shvatimo jeste da mu pripišemo „verovanje" da ima dužnost ili obavezu prema nekom drugom moralnom delatniku, onda tog robota možemo da opišemo kao moralnog delatnika.[25]

Na osnovu toga, Salins zaključuje da ne samo savršeniji roboti budućnosti, već čak i neki interaktivni roboti koje danas konstruišemo, jesu moralni delatnici sa određenim moralnim pravima, kao i obavezama i odgovornostima (vidi, Sullins 2006: 30).[26] Na osnovu analize Salinsovog teksta, glavni utisak do kog sam došla je da je

---

24 Ovo određenje intencionalnosti, kao i određenje odgovornosti koje se navodi u nastavku, ima očigledno eksternalistički karakter.

25 Dejvid Gankel (David Gunkel) – jedan od pobornika teze o moralnim pravima robota – tvrdi da upravo vrste međusobnih odnosa koje će ljudi i napredni roboti u budućnosti imati jeste glavna vrsta razmatranja koju treba uzeti u obzir prilikom dolaženja do odluke o tome da li im treba priznati određena moralna prava ili ne (Gunkel 2012, 2014).

26 Očigledno pitanje koje se ovde može postaviti je koji roboti današnjice bi zadovoljavali opis koji Salins navodi. Iznenađuje činjenica da Salins u tekstu ne nudi odgovor na ovo pitanje, ali na osnovu svega što u tekstu kaže, verujem da bi on tvrdio da samovozeći automobili, čet-botovi i sex-botovi već sada predstavljaju moralne delatnike.

njegov zaključak uverljiv u jednakoj meri u kojoj je uverljiv nivo apstrakcije pojmova autonomije, intencionalnosti i odgovornosti koji je u tom tekstu usvojio. Iako njegov zaključak omogućava velikodušno proširenje klase moralnih delatnika, pitanje koje se teško može izbeći je da li uvođenje različitih nivoa apstrakcije dovode do elementa proizvoljnosti u čitavu problematiku o moralnog statusu robota. Naime, odgovori poput Denetovog, Bringsjordovog ili Nejdouovog mogu nam se činiti radikalnim ili čak potpuno pogrešnim, ali se svakako ne može reći da pružaju nedvosmislen odgovor na pitanje o moralnom statusu robota. Kako mi se čini, Floridi i Sanders, a sa njima i Salins, na ovo pitanje odgovaraju tako što za svakoga nude mogućnost da ga uskladi sa svojim senzibilitetom, po cenu gubljenja objektivnih merila za pripisivanje moralnog statusa određenoj vrsti entiteta. Bar meni ovo ne izgleda kao prihvatljiv ustupak. U skladu s tim, pozicija koja mi se čini najprihvatljivijom i najkonkretnijom jeste upravo Denetova – iako roboti nisu moralni delatnici u trenutnom stepenu tehnološkog razvoja, nema nikakvih načelnih prepreka da to postanu u budućnosti. Konkretnije rečeno, ukoliko roboti u budućnosti budu zadovoljavali lokovski kriterijum ličnosti – tj. ukoliko to budu racionalni, inteligentni entiteti, koji shvataju pojam svog postojanja u toku vremena itd. – sklona sam da prihvatim da bi oni zadovoljili ključni uslov za uključenje u klasu moralnih delatnika.

6. Zaključna razmatranja

U ovom radu sam izložila nekoliko dilema u pogledu moralnog statusa, pri čemu je fokus stavljen na distinkciju između subjekata i objekata moralnog delanja. Moj glavni cilj u radu je prevashodno bio da ukažem na argumente, razloge i filozofske pozicije na osnovu kojih bismo autonomne i inteligentne robote i mašine mogli da uključimo u obe klase koje konstituišu moralni status; tj. u klasu subjekata i objekata moralnog delanja. Konačan zaključak koji sam usvojila u radu može se sumirati na sledeći način: specifična kombinacija utilitarističkog i lokovskog kriterijuma omogućava uključenje inteligentnih veštačkih entiteta budućnosti u klasu OMD, dok bi zadovoljenje lokovskog (tzv. psihološkog) kriterijuma ličnosti bilo dovoljno da ove entitete smatramo moralnim delatnicima.

Mirjana Sokić
Institut za filozofiju
Filozofski fakultet Univerziteta u Beogradu

Literatura

Anderson, S. L. (2008). Asimov's "Three Laws of Robotics" and Machine Metaethics. *AI and Society* 22: 477–493.
Arkin, R. (2009). *Governing Lethal Behavior in Autonomous Robots*. Chapman & Hall/CRC.
Asimov, I. (1990). *The Bicentennial Man and Other Stories*. Gollancz.
Bentham, J. [1780/1789/1823] (2007). *An Introduction to the Principles of Moral and Legislation*. Dover, Mineola, New York.
Bringsjord, S. (2007). Ethical Robots: The Future Can Heed Us. *AI and Society* 22: 539–550.
Bryson, J. (2008). Robots Should be Slaves. Wilks, Y. (ur.), *Close Engagements with Artificial Companions*. John Benjamins Publishing Company, Amsterdam, 63–74.
Danaher, J. (2020). Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Science and Engineering Ethics* 26: 2023–2049.
Dennett, D. (1998). When HAL Kills, Who's to Blame? Computer Ethics. Stork, D. (ur.), *HAL's Legacy: 2001's Computer as Dream and Reality*, MIT Press, 351–365.
Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines* 14: 349–379.
Gunkel, D. (2012). *The Machine Question*. MIT Press, Cambridge
Gunkel, D. (2014). A Vindication of the Rights of Machines. *Philosophical Technoly* 27: 113–132.
Gunkel, D. (2018a). The Other Question: Can and Should Robots Have Rights? *Ethics and Information Technology* 20:87–99.
Gunkel, D. (2018b). *Robot Rights*. Cambridge, MA: MIT Press.
Levy, D. (2008). *Love and Sex with Robots*. Harper, London.
Levy, D. (2009). The Ethical Treatment of Artificially Conscious Robots. *International Journal of Social Robotics* 1: 209–216.
Lin, P., Bekey, G., & Abney, K. (2008). Autonomous Military Robotics: Risk, Ethics, and Design. US Department of Navy, Office of Naval Research. Dostupno online (Pristupljeno 12.15.2020): http://ethics.calpoly.edu/ONR_report.pdf
Lok, Dž. (1962). *Ogled o Ljudskom Razumu*. Kultura, Beograd.
McPherson, T. (1984). The Moral Patient. *Philosophy* 59: 171–183.
Milojević, M. (2018). *Metafizika Lica*. Institut za filozofiju, Beograd.
Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ur.), URL = https://plato.stanford.edu/archives/win2020/entries/ethics-ai (Pristupljeno 15.12.2020.)
Nadeau, J. E. (2006). Only Androids Can Be Ethical. Ford, K., & Glymour, C., (ur.), *Thinking about Android Epistemology*, MIT Press, 241–248.
Naess, A. (1973). The Shallow and the Deep, Long-Range Ecology Movements: A Summary. *Inquiry* 16: 95–100.
Narverson, J. (1965). Pacifism: A Philosophical Analysis. *Ethics* 75: 259–271.
Neely, E. L. (2014). Machines and the Moral Community. *Philosophy & Technology* 27: 97–111.
Nyholm, S., & Frank, L. E. (2017). From Sex Robots to Love Robots: Is Mutual Love with a Robot Possible? Danaher, J., & McArthur, N. (ur.), *Robot Sex: Social and Ethical Implications*. Cambridge, MA: MIT Press.
Schwitzgebel, E., & Garza, M. (2015). A Defense of the Rights of Artificial Intelligences. *Midwest Studies in Philosophy* 39: 89–119.
Singer, P. (1993). *Practical Ethics*. Cambridge University Press.

Sullins, J. (2005). Ethics and Artificial Life: From Modeling to Moral Agents. *Ethics and Information Technology* 7: 139–148.
Sullins, J. (2006). When is a Robot a Moral Agent. *International Review of Information Ethics* 6: 23–30.
Sullins, J. (2009). Telerobotic Weapons Systems and the Ethical Conduct of War. *American Philosophical Association Newsletter on Philosophy and Computers* 8. Dostupno online (Pristupljeno 15.12.2020): http://www.apaonline.org/documents/publications/v08n2_Computers.pdf
Vincent, J. (2017). Pretending to Give Robots Citizenship Helps No One. *The Verge*. Dostupno online (Pristupljeno 15.12.2020): https://www.theverge.com/2017/10/30/16552006/robot-rights-citizenship-saudiarabia-sophia

Mirjana Sokić

Moral Status of Artificial Intelligence
(*Summary*)

My main goal in this paper is to conduct a detailed analysis of the moral status of artificial intelligence. I will start by clarifying the notion of moral status, as well as the dichotomy between moral agent and moral patient, which plays a significant role in a vast number of perplexing dilemmas in applied ethics. This clarification is necessary to get a clearer view of the key issues that I intend to answer in the paper; more specifically, to the question (a) whether we can cause harm, in a morally relevant sense, to an intelligent artificial system, and (b) whether an intelligent artificial system can itself act in a way that can be assessed in moral terms.

KEYWORDS: artificial intelligence, ethical behaviorism, moral agent, moral status.