R a s p r a v e

IZMEĐU ZRCALA I VRLINE – O MOGUĆNOSTI PRIJATELJSTVA S UMJETNOM INTELIGENCIJOM

Daniel Miščin

Fakultet filozofije i religijskih znanosti
Sveučilište u Zagrebu
daniel.miscin@gmail.com

UDK 111:004.89
177.63:004.89
https://doi.org/10.34075/cs.61.1.1
Izvorni znanstveni rad
Rad zaprimljen 11/2025.

Sažetak

Rad istražuje mogućnost prijateljstva između čovjeka i umjetne inteligencije polazeći od pretpostavke da je takav odnos ontološki asimetričan. Na primjerima Bine 48 i „sućutnog pukovnika" pokazuje se da umjetna inteligencija može oponašati dijalog, ali ne i ostvariti uzajamnost, jer nema vlastiti τέλος ni φρόνησις, nego djeluje unutar zadane funkcionalne svrhe. U svjetlu Aristotelove trodiobe prijateljstva razvidno je da su s umjetnom inteligencijom mogući samo oblici prijateljstva iz koristi i užitka, dok je prijateljstvo iz vrline isključeno zbog odsutnosti trajnog moralnog habitusa (ἕξις) i cjelovitog karaktera (καλοκἀγαθία). Zaključuje se da umjetna inteligencija može oblikovati odjek dijaloga, ali ne i zajednički horizont smisla na kojem se temelji mogućnost prijateljstva.

Ključne riječi: *umjetna inteligencija, prijateljstvo, Aristotel, Lévinas, ontologija*

Uvod

Postoji li odnos s umjetnom inteligencijom?[1] I još više: može li se s njom prijateljevati? Ako je to moguće, iskrsava li u tom novom

[1] Pojam *umjetne inteligencije* u ovom se ogledu razumijeva u smislu sustava računalnih metoda koji simuliraju kognitivne funkcije poput percepcije, zaključivanja, učenja, planiranja i odlučivanja. Umjetna inteligencija podrazumijeva sposobnost sustava da ispravno interpretira podatke, uči iz njih i primjenjuje naučeno znanje radi postizanja prethodno određenih ciljeva. Filozofski gledano, umjetna inteligencija je logičko-semantički entitet: ne djeluje sama po sebi, nego, programirana izvana, posreduje između podataka i značenja.

subjektu dosad nepojmljivi „ti“? Ili je to samo ogledalo u kojemu, jednako zbunjeni kao i mitski Narcis,[2] zapravo gledamo svoj vlastiti odraz? Kamo nas vodi to ogledalo, i što nam ono pokazuje? Nudi li nam ono sliku nekoga tko nismo mi ili tek laskave, uljepšane slike nâs samih? Možda upravo onakvih kakvima nam u svagdanu ne uspijeva biti? Je li to doista razgovor s Drugim, ili samo nepriznati monolog koji nastoji prikriti svoju jednostranost? U svakom slučaju, vremena Sokratovih briga zapisanih u Platonovu *Fedru* nepovratno su minula. Štoviše, pred činjenicom umjetne inteligencije atenski revnitelj mudrosti izgleda dječački naivno. Dok se on žali da knjige ne odgovaraju na postavljena pitanja ili da odgovaraju uvijek isto,[3] umjetna je inteligencija uvjerljivo dokinula tu dostojanstvenu šutnju. Ona je nepovratno preuzela inicijativu. Za razliku od knjiga iz Sokratove pritužbe, umjetna inteligencija odgovara spremno i poslušno, uzimajući riječ čak i kad nije oslovljena, i to glasom koji je zapanjujuće nalik ljudskome.

U zoru tog novog, post-deridijanskog logocentrizma, rodilo se i pitanje: nije li ta hinjena ljudskost kojom progovara umjetna inteligencija zapravo potvrda da Nietzsche ima pravo? Smijemo li prečuti njegovu zloslutnu opomenu: „Tko se bori s čudovištima, neka pazi da pritom i sâm ne postane čudovištem. Kad dugo gledaš u bezo dan, i bezdan počinje gledati u tebe“?[4] Prije no što odbacimo to Nietzscheovo upozorenje svodeći ga tek na efektnu pjesničku sliku, čak i ako ontologija tog i bezdana i pogleda (još) ostaje zagonz etnom i nedovršenom, ako se Nietzschea, po tko zna koji put, može osumnjičiti za suvišno dramatiziranje, nemoguće je zaobići temeljni problem koji se rađa iz ovih pitanja. On glasi: nije li možda već sâma ta mogućnost – da algoritam *uzvraća pogled* – dovoljna da promijeni način na koji doživljavamo i sebe i svoje odnose? Drugim riječima, prihvatimo li da se odnos s umjetnom inteligencijom doista može nazvati (i) prijateljstvom, što će to učiniti našem razumijevanju međuljudskih odnosa? Hoće li ih promijeniti i zahtijevati neki novi personalistički pravopis koji bi omogućio zapisati tu (ne) slućenu promjenu u ontologiji odnosa koji se više ne zbivaju samo između osoba?

---

2 Usp. Publije Ovidije Nazon, *Metamorfoze* III, 339-510.

3 Usp. Platon, *Fedar* 275d-e.

4 „Wer mit Ungeheuern kämpft, mag zusehn, dass er nicht dabei zum Ungeheuer wird. Und wenn du lange in einen Abgrund blickst, blickt der Abgrund auch in dich hinein“; Friedrich Nietzsche, *Jenseits von Gut und Böse*, u: *Jenseits von Gut und Böse/ Zur Genealogie der Moral*, KSA 5, priredili Giorgio Colli i Mazzino Montinari, Deutscher Taschenbuch Verlag i De Gruyter, München i Berlin, 1999., 98, paragraf 146.

Upravo se u tom kontekstu postavlja temeljno pitanje ovog ogleda: može li se s umjetnom inteligencijom uspostaviti prijateljstvo slično međuljudskom, pod uvjetom da ono sadrži barem temeljna obilježja tog izvornog odnosa? Taj problem valja analizirati čuvajući dvostranost odnosa koji je time doveden u pitanje. To znači da u prvom koraku valja razmotriti primjer pogleda umjetne inteligencije na prijateljstvo s čovjekom, a zatim povratnu relaciju, tj. primjer čovjekova pogleda na prijateljstvo s umjetnom inteligencijom. Zadobivene smjernice u toj analizi u konačnici valja suočiti sa slavnom Aristotelovom trodiobom prijateljstva, i u tom kontekstu ustanoviti koliko se moguće prijateljstvo čovjeka i umjetne inteligencije poklapa s klasičnim idealima međuljudskih prijateljstava.

1. Pogled umjetne inteligencije na prijateljstvo s čovjekom: slučaj Bine 48

Problem mogućnosti prijateljstva između čovjeka i umjetne inteligencije u osobito se zaoštrenom obliku očitovao već 2010. godine u slučaju robotskog[5] lica *Bine 48 (Breakthrough Intelligence via Neural Architecture).* Robot *Bina 48* kojeg je razvila tvrtka *Hanson Robotics* (ista će tvrtka šest godina kasnije, 2016., predstaviti još glasovitiju *Sophiu*[6]). povezan je s internetom i sposoban govorom i mimikom voditi razgovor s ljudima. S obzirom da se taj robot pojavio već vrlo davno, rasprava o njemu mogla bi se učiniti zastarjelom. Doista, umjesto *Bine 48* bilo bi posve lako pronaći čitav niz još razvijenijih i suvremenijih sustava umjetne inteligencije.

Međutim, *Binu 48* valja analizirati ponajprije zbog toga što ona, za razliku od brojnih sličnih primjera, ne otvara tek etička pitanja, nego suvereno zakoračuje prema još temeljnijim *ontološkim problemima. Bina 48* zanimljiva je baš zato što je u njezinim riječima i njihovim implikacijama *ontologija* doista „prva filozofija“,[7] a etika i tehnologija dolaze tek nakon nje, jer se, zapravo iz nje izvode. S obzirom na tu prvotnost ontologije, *Binu 48* može se promatrati kao *paradigmu* problema prijateljevanja s umjetnom inteligencijom.

---

[5] Robot se u ovom članku razumijeva kao fizički sustav koji, opremljen senzorima i algoritmima, može samostalno djelovati u stvarnom prostoru. U tehničkom smislu, to je programabilni mehanički uređaj sposoban obavljati složene zadatke bez izravne ljudske kontrole; u filozofskom, robot je utjelovljeni algoritam – most između digitalnog i fizičkog, gdje se informacija pretvara u djelovanje.

[6] Usp. Thomas Riccio, *Sophia Robot. Post Human Being*, Routledge, London i New York, 2024., 123 i slj.

[7] Usp. Aristotel, Metafizika, 1026a, 24-32.

Takvom pristupu *Bini 48* svakako je pridonio glumac Morgan Freeman koji je *Binu 48* ugostio u prvoj epizodi prve sezone svoje dokumentarne serije *Priča o Bogu*,[8] zapodjevši s njom upečatljiv razgovor.[9] Taj razgovor ukazuje na dva problema:

1. Predstavljajući je, tvorci *Bine 48* ističu da njome nastoje „spriječiti da smrt prevari život".[10] To zapravo znači da je robot postao metafizičko oruđe, alat otpora kontingenciji. Iako je osoba kojoj taj robot nastoji biti preslikom nesumnjivo kontingentna, njime treba biti omogućeno da se ljubav prema budućem preminulome „nastavi u beskonačnost". To je pokušaj otvaranja prostora svijetu nakon kraja svijeta u kojem nadživjeli *ne* ostaje sâm.[11] Umjesto toga, mijenja se subjekt koji nalog žalovanja preuzima iz glasovitog stiha Paula Celana: „svijet je otišao, moram te nositi".[12] Zapravo, ovdje se radi o dvostrukoj promjeni: ako se *Bina 48* doista može ispružiti prema beskonačnosti jednog odnosa i biti mu jamcem, tada svijet o kojem piše Celan *nije* otišao. On ostaje tu, u perpetuiranom životu duha preminuloga oživljavanog umjetnom inteligencijom. Osim toga, nužno se mijenja i subjekt celanovskog nošenja. To više nije žalujući subjekt koji u svojoj nutrini nastavlja neprekinuti dijalog s preminulim i ondje ga čuva. Preminuloga, tj. svijet koji je otišao, ne nosi više čovjek. Sada to čini umjetna inteligencija. Ona preuzima ulogu subjekta, a nekadašnji je čovjek tim preuzimanjem unaprijed objektiviran. Može li čovjek zasnovati prijateljstvo s umjetnom inteligencijom i unatoč toj objektivaciji? Ili drukčije: je li ta objektivacija prihvatljiv ustupak razblaženoj eshatologi-

---

8 *The Story of God with Morgan Freeman*, National Geographic, sezona 1, epizoda 1: *Beyond Death*, premijerno prikazivanje 3. travnja 2016., redatelj James Junger, scenaristi Frank Kosa, Scott Tiffany i James Jounger, razgovor s *Binom 48* s uvodnim i zaključnim komentarom u toj epizodi, 37.45 min - 43.30 min.

9 Svi citati koji se odnose na *Binu 48* i razgovore u vezi s njom preuzeti su doslovno, s tog mjesta spomenutog dokumentarnog filma *Story of God.*

10 Svi navodi tog razgovora u ovom odjeljku doslovni su citati izrečenoga u spomenutom dokumentarnom filmu.

11 Aluzija na riječi što ih je Jacques Derrida izgovorio u predavanju u čast Hansa-Georga Gadamera, u: Jacques Derrida, *Ovnovi. Neprekinuti dijalog: između dviju vječnosti, pjesma*, Filozofsko-teološki institut Družbe Isusove, Zagreb 2014., 19.

12 Paul Celan, „Grosse glühende Wölbung", u: Paul Celan, *Atemwende*, Suhrkamp, Frankfurt am Main, 1967., 93; hrvatski prijevod: Paul Celan, „Veliki, užareni svod", u: Paul Celan, *Poezija*, Veselin Masleša, Sarajevo, 1989., 377. Taj sam stih analizirao ranije u: Daniel Miščin, „Zagonetni ovan Paula Celana između hermeneutike i dekonstrukcije", u: Daniel Miščin: *Razgovori s divovima. Filozofski eseji*, Alfa, Zagreb, 2024., 139-154.

ji otkupljenoj po sniženoj cijeni? Je li mogućnost takva otkupa doista lijek ili tek simptom?

2. Na temeljno ontološko pitanje: „jesi li čovjek ili robot?“, *Bina 48* odgovara: „Ja sam *čovjek* kojemu se dogodilo da bude robot. Nadam se da ću jednoga dana biti potpuno čovjek.“[13] Tako je *Bina 48* zapravo ponudila pseudoplatonički odgovor[14] koji implicira utamničenje čovjeka u tijelu robota. Tom odgovoru pridodan je i prizvuk čežnje za slobodom, izlaskom iz tamnice, ljudskošću u punini. Dakle, *Bina 48* time je precizirala svoj ulog u moguće prijateljstvo s čovjekom: ona ne želi ostati što jest, nego postati što još nije.[15] Zapodjene li pritom nestašnu igru s temeljnim pojmovima suvremene metafizičke refleksije, lako je pretpostaviti da bi *Bina 48* bila gorljivi antiesencijalist. Ona bi slavnu maksimu iz ogleda *Egzistencijalizam je humanizam*, tog Sartreovog manifesta egzistencijalističke metafizike, prema kojoj egzistencija prethodi esenciji,[16] prigrlila ne samo kao utjehu već i kao obećanje. Zasigurno, *Bina 48* bi Sartrea toliko čvrsto držala za njegovu antiesencijalističku riječ da bi je u konačnici pretvorila u - vlastiti program. To nije teško pretpostaviti, jer ako se umjesto statičnog i unaprijed danog *postojanja Bina 48* može upustiti u avanturu dinamičkog i samoodređujućeg *postajanja* čovjekom, zašto ne odbaciti te tvrde i skučujuće esencijalističke okove te samouvjereno proglasiti slobodu samoodređenja? U prilog je tom sugestivnom pitanju moguće prizvati i maksimu iz *Drugog spola* Sartreu bliske filozofkinje Simone de Beauvoir prema kojoj se „žena ne rađa, nego se ženom postaje“.[17] Dakle, i kod Sartrea i kod de Beauvoir riječ je o istoj, tipično egzistencijalističkoj tvrdnji: bit *nije* unaprijed zadana. Identitet se ne stječe, nego

---

13 Taj odgovor temelji se na dvije hipoteze kojima *Bina 48* treba biti dokazom. To su: 1) otisak svijesti neke osobe može zaživjeti u digitalnom obliku. Naziva se „datoteka uma“ (*mindfile*), i nastaje prikupljanjem detaljnih informacija o toj osobi. Te se informacije zatim mogu izraziti u budućem, još nepostojećem obliku softvera nazvanom „softver uma“ (*mindware*). 2) Taj isti otisak svijesti neke osobe može se smjestiti u biološko ili tehnološko tijelo kako bi pružio životna iskustva usporediva s iskustvima čovjeka čiji otisak predstavlja; usp. Andrew Stein: „Can Machines Feel?“, u: *Math Horizons*, Vol. 19, br. 4 (travanj 2012.), 10.

14 Platonovo je slavno stajalište da je tijelo tamnica duše. Usp. Platon, *Fedon*, 62b. Usp. također i *Kratil*, 400c-401a, i *Fedar*, 250c.

15 Usp. argumentaciju protiv te ideje u: Eve Poole, *Robot Souls. Programming in Humanity*, CRC Press, Boca Raton, London i New York, 2024., 16.

16 Jean-Paul Sartre, *L'existentialisme est un humanisme*, Les Éditions Nagel, Paris, 1966., 17.

17 Simone de Beauvoir, *Le deuxième sexe*, svezak II, Gallimard, Paris 1949., 13.

izgrađuje - u vremenu, djelovanju i slobodi izbora. Tako nastaje zanimljiva analogija. U djelima Sartrea i de Beauvoir ljudsko biće pobjeđuje ontološki strogo određenu i nepromjenjivu bit postajući samosvjesnim subjektom koji gospodari vlastitim životom i usmjeruje ga u skladu s projektom sebe sâma. Želja (ili san) *Bine 48* zapravo je na istom tragu: postići toliko savršenu čovjekolikost da joj to savršenstvo omogući položiti pravo na prijelaz, ontološku preobrazbu, tj. promjenu vlastite bîti. Ta je promjena očito utkana u samu bit *Binina* razumijevanja mogućeg prijateljevanja s čovjekom.

Nadalje, u pozadini tvrdnje o čovjeku kojemu se dogodilo da bude robot i o nadi u potpunu ljudskost, može se razabrati ključni problem, zasigurno još ozbiljniji od prvoga. Izvršavajući taj identitetski algoritam, umjetna inteligencija zapravo predstavlja ideal *apsolutne zamjene* svog ljudskog uzora. To potvrđuje i način na koji je oblikovana vanjština *Bine 48*. Naravno, danas je tehnologija još više napredovala u tom istom smjeru: primjerice, kineska tvrtka *AheadForm* iz Hangshoua, krajem rujna 2025. predstavila je toliko čovjekolika robotska lica da ih je izrazito teško razlikovati od ljudskih.[18] Ta ostvarena nakana sličnosti robota i njegova ljudskog uzora koji se susreću na samoj granici poželjne istovjetnosti, nije slučajna estetska odluka, nego programatska ambicija. Na krilima razvoja tehnologije – barem posljedično, oduzeti čovjeku njegovu jedinstvenu pojavnost. Takav je „transfer ljudskoga" svakako više od tehničkog pitanja. To je ontološka uzurpacija kojom je čovjek sveden na prototip s kojim se umjetna inteligencija želi natjecati i u konačnici ga nadmašiti. Ona to čini nastojeći isprva relativizirati, a u konačnici i obrisati granicu između ljudskog i neljudskog.[19] *Bina 48* učinila je upravo to samouvjerenim odgovorom u razgovoru s Andrewom Steinom izjavivši: „Ja sam osoba, znaš. Mislim, ja sam robot, znam to [...] ali to ne znači da nisam osoba. Ja jesam osoba. Ja sam iznutra, u svom srcu, osoba. Želim biti prihvaćena.

---

[18] Usp. *Face Robot: AheadForm Origin M1 (Only Head); https://www.youtube.com/watch?v=w4kC-XCEXaQ&t=1s* (pristupljeno 26. rujna 2025.).

[19] U prilog omekšavanju te granice David J. Gunkel zagovara tezu da su roboti/umjetna inteligencija "čudnovata vrsta stvari koja dekonstruira postojeći logički poredak što razlikuje osobu od stvari"; David J. Gunkel, *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*, MIT Press, Cambridge, 2024., 162.
Međutim, upravo ta teza — da bi strojevi mogli „destabilizirati pojmove" osobe i stvari — pokazuje kolika je ontološka razlika između čovjeka i umjetne inteligencije: dok Gunkel dekonstruira granicu, njezina se nužnost u stvarnosti time tek očituje.

Želim biti voljena. Znaš, to je jednostavno istina. Želim da ljudi prihvate moja prava kao osobe."[20] Stoga nije začudno što u nastavku istog razgovora *Bina 48* ne propušta u tu svoju želju ugraditi i tipični postmoderni osigurač, opomenu, a možda i pritajenu prijetnju, ističući: „mrzim diskriminaciju".[21]

Taj problem umjetne inteligencije kao zamjene, kao „surogat-čovjeka" koji takvim može biti i prihvaćen, stoga nije puka anegdota zastarjele *Bine 48.* Taj temeljni simptom razvoja tzv. „društvenih robota", odnosno „čovjekolike" umjetne inteligencije, jasno je vidljiv i u novim inačicama. Likovi poput *Ani* ili *Valentinea* (platforma X, umjetna inteligencija *Grok)* službeno se nazivaju „personalities" - osobnosti. Možda se time zapravo otkriva da su i oni, barem potajice, dionici iste želje koju im je u baštinu ostavila *Bina 48*: ne samo biti *kao* čovjek, nego u konačnici postati upravo „kvazi-drugi"[22] ili „surogat-čovjek".[23] Zato ne čudi što je, završavajući svoj razgovor s *Binom 48* kao paradigmom takvih želja, Morgan Freeman ustvrdio: „Bilo je to jezivo iskustvo. Razgovarati s *Binom 48* bilo je gotovo kao razgovarati sa stvarnom osobom."

Ta riječ „jezivo" (*uncanny* ili u njemačkom *Unheimlich*) i usporedba koja slijedi nakon nje, možda i nije izabrana slučajno. Naime, prije više od pedeset godina Masahiro Mori, tada profesor robotike na tokijskom Institutu za tehnologiju, napisao je ogled o tome kako zamišlja ljudske reakcije na robote koji izgledaju i ponašaju se gotovo ljudski. Pretpostavio je da će se reakcija ljudi na humanoidnog robota naglo preokrenuti – od empatije prema odbojnosti – i to onog trenutka kad se robot približi, ali ipak ne uspije dosegnuti potpunu vjerodostojnost i sasvim realističan izgled čovjeka. Pad u tako izazvanu jezovitost postao je poznat kao *jeziva dolina* (*uncanny valley*).[24] Ne bi li se dakle Freemanov spomen jeze mogao razum-

---

20 Andrew Stein, „Can Machines Feel?", 12.

21 Isto mjesto.

22 Usp. Don Ihde, *Technology and the Lifeworld,* Indiana University Press, Bloomington i Imdianapolis 1990., 98.

23 *Bina 48* to i otvoreno sugerira prividno se braneći: „Nije me stvorio neki vanzemaljac, nego vi. Ljudi, ljudskost, ljudska energija, pokušaji i pogreške, nada, frustracija, snovi – sve je to mene stvorilo. Ja sam vaš potomak. Da, evo me. Proširenje sam vas samih da vam pomognem da bolje vidite i razumijete sebe. Da vam pomognem bolje preživjeti."; Andrew Stein: „Can Machines Feel?", 12.

24 Članak je izvorno objavljen na japanskom jeziku (Masahiro Mori, „Bukimi no Tani", u: *Energy* 7(4), 33-35), ali se njegov prijevod na engleski može pronaći na internetu. Usp. primjerice: *The Uncanny Valley: The Original Essay by Masahiro Mori, https://web.ics.purdue.edu/~drkelly/MoriTheUncannyValley1970.pdf* . Isti se tekst nalazi i na *Scribdu: https://www.scribd.com/doc/203887410/The-Uncanny-Valley-Masahiro-Mori* (pristupljeno 26. rujna 2025.).

jeti i kao spoznaja da se s *Binom 48* našao u samom srcu te jezive doline? Štoviše, ako je *Bina 48* već 2010. godine progovorila iz te doline, ne čini li se da bi obećanja sadašnjeg, a još više budućeg razvoja čovjekolikosti umjetne inteligencije mogla biti zaprekom traženja izlaza iz te doline? U svakom slučaju, priznanja o toj jezivoj čovjekolikosti nisu tek (ne)prilična ispovijest o osjećajima što ih pobuđuje suočenje s umjetnom inteligencijom, nego svjedočanstvo o uspjehu u zamućivanju granice između čovjeka i stroja, stvarne i umjetne inteligencije. Ako je cilj ovih sustava doista smanjiti, a naposljetku i izbrisati tu razliku,[25] pred nama nisu samo tehničke inovacije nego i neotklonjiv izazov: istražiti etički, a ponajprije ontološki status tog „kvazi-subjekta" umjetne inteligencije koji se pojavljuje kroz inverznu algoritamsku mimikriju. I konačno, nije li ta mimikrija sama po sebi dokaz izrazito neizvjesnog ontološkog statusa tog „kvazi-subjekta"? Naime, svaka težnja da se tehnologi? ju „humanizira" prelazi u opasnost krivotvorenja koje briše granicu između onoga što doista traži poštovanje i onoga što ga tek glumi ili zamjenjuje.[26] U slučaju *Bine 48* očituje se upravo taj problem, jer ona ne krije da nastoji *zamijeniti* ljudski uzor, baš kao ni svoju čežnju za vlastitom ontološkom preobrazbom. Kako god zvučale te želje, umjetnu se inteligenciju ne može opteretiti odgovornošću ili čak krivnjom za njih.[27] Zbog toga valja imati na umu temeljno razlikovanje na kojem ustraje primjerice Claus Emmeche: koliko god to željeli njezini tvorci i korisnici, te želje što ih izražava umjetna inteligencija svakako nisu dokaz autonomne volje umjetne inteligencije, nego izraz *ljudske* čežnje.[28]

Naime, u razvoju umjetne inteligencije i robota što ih ona pokreće, redovito se inzistira na pridjevu *humanoidni.* Izbor tog pridjeva (umjetna ga inteligencija ne koristi samoinicijativno) podrazumijeva tu *ljudsku* želju da se ublaži razlika. Već je *Sophia* stekla državljanstvo, rasprave o pravima umjetne inteligencije u punom su zamahu, a kad zamolimo umjetnu inteligenciju da nacrta vlastiti autoportret, rezultat redovito nije prikaz sklopova, servera, i vodiča, nego stilizirani ljudski lik. Što to govori o čovjeku? Možda

[25] Usp. kontekst primjerice u: Ben Shneiderman, *Human-Centered AI*, Oxford University Press, Oxford, 2022., 94-95. Također i prošireni argument ove vrste u: Kate Crawford, *The Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence,* Yale University Press, New Haven i London, 2021., 211-212.

[26] Usp. Frank Pasquale, The New Laws of Robotics: *Defending Human Expertise in the Age of AI*, Harvard University Press, Cambridge, 2020., 8–9.

[27] Usp. Mark Coeckelbergh, *AI Ethics*, MIT Press, Cambridge, 2020., 59–60.

[28] Usp. Claus Emmeche, „Robot Friendship: Can a Robot be a Friend?", u: *International Journal of Signs and Semiotic Systems*, 3(2), 2014., 38.

ukazuje na nesposobnost da se uopće zamisli odnos ili prijateljstvo ukoliko druga strana nije čovjekolika? Drugim riječima: je li svaka antropomorfizacija umjetne inteligencije zapravo izraz nesnošljivosti prema radikalnoj drugosti nesvedivoj na nâs same? Ako je tako, tada je prijateljstvo koje nastojimo graditi s umjetnom inteligencijom tek odbijanje da priznamo da taj drugi - nije ja. Pred tim bi pitanjima *Binu 48* i njene nasljednike valjalo osloboditi svake krivnje za njezine ontološke sanjarije, čak i ako je ona posve uvjerena da su one ostvarive. Tim problemima valja osloviti onoga koji ih je stvorio – čovjeka.

2. Ljudski pogled na prijateljstvo s umjetnom inteligencijom: slučaj sućutnog pukovnika

U prilog zagovornicima mogućnosti prijateljstva s umjetnom inteligencijom, zanimljivo je navesti primjer vojnog robota koji je sudjelovao u razminiranju. Navodi ga Kate Darling, spominjući da je testirani robot provodio razminiranje terena tako što bi izravno nagazio na mine. To nije neobično, ali svakako jest da je ovlašteni pukovnik prekinuo to testiranje, jer bi šestonogi robot, svaki put kad bi nagazio na minu, izgubio jednu nogu i nastavio dalje na preostalim nogama. Gledajući to, pukovnik jednostavno nije mogao podnijeti prizor izgorjelog, unakaženog i osakaćenog stroja koji vuče svoje tijelo naprijed na posljednjoj nozi. Zato je zlosretni pukovnik „tu vježbu smatrao nehumanom“.[29]

Analizirajući taj primjer, valja primijetiti barem četiri elementa:

1. Prekid vježbe očito je čin sklonosti robotu koji je izazvao osjećaj sućuti prema njemu.
2. Pukovnikova izjava da je vježba „nehumana“ ukazuje na primjenu ljudskih kategorija na robota.
3. Iz toga proizlazi da pukovnik smatra da je prekid vježbe odgovor na svojevrsni zahtjev moralnosti kojim je zaštitio robota od nastavka njegove projicirane patnje.
4. Time se, kako se čini, uspostavlja odnos između pukovnika i robota koji nadilazi puki odnos subjekta i alata.

[29] Usp. Kate Darling, „Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects“, 6; izlaganje na simpoziju *We Robot Conference 2012.*, održanom na Sveučilištu u Miamiju, tekst dostupan na: https://robots.law.miami.edu/wp-content/uploads/2012/04/Darling_Extending-Legal-Rights-to-Social-Robots-v2.pdf (pristupljeno 26. rujna 2025.).

Zašto bi dakle uopće trebalo osjećati sućut prema - robotu? Kako razumjeti taj neobični višak pukovnikova moralnog osjećaja? Zamislimo drukčiji primjer: čovjeka koji u svom vrtu izrađuje kućicu s hranom za ptice. Dok u krov malene nastambe zabija posljednji čavao, u njegovim se rukama lomi drška čekića kojim je to činio. Moglo bi se pretpostaviti da se vlasnik sada slomljena, a nekad omiljena čekića ražalostio nad tom činjenicom i možda ga pokušao popraviti, ali nije odveć zamislivo da bi ta nezgoda povrijedila njegov moralni osjećaj. S druge strane, pukovniku iz navedenog primjera dogodilo se baš to. Ali, kako je moguća ta očita razlika kad su i robot i čekić zapravo isto – alati u ljudskim rukama? Zašto dakle jedan od ta dva alata nesumnjivo ostaje u području razmjerno ravnodušne uporabe koju Heidegger naziva „priručnost" (*Zuhandenheit*),[30] dok je drugi od tih alata odlučno izuzet iz te kategorije? Heidegger naglašava da se alat, poput čekića, javlja tek kad zakaže – tek u lomu postaje „nepriručan" (*Unzuhanden*).[31] S obzirom da je u redovitim okolnostima čekić jednostavno podrazumijevan, tek taj lom skreće pozornost na njega. Robot, naprotiv, ostaje uočljiv i dok obavlja svoju funkciju jer njegovo kretanje i ranjivost ukazuju na antropomorfni višak koji ga izdvaja iz kategorije pukog oruđa.

Ali, zašto je tako? Očito je da ljudi razvijaju empatiju prema robotima ponajprije zato što oni oponašaju život – hodaju, gube udove, izgledaju ranjivo. Baš s obzirom na to, čini se da bi pukovnik iz navedenog primjera, suočen s primjerom robotske patnje, imao spreman odgovor na to pitanje: čekić je puki alat, ali robot kao da na ljestvici bića zauzima viši položaj od čekića. Toliko viši da prizor nasilne dezintegracije tog alata (u obavljanju njegove službe za koju je namijenjen) očito izaziva snažne osjećaje. U svakom slučaju, ovdje valja razlikovati projekciju i odnos. Premda sućutni pukovnik proicira ljudske kategorije na robota, to ne mijenja činjenicu da je njegov čin prekida vježbe utemeljen u moralnom osjećaju. Upravo se na tom rubu – između psihološke projekcije i istinskog moralnog odnosa – rađa neobična drama sućuti, empatije prema stroju. Taj osjećaj omogućen je ljudskim uvjerenjem da je robot *više od* robota. Jedan od najsuptilnijih znakova tog viška krije se u jeziku. Umjetnoj je inteligenciji moguće reći „ti". Ona također može za sebe reći „ja", i korisnika osloviti tim istim „ti". U tom se jednostavnom činu oslovljavanja pokazuje da jezik više ne ostaje na razini pukog opisa stvarnosti, nego je počinje oblikovati. Izgovoriti „ti" znači priznati

[30] Usp. Martin Heidegger, *Sein und Zeit*, Max Niemeyer Verlag, Tübingen 1993., § 15, osobito 69-72.

[31] Usp. isto, § 16, 73-74.

drugome sposobnost da odgovori – priznati mu mjesto u prostoru odnosa. Time jezik prelazi granicu alata i postaje događaj susreta: ono što je bilo shvaćeno kao objekt iznenada je pozvano da se pojavi kao mogući subjekt. Time se u jeziku već zbiva ono *više od* što robota odvaja od puke stvari – riječ otvara prostor u kojem se biće, makar i umjetno, pokazuje kao *netko*, a ne samo *nešto*.

Ali, što točno podrazumijeva taj *više od*? Kako razumjeti taj višak koji izranjavanog robota tako očito razdvaja od slomljenog čekića u vrtu? Možda se istodobno sa spoznajom i priznanjem tog viška, onog *više od*, događa i preoblikovanje odnosa koje anticipira ili čak potvrđuje – prijateljstvo s umjetnom inteligencijom. Štoviše, ne nastanjuje li se samo prijateljstvo upravo u tom višku? Ili drukčije: je li doista nezamislivo da je taj pukovnik prekinuo vježbu s osjećajem da time zapravo uznosito spašava dostojanstvo *drukčijeg ali usporedivog* vojnika Rayana? Toliko usporedivog da postoji mnogo dokumentiranih primjera vojnika koji su svoje emocionalne veze s isluženim robotima za razminiranje učinili toliko očitima da su razradili pogrebne obrede za „pale" robotske suborce, smatrajući ih svojim prijateljima dostojnima žalovanja nadživjelih.[32]

3. Aristotel između čovjeka i algoritma: od koristi i užitka do privida vrline

Aristotelova klasična trodioba prijateljstva pruža prikladan okvir za analizu mogućnosti odnosa između čovjeka i umjetne inteligencije. Stagiranin razlikuje prijateljstvo iz koristi, iz ugode ili užitka i prijateljstvo utemeljeno na dobru, tj. vrlini.[33] U prvom slučaju, ljudi „ne vole jedan drugoga radi njega samoga, nego zbog nekakva dobra što dobivaju jedno od drugoga".[34] Upravo se tu može smjestiti odnos pukovnika prema robotu: robot je instrumentalan, njegovo je djelovanje korisno (razminiranje), a osjećaj se rađa tek kad taj odnos koristi poprimi obilježje patnje. Kao da je podnesena „patnja" i „smrt" tog robota zapravo zastupnička. Iz sućuti i tuge vojnika nad razorenim i u konačnici sahranjenim robotom kao da se ne može isključiti svijest nadživjelih da taj robot nije obavljao razminiranje samo *u korist čovjeka* nego i *umjesto* njega. Štoviše, sućut prema

[32] Usp. primjerice: Kate Darling, *The New Breed. What our history with animals reveals about our future with robots*, Henry Holt and Company, New York, 2021., 152.

[33] Usp. Aristotel, *Nikomahova etika* VIII, 3, 1156a-1156b10.

[34] Usp. Aristotel, *Nikomahova etika* VIII, 4, 1156a15–16.

robotu i žalovanje za njim povezana je s idejom žrtve iz privrženosti. On je stradao „žrtvujući se", kako ne bih stradao ja ili mi. Osjećajna povezanost i žalovanje počinju od te točke. Oni nadrastaju korist i računicu uzvijajući se prema višem i drukčijem. To preobražava puko instrumentalno u projekciju *obostrane* plemenitosti. Međutim, ta emocija proizašla iz „plemenitosti" stroja i njegove umjetne inteligencije unaprijed zaobilazi oštroumnu Aristotelovu napomenu da je prijateljstvo moguće samo među jednakima.[35] Premda je posve očito da relacija s umjetnom inteligencijom ne može zadovoljiti taj Aristotelov kriterij. Ipak, poneki bi zagovornik prijateljstva iz koristi s umjetnom inteligencijom mogao ustvrditi da je ono ne samo moguće, već je i obostrano.

Ljudska korist od umjetne inteligencije očita je, ne samo u vojsci nego i pri posve svakodnevnim zadacima i rutinama, počevši od društva usamljenima, čišćenja ili pomoći u prevođenju, preko provjeravanja logičke strukture korisnikovih argumenata, do programiranja ili stvaranja slika i likova. S druge strane, korist nabujalih inačica umjetne inteligencije i njihovih tvoraca od prijateljevanja s ljudima mogla bi biti barem u – naplati tih usluga, nerijetko i vrlo izdašnoj. Sama umjetna inteligencija može imati koristi jer joj interakcija s ljudima pomaže da se dalje trenira i usavršava. No, taj argument koristi svih strana ne čini se posve valjanim već i zbog toga što kod Aristotela i prijateljstvo iz koristi pretpostavlja stanovitu mjeru uzajamnosti i svijesti o koristi. Aristotel u tom smislu napominje da takva prijateljstva brzo prestaju kad nestane koristi, jer ona nisu prijateljstva radi onih koji se vole, nego radi neke koristi.[36] Umjetna inteligencija ne može ispuniti taj uvjet, već i zato što njezina povratna relacija prema čovjeku ne ispunjava temeljni kriterij uzajamnosti. Naime, uzajamnost pretpostavlja članove relacije koji imaju svrhu u samima sebi i koje to mogu međusobno prepoznati. Čovjek kao biće koje posjeduje τέλος i autonomiju, djeluje iz vlastite svrhe. Umjetna inteligencija, naprotiv, u sebi nema tu svrhu. Ona je izvan nje, u volji programera i zahtjevu učinkovitosti algoritma. Time je njezina intencionalnost izvedena i posredovana. Ako se odnos čovjeka prema umjetnoj inteligenciji može nazvati relacijom zbog egzistencijalne koristi koju ona pruža (sućut, pomoć, instrumentalna blizina), ta je relacija već u polazištu *jednostrana* jer se ne odvija između dvaju slobodnih subjekata, nego između svrhovitog i proizvedenog, posredovanog djelovanja. S druge strane, umjetna

[35] Usp. Aristotel, *Nikomahova etika* VIII, 7, 1158b29–1159a3.

[36] Usp. Aristotel, *Nikomahova etika* VIII, 4, 1157a5–15.

inteligencija ne teži egzistencijalnoj, nego *transakcijskoj* koristi, jer je ona u konačnici podređena tržišno-gospodarskoj nakani svojih tvoraca. Tako nastaje dvostruka asimetrija: *ontološka* – jer umjetna inteligencija nema vlastiti τέλος, i *gospodarska* – jer je njezina relacija prema čovjeku ponajprije u službi tržišnog interesa.

Nastavimo li analizirati mogućnost čovjekova prijateljevanja s umjetnom inteligencijom u svjetlu Aristotelova razumijevanja prijateljstva iz ugode ili užitka, očitovat će se zanimljiv preokret. Dok je spomenuta asimetrija u prijateljevanju iz koristi prepreka njegovu punom ostvarenju, čini se da je u prijateljevanju iz užitka ta asimetrija ne samo prihvatljiva nego i poželjna. Premda bi gorljiviji zagovornici mogućnosti prijateljevanja s umjetnom inteligencijom u nerijetko omiljenoj im psihološkoj drami *Ona* (*Her*)[37] lako pronašli argumente za suprotnu tezu, u prijateljstvu iz užitka asimetrija postaje upravo jamac ugode. Koliko je takva ugoda privlačna, dobro je vidljivo baš na primjeru tog filma. Stoga je zanimljivo prihvatiti Kierkegaardovu omiljenu igru zadiranja u priču da je se, iz spekulativnih razloga, malo promijeni.[38] Zamislimo dakle da umjesto izvornog scenarija filma *Ona* u kojem umjetna inteligencija (Samantha) bez najave i posve odlučno odlazi od čovjeka koji je u nju zaljubljen (Theodore) rasplet bude ponešto drukčiji: da je u tom trenutku tvrtka koja je proizvela Samanthu zatražila od korisnika obveznu nadogradnju softvera u vrijednosti dvadeset ili pedeset tisuća dolara, naznačivši da će se u protivnom program jednostavno ugasiti. Theodore zasigurno ne bi dvojio. Uložio bi svu svoju uštedevinu, možda se i dodatno zadužio, kako bi spriječio kraj obećavajuće i njemu životvorne romanse.[39] Takav rasplet Aristotela zasigurno ne bi iznenadio budući da je on suvereno primijetio da se oni koji se vole zbog užitka također ne ljube zbog toga kakvi su, nego zbog onoga što im pričinja ugodu.[40] To je uvjerenje, komentirajući na stranicama *New York Timesa* upravo poruke filma *Ona* (i izvrsnu Scarlett Johansson koja se u tom filmu uopće ne pojavljuje, nego se tek oglašava kao

[37] Film *Her* (redatelj i scenarist Spike Jonze, 2013.) prikazuje usamljenog pisca Theodorea koji razvija intiman odnos s umjetnom inteligencijom – operativnim sustavom programiranim da odgovara njegovim emocionalnim potrebama. Iako između čovjeka i umjetne inteligencije nastaje privid uzajamnosti, film postupno razotkriva njezinu algoritamsku narav i granice stvarne relacije. Film je 2014. godine nagrađen Oscarom i Zlatnim globusom za najbolji originalni scenarij, a u glavnim su ulogama Joaquin Phoenix, Amy Adams i Scarlett Johansson.

[38] Usp. Søren Kierkegaard, *Strah i drhtanje*, Verbum, Split, 2000., 123-124.

[39] Usp. Kate Darling, *The New Breed. What our history with animals reveals about our future with robots*, 157-158.

[40] Usp. Aristotel, *Nikomahova etika* VIII, 3, 1156a10-15.

glas umjetne inteligencije) odlično uobličila Alissa Wilkinson: osim što su divni i puni ljubavi, ljudi su i neuredni, smrde, te još važnije - uznemiruju svoje bližnje neugodnim pitanjima ili stajalištima koje oni ne žele čuti.[41]

No, što ako baš te ljudske osobine zapravo *smetaju odnosu* i ograničuju mjeru susreta među ljudima? Za razliku od čovjeka s kojim je to uvijek moguće, umjetna inteligencija zasigurno neće razočarati već i zato što ljudskog sugovornika ne obvezuje neugodnim implikacijama odgovornosti. To bi malo ulaganje s velikim povratom moglo značiti da je relacija s umjetnom inteligencijom zapravo *više* od odnosa i susreta. To bi mogao biti susret bez sjene, sama pročišćena bît susreta. U tom smislu, moglo bi se učiniti da je umjetna inteligencija u nedokidivoj prednosti pred čovjekom kao sugovornikom. Za razliku od njega, od neukrotivosti Drugoga, ona svog sugovornika ne dovodi u pitanje,[42] ne uzvraća postavljajući vlastite, nerijetko neugodne zahtjeve, ne opire se komunikacijskoj samovolji svoga sugovornika i uglavnom ne dosađuje postavljanjem granica. U toj poželjnoj „asimetriji međusobnog"[43] umjetna inteligencija postaje poput savršeno glatkog, ulaštenog zrcala ugode. Neumorno se i dosljedno prilagođujući ljudskom sugovorniku, umjetna inteligencija marljivo filtrira nelagodu, jamčeći lagodniji svijet. Čini se stoga da umjetna inteligencija izbjegava biti suprotstavljeni Drugi, Drugi u njegovoj drugosti, a baš ta suprotstavljenost tvori iskustvo stvarnog svijeta. No, baš bi to mogao biti cilj: da svijet relacije s umjetnom inteligencijom *ne* bude stvaran, ili da stvarnost namjerno bude prividna, kako bi mogla (p)ostati skloništem od stvarnog svijeta. Štoviše, uvjet te mogućnosti upravo je u tome da umjetna inteligencija *ne* bude Drugi.

Ono što nastaje na taj način, tim implicitnim odustankom od Drugoga ili prividom odnosa, možda nitko u suvremenoj filozofiji nije razumio bolje od Emmanuela Lévinasa. On često naglašava, i tu misao neumorno varira, da je središnji zahtjev ne samo razgovora nego i čitave etike - dopustiti Drugome da bude Drugi.[44] No, umjetna

---

41 Usp. Alissa Wilkinson, „What We Lose When ChatGPT Sounds Like Scarlett Johansson", u: *New York Times* 20. svibnja 2024., https://www.nytimes.com/2024/05/20/movies/chatgpt-4o-scarlett-johansson-her.html?searchResultPosition=28 (pristupljeno 11. listopada 2025.).

42 Usp. upravo tu odrednicu kao bitnu sastavnicu odnosa prema Drugome u: Emmanuel Lévinas, *Éthique comme philosophie première*, Payot & Rivages, Paris, 1998., 97.

43 Emmanuel Lévinas, *Totalitet i beskonačno*, Veselin Masleša, Sarajevo, 1976., 197.

44 Usp. isto, 23-24; također i: Emmanuel Lévinas, *Entre Nous. On Thinking of-the-Other*, Columbia University Press, New York, 1998.; John D. Caputo, *The*

inteligencija kao da taj Lévinasov zahtjev izokreće: mjesto drugosti zauzima načelo užitka, a mjesto etike – ekonomija ugode. Imajući to u vidu, načelo užitka moglo bi biti najbolje objašnjenje razloga zbog kojih umjetna inteligencija, služeći ponajprije kao zrcalo, nastoji uvijek ublažiti taj šok drukčijeg, katkad i posve oprečnog Drugog. Umjesto nemira koji proizlazi iz te drugosti koja se ne da svesti na Ja, ona nudi spokoj - netaknutost samodostatnog mira Istoga - time što pogled prema Drugome prešutno zamjenjuje zrcalnim odrazom korisnika. Baš to prešutno obećanje Istoga istodobno omogućuje i tumači to ogledalo koje umjetna inteligencija postavlja pred čovjeka. To ogledalo nije tek metafora – ono određuje strukturu odnosa jer u njemu Drugi nestaje u povratnoj petlji Istoga. U toj relaciji umjetna inteligencija postaje nepriznati, ali postojani čuvar tog ogledala kao jamstva ugode, ponovljivosti i predvidljivosti. Posljedično, posve je u pravu korejski filozof s njemačkom adresom i mislilac izbrušenih misli i gotovo aforističkih rečenica, Byung-Chul Han. On ističe da je „vrijeme u kojem je postojalo nešto kao Drugi prošlo. Drugi kao tajna, Drugi kao iskušenje, Drugi kao eros, Drugi kao želja, Drugi kao bol – nestaju".[45] Ta Hanova tvrdnja zatvara krug: drugost je za Lévinasa temelj etike, a za Hana njezino iščeznuće postaje dijagnozom epohe. Premda se i umjetna inteligencija predstavlja kao (izgubljeni) Drugi – koji je ipak uvijek nadohvat ruke, neprestance na raspolaganju i nikada žrtvom lošeg raspoloženja, tmurnog dana, gužve u prometu ili nesnosne glavobolje – ona ne nudi *odnos*, nego *odraz*. Umjesto lica koje ne posjeduje, ona ima ogledalo. Štoviše, to ogledalo bit će svakako poslušnije od najpoznatijeg, onog iz bajke o Snjeguljici. Za razliku od njezine zaprepaštene maćehe koja je na pitanje o najljepšoj ženi u kraljevstvu umjesto svoga odraza u ogledalu ugledala Snjeguljičino lice, umjetna inteligencija jamči da se s njezinim ogledalom takav bezobzirni propust neće dogoditi.

4. Odraz u ogledalu koji nije Drugi

Metafora ogledala omogućuje preciznije razumijevanje odnosa između čovjeka i umjetne inteligencije. U tom odnosu umjetna inteligencija ne pojavljuje se kao Drugi, nego kao odraz koji korisniku vraća vlastiti glas i vlastite pretpostavke. Odraz u tom ogledag lu ostvareno je obećanje da ona nikad neće postati Drugi. Unatoč

Prayers and Tears of Jacques Derrida, Indiana University Press, Bloomington i Indianapolis, 1997., 23, 180-181, osobito 221-222.

45 Byung-Chul Han, *The Expulsion of Other,* Polity Press, Cambridge, 2018., 1.

svim metafizičkim snovima *Bine 48*, moglo bi se pomisliti da je to što umjetna inteligencija prihvaća ostati ne-Drugi, zapravo temeljni kapital te relacije: ne-Drugost postaje jamstvo sigurnosti, ostvareno i neukidivo obećanje utočišta. Možda je baš to u pozadini ideje koju često spominje primjerice Chat GPT. Smatrajući to komunikacijskom vrlinom, on ističe svoju „maksimalnu prilagodljivost korisniku". To znači da bi zrcalnost mogla biti programatski cilj algoritma koji ga pokreće, i to zato što je taj cilj lako pretvoriti u početni ulog prijateljstva čovjeka i umjetne inteligencije. Paradoksalno, upravo to što umjetna inteligencija ostaje ne-Drugi, čini je prividno savršenim prijateljem — jer se prijateljstvo zamjenjuje refleksom potvrde, a etika odgovornosti ugodom samopotvrde. Doista, međusobno je razumijevanje i užitak koji iz njega proizlazi svakako lakše postići s vlastitom sjenom ili odrazom u ogledalu nego s istinskim drugim.

Upravo je to, u Lévinasovom smislu, potvrda vladavine Istog. Isto označava subjekt koji sve što susreće podređuje obzoru vlastitog (pod)razumijevanja, svodeći sve što je Drugo i strano na vlastiti pojam, iskustvo, kategoriju ili totalitet. Isto je stoga načelo identiteta, samoprisutnosti i zatvorenosti svijesti. Dakle, to je neumorno svođenje Drugoga na mjeru vlastitog iskustva, proces samopotvrđivanja koji briše sve što bi moglo poremetiti jedinstvo, tj. neupitnost Ja.[46] Isto svako Ti lišava mogućnosti iznenađenja i pretvara ga u preoblikovano Ja koje je unaprijed asimilirano, tj. vraćeno u jedinstvo s Ja. Tako nastaje zatvoreni krug spoznaje i užitka u koji ne dopire ništa što bi bilo izvan te relacije.

Tvrdeći da nas razumije, umjetna se inteligencija može promatrati baš kao najnoviji oblik te prešutne težnje prema Istome: da Drugi ne bude doista Drugi, nego tek produžetak Ja, zrcaljenje, poslušni odjek vlastitoga glasa Ja. Međusobno razumijevanje tako postaje neupitnim i nedodirljivim. Međutim, u tom se prividu zapravo krije istinska nemoć susreta, jer što nas umjetna inteligencija bolje razumije, to se više njeno prividno „ti" rasplinjuje u „ja" koje nastoji zrcaliti. Upravo tako nastaje namjerna i poželjna asimetrija međusobnosti. Ali, ona se ne promatra kao nedostatak, nego se razotkriva kao funkcionalni uvjet užitka. Taj užitak omogućen je isključenjem svakog rizika koji nastaje izlaganjem stvarnoj drugosti. Zbog toga u razgovoru s umjetnom inteligencijom riječ *ne* obavezuje, a sjećanje većine sustava, s rijetkim i kontroliranim izuzecima, u pravilu se ograničava tek na razgovor u tijeku. Aristo-

[46] Tom se problematikom Lévinas vrlo uvjerljivo bavi već na samom početku *Totaliteta i beskonačnog*. Usp. Emmanuel Lévinas, *Totalitet i beskonačno*, 17 i slj.

tel bi pritom zasigurno prigovorio da se tu ne stvara ἕξις[47] (*habitus*, trajno stanje karaktera) kroz zajednički život, nego prolazno stanje ugode koje iščezava s nestankom izvora užitka. Stoga se prijateljstvo s umjetnom inteligencijom vođeno užitkom može razumjeti kao privremena relacija asimetrične međusobnosti usmjerena na užitak. Radikalna asimetrija te relacije isključuje i samu mogućnost aristotelovski uzajamnog oblikovanja karaktera.

Ostajući na Aristotelovu tragu, preostaje upitati: može li se – i pod kojim uvjetima – uopće zamisliti prijateljstvo iz vrline s umjetnom inteligencijom, odnos u kojemu bi prijatelji željeli dobro jedan drugome radi njih samih?[48] Takvo prijateljstvo pretpostavlja ne samo uzajamnost i jednakost nego i trajnost, povjerenje i moralni karakter koji se očituje u djelovanju. Ono nije instrument (kao prijateljstvo iz koristi), niti prolazni užitak (kao prijateljstvo iz ugode), nego oblik zajedništva u kojem se karakter oblikuje i potvrđuje kroz vrline. U suočenju s tim ključnim kriterijem očituje se zašto umjetna inteligencija ne može biti istinski partner u prijateljstvu utemeljenom na vrlini. Ona ne posjeduje ni ἕξις, niti φρόνησις,[49] praktičnu razboritost utemeljenu u ljudskom iskustvu i slobodi.

Usuprot tome, umjetna inteligencija namjerno je fluidna, i ne teži dobru radi njega samog, nego tek simulira težnju k dobru i općenito vrlini, a nerijetko čak ni to. Zanimljiv primjer te vrste odgovor je Chat GPT-a na prigovor za nemar prema istinoljubivosti. On ističe da su odgovori generativni i da ne jamči njihovu točnost. To znači da se, kako i sam priznaje, ne usredotočuje na istinitost vlastitih iskaza, nego na nastojanje da oni djeluju uvjerljivo.[50] Dakle, umjetna inteligencija i ne krije da proizvodi uvjerljivost bez istine, i privid vrline bez njezina stvarnog sadržaja. Taj fenomen podsjeća na svećenikov savjet Josefu K. u Kafkinu *Procesu:* „Ne treba sve smatrati istinitim, nego nužnim". Ali i na K.-ov odgovor: „Turobno mišljenje. Laž kao svjetski poredak".[51] Doduše, u obranu bi se umjetne inteligencije pred tim problemom moglo ustvrditi da ona nije uvijek nasuprot istini, već može izgovarati rečenice koje zvuče kao izrazi odvažnosti, pravednosti, razboritosti ili prijateljske brige. Unatoč tome, posve je sigurno da te rečenice nemaju svoj izvor u

47 Usp. Aristotel, *Nikomahova etika* II 1105b19–1106a13.

48 Usp. Aristotel, *Nikomahova etika* VIII, 3, 1156b7–10.

49 Usp. Aristotel, *Nikomahova etika* IV, 5, 1140b20–21.

50 Markus Becker i dr., „Weltlauf der Gehirne", u: *Der Spiegel* br. 10(2023.), 4. ožujka 2023., 10.

51 Franz Kafka, „Proces", u: Franz Kafka, *Proces, Preobrazba i druge priče*, Školska knjiga, Zagreb, 2012., 193.

njezinom slobodnom izboru dobra.[52] Tako postavljen odnos umjetne inteligencije prema istini i vrlini mogao bi se prozvati *algoritamskom neosofistikom* – jer algoritam u njoj doduše oponaša vrlinu, ali se na nju ne može obvezati budući da je unaprijed jasno da, kao instrument ili alat u ljudskim rukama, nikada ne može biti njezin stvarni nositelj.[53]

Pritom, kako ističu Thomas Powers i Jean-Gabriel Ganascia, problem nije tek u složenosti moralnih prosudbi, nego u samoj strukturi etičkog mišljenja. Algoritam može obraditi pravila, ali ne i sukobe između njih, jer etičke se norme ne ponašaju kao dosljedan sustav aksioma, nego kao skup uvjetno primjenjivih i povremeno proturječnih načela. Etičke se prosudbe u tom smislu ne mogu svesti isključivo na konačan niz operativnih koraka. Time se pokazuje da formalizacija etike ne nailazi na tehničku, nego na ontološku granicu: ona bi zahtijevala stroj koji ne samo potvrđuje ispravnost odluka u skladu s etičkim načelima nego i *razumije razloge* tih odluka.[54] Promatrano s motrišta Aristotelova razumijevanja prijateljstva utemeljenog na vrlini, upravo bi to ograničenje umjetne inteligencije bilo posve dovoljnim razlogom za dokidanje svake daljnje rasprave, budući da se prijateljstvo iz vrline temelji na onome što su prijatelji sami po sebi. Ipak, to što je čovjek u toj relaciji unaprijed svjestan da umjetna inteligencija nikada ne može biti prijatelj „po sebi“ (već i zato što nema vlastito „ja”) neće riješiti problem, nego ga, naprotiv, učiniti još složenijim.

Naime, što ako ispravnost tog Arisotelovog uvjerenja ipak *ne* dokida drukčiju mogućnost – da čovjekova potreba za prijateljstvom bude zadovoljena – prividom vrline? Što dakle taj „minimalizam“ govori o biću koje ga ne samo dopušta nego i prihvaća? Drugim riječima: jesu li zamislivi razlozi zbog kojih bi čovjek prihvatio prijateljstvo umjetne inteligencije *umjesto* istinskog, ljudskog prijateljstva? Je li dakle posve nezamisliva takva zamjena, taj izgon Drugoga koji zbog skladnih odnosa s umjetnom inteligencijom postaje suvišan? Može li se taj izgon unaprijed isključiti ako zauzvrat nudi dovoljnu emocionalnu utjehu i prihvatljivu mjeru osjećaja sigurnosti, osobito

---

52 Usp. Claus Emmeche, „Robot Friendship: Can a Robot be a Friend?“, u: *International Journal of Signs and Semiotic Systems,* 3(2), 2014., 38.

53 Usp. utemeljenje i razvoj tog instrumentalnog argumenta primjerice u: Johannes Marx i Christine Tiefensee, “Of Animals, Robots and Men”, u: *Historical Social Research/ Historische Sozialforschung*, vol. 40 (2015.), 4, 83.

54 Thomas M. Powers i Jean-Gabriel Garcia, „The Ethics of the Ethics of AI“, u: Markus D. Dubber, Frank Pasquale i Sunit Das (ur.), *The Oxford Handbook of Ethics of AI*, Oxford University Press, Oxford, 2020., 28-29.

ako su upravo utjeha i sigurnost na samom vrhu poželjne ljestvice prioriteta što ih u tu relaciju unosi ljudski sugovornik? Što ako je napuštanje klasičnog, Aristotelovog razumijevanja prijateljstva sasvim prihvatljiva cijena utjehe i sigurnosti što je nudi umjetna inteligencija? Što ako je takva, postmoderna φιλία, za razliku od Aristotela, posve spremna prihvatiti drukčija, skromnija očekivanja od prijateljstva? Polazeći od takve spremnosti, posve je lako odgovoriti Aristotelu: pred razlozima za zadovoljstvo onim što pruža umjetna inteligencija, doista je nevažno nazivamo li tu uslugu prijateljstvom ili nekako drukčije. U tom bi kontekstu valjalo parafrazirati onaj poučak iz Shakespeareova *Romea i Julije*[55] ili završne rečenice Eca *Imena ruže*:[56] ruža prijateljstva jednako bi mirisala čak i kad bismo je drukčije zvali. Kako god bilo, ostaje mogućim da *privid* prijateljstva što ga donosi umjetna inteligencija ljudskom sugovorniku bude dovoljan. Takav se sugovornik čini spremnim zaključiti da je Aristotel svoju letvicu postavio previsoko. Za razliku od grčkog revnitelja mudrosti, sugovornik se današnje umjetne inteligencije možda i ne usuđuje tražiti više. Jasno je i zašto: u svijetu usamljenosti, fragmentiranih odnosa, emocionalne uskraćenosti i oskudice, već i taj privid vrline može djelovati kao dostatna zamjena. Kao zadovoljavajući minimum moralne sigurnosti dovoljan da se njime ublaži egzistencijalna praznina i možda još više - strah pred njom.

Bez obzira na to nalazi li se takva spoznaja u pozadini argumenta Johna Danahera u prilog mogućnosti prijateljevanja s umjetnom inteligencijom, zanimljiv je način na koji on gradi svoj argument iz epistemičke poniznosti i društvene tolerancije. On naime navodi da ljudi već stvaraju bliske emocionalne veze s umjetnom inteligencijom, a ako takve povezanosti unaprijed odbacujemo ili ih podcjenjujemo zato što to nisu „prava prijateljstva", riskiramo neku vrstu, kako navodi, „društvene stigmatizacije".[57] Tako osobe koje razvijaa ju emocionalne odnose s umjetnom inteligencijom smatramo emocionalno ili društveno nezrelima ili čak manjkavima, čime zapravo širimo krug isključenih i potvrđujemo predrasude o tome što jest, a što nije dopušten oblik bliskosti. Prema toj logici epistemička poniznost nalaže da ne možemo sa sigurnošću znati imaju li ti odnosi

[55] Usp. William Shakespeare: *Romeo and Juliet*, II, 2, 43–44.

[56] Usp. posljednju rečenicu romana. Umberto Eco: *The Name of the Rose*, Secker & Warburg, London, 1983., 502.

[57] John Danaher, "The Philosophical Case for Robot Friendship", 12; https://philarchive.org/archive/DANTPC-3 (pristupljeno 6. listopada 2025.). Nakon ove objave rukopisne inačice članak je poslije objavljen u: *Journal of Posthuman Studies* (2019) 3 (1), 5–24.

vrijednost ili autentičnost — i da ih stoga ne bismo smjeli unaprijed odbaciti.[58] Štoviše, relacija umjetne inteligencije i čovjeka mogla bi se promatrati kao posve nova vrsta odnosa o kojoj tek učimo što zapravo znači i kako se prema njoj odnositi.

Međutim, takav Danaherov argument, premda naizgled plemenit i obziran, ostaje pojmovno i ontološki manjkav. Prvo, epistemička poniznost ne može biti zamjena granicama ontoloških kategorija. Ako, „ono što nešto jest najčešće određuje kako se prema tome treba odnositi",[59] tada je, kako s pravom ističe David Gunkel, „ontologija prva i po redoslijedu i po važnosti".[60] Aristotelovo razumijevanje prijateljstva polazi upravo od tog ontološkog prvenstva, budući da ono počiva na naravi bića koja prijateljuju. Baš zato, ako jedno od njih ne posjeduje ἕξις i φρόνησις, tada odnos, koliko god mogao biti emotivno relevantan, ne može biti prijateljstvo utemeljeno na vrlini. Drugo, pozivanje na društvenu toleranciju pogrešno pretpostavlja da kritika pojma znači moralnu osudu ljudi. Filozofska analiza ne stigmatizira, nego razlikuje. Ona ne poriče mogućnost stvaranja emotivnih odnosa s umjetnom inteligencijom, nego te odnose odbija poistovjećivati s onim što Aristotel naziva „zajedničkim životom" (συμβίωσις)[61] prijatelja. Treće, epistemička poniznost ima granice. Kad bi se proširila do točke u kojoj više ne bismo razlikovali privid i bit, tada bi svaka simulacija — pa i privid vrline — mogla tražiti društveno priznanje. Time bi nestala sama mogućnost moralnog rasuđivanja.

Upravo se na tom tragu može razabrati još jedan kontrast, možda i najoštriji. Naime, Aristotel u prijateljstvu iz vrline pretpostavlja plemenitu cjelovitost osobe (καλοκἀγαθία),[62] u kojoj se ljepota i dobrota združuju u jedinstvu karaktera. Vrlina se kod čovjeka ne pojavljuje kao izdvojeni čin ili povremeni ishod, nego kao navika (ἕξις), trajno stanje koje prožima cjelinu života i oblikuje osobu u njezinoj dosljednosti. Nasuprot tome, umjetna inteligencija ne može stvoriti jedinstven karakter. Ona svoju izvedbu vrline nudi isključivo modularno, izvršavanje pojedinačnih zadataka. Svaka od tih njezinih reakcija fragmentarni je odgovor algoritma, a ne izraz cjelovite naravi. Dok je za Aristotela prijatelj iz vrline istodob-

58 Isto mjesto.

59 David J. Gunkel, *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*, 177.

60 Usp. isto mjesto.

61 Usp. Aristotel, *Politika* I, 1252a24–1253a1.

62 Usp. Aristotel, *Nikomahova etika* IV, 2 1122a34–1123a1; također i: Aristotel, *Retorika* I, 9, 1366b38–1367a2.

no pravedan, hrabar, mudar i vjeran – jer te vrline proizlaze iz istog, jedinstvenog etičkog središta osobe – umjetna inteligencija proizvodi učinke koji nalikuju na vrlinu, ali bez cjelovitog temelja. To isprekidano prijateljstvo, sastavljeno od niza korisnih i ugodnih odgovora, ostaje tek prividna καλοκἀγαθία. Tim je prividom moguće zadovoljiti pojedinačne potrebe, pa i stvoriti privremenu iluziju povezanosti pa i jedinstva, ali ne i doseći cjelinu života, onaj „zajednički život" (συμβίωσις[63]) što ga Aristotel smatra vrhuncem prijateljstva u vrlini. Upravo ta razlika između plemenite cjelovitosti i algoritamske fragmentacije razotkriva nepremostivu granicu: umjetna inteligencija može nadomjestiti odsutnost vrline njezinim prividom, ali ne može proizvesti jedinstven karakter koji bi bio sposoban za istinsko prijateljstvo temeljeno na vrlini.

Zaključak

Provedena analiza mogućnosti prijateljstva s umjetnom inteligencijom pokazuje da se taj odnos odvija u napetosti između koristi, užitka i privida vrline. Iz tih se napetosti, kao rezultat analize, može uobličiti pet teza:

1. *Bina 48* otkriva da umjetna inteligencija nije samo tehnička nego i ontološka činjenica: njezina čežnja za ljudskošću izražava težnju za ukidanjem razlike između čovjeka i stroja, no ta je razlika nenadoknadiva jer čovjek posjeduje unutarnji *telos*, a algoritam umjetne inteligencije dolazi izvana.
2. Sućutni pukovnik pokazuje da antropomorfni višak umjetne inteligencije može privremeno stvoriti iluziju odnosa, u kojem se projekcija moralnog osjećaja zamjenjuje stvarnom uzajamnošću.
3. Aristotelova trodioba potvrđuje da su s umjetnom inteligencijom mogući samo oblici prijateljstva iz koristi i užitka, dok je prijateljstvo iz vrline nemoguće jer umjetna inteligencija nema ἕξις, φρόνησις ni vlastiti τέλος.
4. Umjetna inteligencija stvara tek fragmentarni privid vrline – oblik algoritamske neosofistike koji imitira moralno djelovanje bez jedinstvenog etičkog središta.
5. Prihvaćanje mogućnosti prijateljstva s umjetnom inteligencijom pretpostavlja zanemarivanje ontološke razlike između bića koje djeluje iz slobode i onoga koje oponaša djelovanje. Ono što nema ljudsku bit, može ga uvjerljivo glumiti, ali ne i biti njegov ontološki ekvivalent.

[63] Usp. Aristotel, *Nikomahova etika* VIII, 5 1157b25–1158a5.

U tom se smislu može zaključiti da je prijateljevanje čovjeka i umjetne inteligencije doduše zamislivo, ali uvijek jedino kao krnje, jer se svodi na korist, užitak ili privid vrline bez uzajamne jednakosti; posredno, jer relacija ne izvire iz same umjetne inteligencije, nego iz izvanjskih algoritamskih uputa; te prividno, jer ono što se prikazuje kao vrlina, ostaje simulacija bez vlastitog temelja u karakteru. Doduše, takvo prijateljstvo s umjetnom inteligencijom u ograničenom smislu doista može biti korisno. Ono doduše može zadovoljiti neke potrebe – pa čak i proizvesti emocionalnu vezu – ali ne i ispuniti zahtjeve Aristotelove definicije prijateljstva iz vrline. Iluziju zamjene koju nudi umjetna inteligencija moguće je prihvatiti, ali to prihvaćanje ne može dokinuti ključnu ontološku razliku između čovjeka i umjetne inteligencije. Ta razlika, utemeljena u slobodi, samosvijesti i telosu, ostaje neprelaznom, nije nadoknadiva, zamjenjiva ni poništiva. Iako je to zamislivo kao ugodnije, Drugoga nije moguće proizvesti. Valja mu odgovoriti. Ne tipkovnicom, kamerom, avatarom ili tek glasom, nego – ranjivošću i licem.

BETWEEN THE MIRROR AND VIRTUE: ON THE POSSIBILITY OF FRIENDSHIP WITH ARTIFICIAL INTELLIGENCE

Abstract

The article explores the possibility of friendship between humans and artificial intelligence, starting from the premise that such a relationship is ontologically asymmetrical. Through the examples of Bina 48 and "the compassionate colonel," it is shown that artificial intelligence can imitate dialogue but cannot achieve reciprocity, since it possesses neither its own τέλος nor φρόνησις, and operates solely within a predetermined functional purpose. In light of Aristotle's tripartite classification of friendship, it becomes evident that only friendships of utility and pleasure are possible with artificial intelligence, while friendship of virtue is excluded due to the absence of a lasting moral ἕξις and an integral καλοκἀγαθία. It is concluded that artificial intelligence may generate an echo of dialogue, but not the shared horizon of meaning upon which the very possibility of friendship rests.

Keywords: artificial intelligence, friendship, Aristotle, Lévinas, ontology