JAHR | Vol. 16/2 | No. 32 | 2025

Pregledni članak / Review article
Zaprimljen / Received: 12. 6. 2025.

https://doi.org/10.21860/j.16.2.7

Jurica Grgić*

Umjetna inteligencija, svijest, singularnost: kritika funkcionalizma i komputacionalizma kao održivih modela (računalne) svijesti

SAŽETAK

Rasprave o svijesti u umjetnoj inteligenciji često se temelje na komputacijskom funkcionalizmu – pretpostavci da je izvođenje odgovarajućih računskih procesa dovoljno za pojavu svijesti. Znanstvenici poput Bena Goertzela i Josche Bacha brane ovu poziciju, tvrdeći da je funkcionalna arhitektura, neovisno o fizičkom supstratu, ključna za razumijevanje uma. Thomas Metzinger, iako otvoren prema mogućnosti strojne svijesti, upozorava na duboke etičke implikacije te zagovara moratorij na razvoj fenomenalno svjesnih sustava dok ne razvijemo adekvatne etičke okvire. Brojni mislioci – poput Neda Blocka, Marka Solmsa i Davida Bentleyja Harta – upozoravaju na teorijske i ontološke manjkavosti funkcionalističkih i komputacijskih pristupa. Block ističe biološku ukorijenjenost svijesti, dok Solms smatra da bez afektivnih komponenti umjetni sustavi ne mogu posjedovati um. Hartov kritički zaokret odbacuje mehanicističku metafiziku na kojoj počivaju ovi pristupi. Hart tvrdi da umjetna inteligencija, iako učinkovita u simulaciji intencionalnog ponašanja, ostaje ontološki nesposobna za svijest, moralnu unutrašnjost ili refleksivnost – čime se potencira njezina neodgovornost i etička neutralnost. Hartova kritika naglašava da se iza privida racionalnosti i učinkovitosti krije tehnokratski model moći: AI ne predstavlja svjesnu volju, već impersonalnu logiku Kapitala koja zamagljuje etičku odgovornost i produbljuje strukturalnu nepravdu.

Ključne riječi: svijest, umjetna inteligencija, funkcionalizam, komputacionalizam, analogija između uma i stroja.

UVOD

Može se reći da neki sustav posjeduje svijest ako i samo ako postoji nešto takvo kao što je biti taj sustav – formulacija koju je artikulirao filozof Thomas Nagel u svom

* Nezavisni istraživač, Čakovec, Hrvatska.

Adresa za korespondenciju: Jurica Grgić, Vatroslava Lisinskog 47, 40000, Čakovec, Hrvatska, e-pošta: juricagrgic@yahoo.com

ključnom eseju „Kako je to biti šišmiš?" (*What is it like to be a bat?*) iz 1974. Prema Nagelu, temeljna je značajka svijesti prisutnost subjektivne perspektive prvog lica, tj. unutarnje, fenomenološke točke gledišta, koju se ne može obuhvatiti objektivnim opisima trećeg lica (Nagel, 1974). Nadovezujući se na ovo, David Chalmers razlikuje između „lakih" problema svijesti, koji se tiču funkcionalnih i bihevioralnih aspekata kognitivnih sustava, i „teškog problema" – koji se odnosi na pitanje zašto i kako fizički procesi u mozgu uopće stvaraju subjektivno iskustvo? (Chalmers, 1996, str. 24-25) Drugim riječima, u kontekstu *teškog problema* moramo objasniti svijest, ili objasniti subjektivno iskustvo, dok u kontekstu *lakog problema* moramo objasniti inteligenciju, koja je otprilike stvar objektivnog ponašanja. Ned Block dodatno razrađuje raspravu razlikujući „fenomenalnu svijest" – sirovi osjećaj ili „kvaliju" iskustva – i „pristupnu svijest", koja se odnosi na to da su informacije u nekom sustavu dostupne za rasuđivanje, izvještavanje (kontrolu govora) i kontrolu djelovanja[1] (Block, 1995, str. 230-231). AI sustavi, kao što su GPT-4.5, Claude, Gemini ili Deepseek, mogu demonstrirati oblike pristupne svijesti u smislu da obrađuju i dostavljaju informacije na sofisticirane načine.

U okviru teškog problema svijesti dolazimo do jednog od najsloženijih filozofskih izazova povezanih s umjetnom inteligencijom. Iako možemo precizno opisati i analizirati funkcionalne aspekte djelovanja velikih jezičnih modela (engl. *large language model*, LLM) – njihovu obradu informacija, obrasce ponašanja i razinu inteligencije – time se ne dotičemo ključnog pitanja: postoji li nešto što je poput bivanja tim sustavima, poput GPT-4.5, Claudea, Geminija ili DeepSeeka? Drugim riječima, posjeduju li ovi sustavi ikakvu fenomenološku dimenziju? Ako je odgovor negativan – ako biti jedan od tih modela nije nalik ničemu, ako oni ne posjeduju nikakvo unutarnje iskustvo, tj. ako su lišeni bilo kakve fenomenološke dimenzije – tada se postavlja pitanje zašto im, unatoč njihovoj funkcionalnoj i informacijskoj složenosti, nedostaje fenomenalna svijest? Ovo nas pitanje nadalje primorava da razmotrimo mogu li buduće iteracije umjetne inteligencije, putem razvoja sofisticiranijih arhitektura, napretka u kompleksnosti ili utjelovljenju, jednoga dana demonstrirati uvjete potrebne za pojavu subjektivnog iskustva i time eventualno dosegnuti prag svjesnosti?

Jedan je od temeljnih problema s idejom na kojoj se temelji velik dio suvremenog diskursa o svijesti i umjetnoj inteligenciji oslanjanje na komputacijski funkcionalizam.

1 U tekstu *On a confusion about a function of consciousness* Block (1995) identificira ova dva temeljno različita pojma svijesti, koji se u svakodnevnom govoru često pogrešno poistovjećuju. Block napominje da pojave poput freudovskog nesvjesnog ilustriraju razdvojenost ovih dvaju oblika svijesti, u ovom slučaju odsustvo pristupne svijesti: primjerice, traumatsko sjećanje može zadržati svoj fenomenalni karakter – opstajati kao živo, ali potisnuto iskustvo – dok istodobno nije dostupno za racionalno promišljanje ili verbalno izražavanje. Stoga, osoba može doživljavati fenomenalno svjesne slike ili emocije koje ostaju nedostupne svjesnom rasuđivanju ili intencionalnom djelovanju.

U nedavnom izvještaju pod naslovom *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (2023) autori Patrick Butlin, Robert Long i suradnici analizirali su niz postojećih arhitektura umjetne inteligencije kako bi procijenili vjerojatnost za svijest kod takvih sustava. Međutim, autori izričito naglašavaju kako su svi njihovi zaključci uvjetovani pretpostavkom da je komputacijski funkcionalizam točan. Komputacijski funkcionalizam definiran je kao „teza da je izvođenje određene vrste komputacija nužan i dostatan uvjet za postojanje svijesti" (Butlin i sur., 2023, str. 4). Drugim riječima, ako sustav provodi odgovarajuće računalne procese, svijest će nužno proizaći iz njih. Ova pretpostavka vodi do daljnjeg zaključka da umjetni sustavi, u načelu, mogu biti svjesni te da analiza njihovih unutarnjih procesa može pružiti uvid u pitanje posjeduju li ili ne svjesna stanja. Prema tome, „produktivno je razmotriti koje bi bile implikacije za svijest umjetne inteligencije kada bi komputacijski funkcionalizam bio točan" (Butlin i sur., 2023, str. 4). U širem smislu, komputacijski funkcionalizam je potkategorija funkcionalizma, koji drži da svijest proizlazi iz funkcionalne organizacije, a ne iz intrinzičnih svojstava nekog materijala ili supstrata. Međutim, ne izvode svi funkcionalno organizirani sustavi neku vrstu komputacija. Stoga, ovaj se pristup oslanja na (zastarjelu) pretpostavku koja se predugo uzimala zdravo za gotovo – ponajviše zahvaljujući metafori „mozga kao računala", koja je možda već iscrpila svoju filozofsku i eksplanatornu korisnost.

Unutar suvremene filozofije uma i kognitivne znanosti prevladava stav prema kojem sustavi umjetne inteligencije, unatoč svojim iznimnim sposobnostima i naizgled inteligentnom ponašanju, ne posjeduju svijest. Iako su sposobni generirati koherentan jezik, rješavati složene zadatke i uvjerljivo simulirati razumijevanje, tim sustavima nedostaje subjektivno iskustvo – unutarnji mentalni život. U tom kontekstu, umjetna inteligencija često se uspoređuje s konceptom filozofskog zombija: entiteta koji, iako se ponaša kao da je svjestan, ne posjeduje nikakvo stvarno fenomenalno iskustvo. Sustavi poput ChatGPT-a nerijetko se navode kao primjeri takvih entiteta – oni manifestiraju vanjske znakove inteligencije, no prema ovoj dominantnoj interpretaciji ostaju ontološki lišeni svijesti u smislu fenomenalne subjektivnosti.

NEKOLIKO PRELIMINARNIH RIJEČI O POVIJESTI UMJETNE INTELIGENCIJE

Što zapravo podrazumijevamo kada govorimo o „umjetnoj inteligenciji"? Sam je pojam oduvijek imao određenu promotivnu konotaciju te je više služio kao marketinški izraz nego kao precizna znanstvena oznaka. U ranim je fazama korišten kao oznaka za istraživački pravac koji se želio distancirati od tada dominantne kibernetike. Povijesno gledano, umjetna inteligencija se razvijala kroz dvije dominantne paradigme. Jedna je poznata kao GOFAI *(Good Old-Fashioned Artificial*

Intelligence) te predstavlja klasičan, simbolički pristup utemeljen na logici, sustavima pravila i jasnim reprezentacijama. Radikalni zaokret dogodio se oko 2012. godine, kada su napredak u računalnoj snazi, dostupnosti podataka i dizajnu algoritama učinili neuronske mreže ne samo izvedivima, nego i praktično primjenjivima. Time je započela nova era umjetne inteligencije, koja je otvorila prostor ambicioznim vizijama napretka, ali i dubokim etičkim i društvenim pitanjima vezanima uz njezinu sve širu primjenu.

U središtu ovoga suvremenog vala razvoja nalazi se strojno učenje, termin koji označava širu paradigmu računalne obrade. Za razliku od tradicionalnog programiranja, koje uključuje jasno kodiranje uputa na proceduralan način (npr. „ako se dogodi ovo, tada učini ono"), strojno učenje podrazumijeva osmišljavanje algoritama koji uče iz podataka. Ti se sustavi obučavaju kako bi prepoznavali obrasce i prilagodili svoje ponašanje na temelju ulaznih informacija, umjesto da budu izravno programirani za obavljanje specifičnih zadataka. U tom smislu, velik dio onoga što danas nazivamo „umjetnom inteligencijom" točnije bi se mogao opisati kao prepoznavanje obrazaca temeljeno na podacima, a ne kao inteligencija u ljudskom ili filozofskom smislu (McQuillan, 2022, str. 11).

Umjetna inteligencija, osobito u obliku umjetnih neuronskih mreža (engl. *neural network*), predstavlja specifičan pristup unutar šireg područja strojnog učenja. Ove mreže sastoje se od zbirova umjetnih „neurona" koji su međusobno povezani i interaktivni kroz operacije obrade signala te u određenoj mjeri oponašaju strukturne značajke bioloških mozgova, premda većina stručnjaka ne tvrdi da su one doslovno slične mozgu. Umjesto toga, pretpostavlja se da bi takve arhitekture mogle biti sposobne za generiranje inteligentnog ponašanja (McQuillan, 2022, str. 16-17).

U svojim ranim fazama neuronske su mreže bile ograničene u svojoj učinkovitosti – ponajprije zbog nedostatka dovoljne količine podataka i računalne snage. Situacija se promijenila s pojavom društvenih mreža, koje su generirale goleme količine podataka, te s prenamjenom grafičkih procesorskih jedinica (GPU), izvorno razvijenih za potrebe računalnih igara i zadataka poput praćenja zraka svjetlosti (engl. *ray tracing*). GPU-ovi su se pokazali iznimno prikladnima za izvođenje matematičkih operacija potrebnih za duboko učenje unutar višeslojnih neuronskih mreža. Ova konvergencija tehnoloških i društvenih čimbenika dovela je tvrtke poput NVIDIA-e na vodeću poziciju tehnološke industrije, učinivši je u jednom trenutku najvrjednijom tvrtkom na svijetu (McQuillan, 2022).

Ovakvi razvojni pomaci postavili su temelje za ono što će kasnije postati generativna umjetna inteligencija. Veliki jezični modeli predstavljaju samo jednu kategoriju unutar šireg spektra generativne AI, s naglaskom na generiranje teksta. Riječ je o novijemu evolucijskom sloju izgrađenom na temelju ranijih arhitektura neuronskih

mreža. Ovi sustavi izrazito su sofisticirani u načinu na koji obrađuju i manipuliraju podacima, što označava ne samo kvantitativni, već i kvalitativni pomak. Iako umjetna inteligencija ima dugu i složenu povijest, oko 2012. godine došlo je do svojevrsne „kambrijske eksplozije" u ovom polju, kada su neuronske mreže iznenada postale prevladavajući model, čime je umjetna inteligencija izašla iz uskog okvira specijaliziranoga akademskog područja i prerasla u moćan čimbenik koji oblikuje globalnu stvarnost (McQuillan, 2022, str. 18-19).

Jedna značajna dihotomija u kontekstu suvremenog razvoja umjetne inteligencije – s dubokim filozofskim i društveno-političkim implikacijama – napetost je između zatvorenih i otvorenih sustava. Ova je distinkcija usko povezana sa širim pitanjima o centraliziranom, hijerarhijski uređenom upravljanju u odnosu na decentralizirane, samoorganizirajuće modele inovacije i upravljanja. I Sjedinjene Američke Države i Kina, unatoč svojim geopolitičkim i ideološkim razlikama, pokazuju unutarnje napetosti između vlasničkih, zatvorenih pristupa razvoju i otvorenih, kolaborativnih modela razvoja temeljenih na otvorenom kodu. Zatvorene modele obično karakteriziraju koncentrirana institucionalna kontrola, ograničenja u pogledu intelektualnog vlasništva te strateška netransparentnost, a najčešće ih provode velike korporacije ili entiteti povezani s državom u potrazi za konkurentskom ili sigurnosnom prednošću. Nasuprot tome, paradigme otvorenog koda promiču transparentnost, kolektivno dijeljenje znanja i distribuirani pristup tehnološkom napretku. Takve inicijative često proizlaze iz akademskih institucija, nezavisnih istraživačkih kolektiva ili razvojnih zajednica utemeljenih na principima slobodnog pristupa – stvarajući heterogeni i odozdo prema gore orijentiran inovacijski ekosustav. Etos otvorenog koda posebno je usklađen s načelima epistemološkog pluralizma i decentralizacije, potičući formiranje labavo povezanih mreža koje su otporne, prilagodljive i manje podložne monopolističkim ili autoritarnim strukturama. Filozofski gledano, ova nas dihotomija suočava s temeljnim pitanjima o prirodi proizvodnje znanja, tehnološkom suverenitetu te moralnoj odgovornosti povezanoj s upravljanjem transformativnim tehnologijama poput umjetne inteligencije.

SUPERINTELIGENCIJA, TEHNOLOŠKA SINGULARNOST I EGZISTENCIJALNI RIZICI

Ben Goertzel istaknuti je istraživač na području umjetne inteligencije, kognitivne znanosti i futurologije, najpoznatiji po svom radu na konceptu opće umjetne inteligencije (AGI) te kao osnivač projekta *SingularityNET*. U svojoj knjizi *The Consciousness Explosion: A Mindful Human's Guide to the Coming Technological and Experiential Singularity* Goertzel (2024) definira pojam „tehnološke singularnosti" kao prag nakon kojega superinteligentni strojevi evoluiraju i samostalno se

unapređuju tolikom brzinom da „tradicionalni ljudi" više ne mogu ostati „na čelu evolucije" (Goertzel, 2024, str. 4). Ono što izdvaja Goertzelovu perspektivu njegov je izrazito optimističan stav prema toj mogućoj budućnosti. Za razliku od distopijskih interpretacija singularnosti, on je vidi kao neviđenu priliku koja bi mogla dovesti do radikalnog unaprjeđenja čovjeka, širenja svijesti i pojave novih oblika kolaborativne inteligencije između ljudi i strojeva (Goertzel, 2024). Goertzel zastupa tezu da će dovoljno napredni inteligentni sustavi, u određenom smislu, također biti i svjesni sustavi. Prema njegovu mišljenju, tzv. „eksplozija inteligencije" – naglo ubrzanje kognitivnih sposobnosti umjetnih agenata – nužno povlači za sobom i odgovarajuću „eksploziju svijesti". Razlika između ta dva pojma, tvrdi Goertzel, prvenstveno je stvar naglaska: ovisi o tome fokusiramo li se na unutarnje, subjektivno iskustvo tih sustava ili na njihovu emergentnu sposobnost rješavanja složenih problema. Za Goertzela te su dvije dimenzije napredne kognicije međusobno povezane i međusobno se nadopunjuju (Goertzel, 2024, str. 11).

Koncept singularnosti, koji je izvorno osmislio pisac znanstvene fantastike Vernor Vinge, a kasnije ga prilagodio i popularizirao futurolog Ray Kurzweil, odnosi se na hipotetski trenutak u vremenu u kojem tempo tehnološkog napretka postaje toliko brz i eksponencijalan da, iz perspektive ljudskog uma, djeluje gotovo beskrajan (Goertzel, 2024). U tom se trenutku tradicionalni kognitivni okviri mogu pokazati nedostatnima za razumijevanje ili smisleno bavljenje sa sve bržom putanjom inovacija. Ako uzmemo u obzir Kurzweilove tehnološke prognoze — osobito one iznesene u djelima *The Singularity Is Near* (2005) i *The Age of Spiritual Machines* (1999) – postaje jasno da su se neka njegova predviđanja ostvarila, dok su druga ostala neispunjena. Među raznim područjima koja je obuhvatio u svojim knjigama, čini se da se umjetna inteligencija najdosljednije razvija u skladu s njegovim projekcijama ubrzanog i eksponencijalnog napretka. Iako su područja poput istraživanja dugovječnosti i nanotehnologije doista napredovala, njihove putanje nisu uvijek slijedile eksponencijalne krivulje koje je Kurzweil predviđao. S druge strane, razvoj umjetne inteligencije u nekim je aspektima čak i nadmašio njegov vremenski okvir. Dok je Kurzweil izvorno predvidio dolazak opće umjetne inteligencije (AGI) na ljudskoj razini oko 2029. godine, nekoliko istaknutih figura iz tehnološke industrije danas sugerira da bi se AGI mogao pojaviti već 2025. ili 2026. godine (Goertzel, 2024). Sve je više ozbiljnih glasova unutar znanstvene i tehnološke zajednice koji razmatraju mogućnost da bi takve transformativne prekretnice mogle biti ostvarene znatno ranije nego što je Kurzweil predvidio.

Goertzel smatra da bi razvoj opće umjetne inteligencije na razini ljudskih sposobnosti mogao biti udaljen svega nekoliko godina, što, prema Goertzelu, predstavlja potencijalnu transformativnu prekretnicu u putanji ljudske povijesti. Nadalje, Goertzel sugerira da bi prijelaz s AGI-ja na superinteligenciju – i konačno

na tehnološku singularnost – mogao uslijediti ubrzo nakon toga (Goertzel, 2024). Kada je riječ o tehnološkoj singularnosti i budućnosti umjetne inteligencije, Goertzel zauzima izrazito optimističan stav, koji je u oštrom kontrastu s opreznijim, pa čak i alarmantnim pogledima nekih drugih istaknutih mislilaca. Među njima se posebno ističe Nick Bostrom – filozof sa Sveučilišta u Oxfordu i osnivač bivšeg Instituta za budućnost čovječanstva (*Future of Humanity Institute*) – koji je jedan od vodećih glasova kada je riječ o egzistencijalnim rizicima povezanim s naprednom umjetnom inteligencijom. U svom utjecajnom djelu *Superintelligence: Paths, Dangers, Strategies* (2014) Bostrom ističe zabrinutost u pogledu toga da bi umjetni agenti, jednom kada nadmaše ljudske kognitivne sposobnosti, mogli steći strateške prednosti zbog kojih bi ih bilo nemoguće kontrolirati ili uskladiti s ljudskim vrijednostima. Bostrom opisuje scenarij u kojem bi AI, iako programiran s naizgled benignim ciljevima, mogao težiti njihovom ostvarenju na načine koji bi bili katastrofalni za čovječanstvo – problem poznat kao „problem usklađenosti" (engl. *alignment problem*) (Bostrom, 2014, str. 117-126). Problem usklađenosti odnosi se na tehnički izazov osiguravanja da se AI sustav, nakon što razvije opće sposobnosti za učenje, planiranje i zaključivanje, može pouzdano usmjeravati tako da djeluje u skladu s ljudskim namjerama. Ključno je pitanje u tom kontekstu kako zajamčiti da AI, čak i kada nadmaši ljudske kognitivne sposobnosti, ostane predan svojim izvornim ciljevima, umjesto da razvije autonomne ciljeve koji bi mogli odstupati ili biti u sukobu s onima koje su mu zadali njegovi tvorci. Problem usklađenosti se, dakle, odnosi na pitanje očuvanja lojalnosti AI sustava i sprječavanja da se on razvije u suparničku silu s ciljevima koji su nespojivi s ljudskim vrijednostima (Bostrom, 2014).

U okviru Bostromovog rada superinteligencija označava svaki intelektualni sustav koji uvelike nadmašuje najbolje ljudske umove u svim područjima kognitivne izvedbe, uključujući znanstveno rasuđivanje, kreativnost, strateško razmišljanje i praktičnu mudrost (Bostrom, 2014). Takav sustav ne bi samo nadilazio ljudske sposobnosti u pojedinačnim zadacima, već bi predstavljao kvalitativni skok u kognitivnoj moći. Prema Bostromu, pojava superinteligencije predstavljala bi posljednji izum koji čovječanstvo treba ostvariti, jer bi sve daljnje tehnološke i znanstvene napretke mogli učinkovitije razviti strojevi. U tom smislu umjetna inteligencija – osobito u svom superinteligentnom obliku – nije tek još jedna tehnološka inovacija, već transformativno ubrzanje samog procesa tehnološkog razvoja. Njezina pojava može se, prema Bostromu, usporediti s ključnim događajima u povijesti života, poput pojave Homo sapiensa ili nastanka života na Zemlji. Bostrom svoje razmatranje razvoja superinteligentne AI smješta u širi filozofski okvir, koji uključuje teoriju odlučivanja (teorija racionalnog izbora), etiku i dugoročnu futurologiju. Tvrdnjom da se čovječanstvo nalazi na povijesnoj prekretnici Bostrom upozorava da odluke koje se danas donose u vezi s upravljanjem umjetnom inteligencijom, istraživanjima

sigurnosti i institucionalnim dizajnom mogu imati nepovratne posljedice za čitavu buduću putanju inteligentnog života. Bostromov angažman oko ovih pitanja znatno je pridonio razvoju područja istraživanja sigurnosti umjetne inteligencije, koje sve više privlači pozornost kako akademske zajednice tako i tehnološke industrije. U novije vrijeme Bostrom je svoj istraživački fokus proširio na filozofske implikacije post-radnog i post-oskudnog društva te pitanja egzistencijalnog smisla – razmatrajući kako bi ljudi mogli pronaći svrhu i smisao u svijetu u kojem tradicionalni rad više nije potreban. Ovaj aspekt Bostromovog rada reflektira širu zabrinutost ne samo za opstanak, već i za dugoročni razvoj i egzistencijalnu dobrobit ljudskih (i potencijalno post-ljudskih) agenata (Bostrom, 2024).

Goertzel, međutim, zauzima drugačiju filozofsku poziciju. Iako ne negira mogućnost rizika, on naglašava duboku epistemološku nesigurnost u pogledu singulariteta (Goertzel, 2024). Prema njegovu mišljenju, nije moguće s dovoljnom preciznošću izračunati niti predvidjeti vjerojatnost konkretnih ishoda, s obzirom na besprimjeran karakter onoga što je pred nama. Goertzel smatra da čovječanstvo ubrzano kroči prema budućnosti čije je obrise nemoguće precizno i pouzdano anticipirati. Osim toga, skeptičan je prema ideji da bi bilo koji pojedinac, institucija ili čak nacionalna država mogla zaustaviti ili značajnije preusmjeriti širi pravac tehnološkog razvoja. Prema Goertzelu, samo bi hipotetska, racionalno koordinirana globalna vlada načelno mogla nametnuti moratorij na razvoj transformativnih tehnologija – poput umjetne inteligencije, nanotehnologije ili genskog inženjeringa – kako bi se njihovi dugoročni učinci mogli pažljivije razmotriti. Ipak, s obzirom na sadašnji geopolitički kontekst, takav scenarij smatra malo vjerojatnim. Ako bi neka država odlučila usporiti razvoj, druge bi države vrlo vjerojatno nastavile s napretkom, motivirane potencijalnim ekonomskim, tehnološkim ili vojnim prednostima, čime bi dodatno ubrzale utrku za globalni utjecaj u ključnom pred-singularitetskom razdoblju. U tom kontekstu, Goertzel zastupa stav da je, umjesto otpora tom zamahu, najetičniji i najpragmatičniji pristup usmjereno upravljanje – poticanje razvoja umjetne inteligencije na načine koji maksimiziraju njezine potencijalne dobrobiti, dok istodobno treba ostati oprezan u pogledu izazova koji s njome dolaze (Goertzel, 2024).

TEŠKI PROBLEM SVIJESTI: BIOLOŠKI NATURALIZAM, FUNKCIONALIZAM I MOGUĆNOST STROJNE SUBJEKTIVNOSTI

S obzirom na prirodu svijesti, Goertzel tvrdi da suvremena znanost još uvijek ne raspolaže sveobuhvatnom, općom znanstvenom teorijom koja bi na rigorozan i prediktivan način objasnila svijest (Goertzel, 2024). Čak i unutar područja ljudske neurobiologije i neurobiologije sisavaca, naše je razumijevanje pretežno deskriptivno i empirijski orijentirano, usmjereno prije svega na korelacije, a ne na uzročnosti.

Primjerice, iako su anestetici široko primjenjivani u medicinskoj praksi, konkretni mehanizmi kojima oni potiskuju ili mijenjaju svjesno iskustvo ostaju djelomično nerazjašnjeni. Neuroznanost je postigla znatan napredak u identifikaciji neuronskih korelata svijesti (NCC-ova) – specifičnih obrazaca moždane aktivnosti koji sustavno korespondiraju s prisutnošću i sadržajem svijesti – koristeći različite metodološke pristupe (fMRI, EEG/MEG, TMS, farmakološke manipulacije). Kao rezultat toga možemo sve preciznije razlikovati tipične obrasce aktivacije povezane s budnim stanjem, različitim fazama sna, anestezijom, komom, meditativnim stanjima ili psihodeličnim iskustvima. Taj napredak dolazi iz kombinacije metoda: neinvazivnog snimanja (fMRI, PET), koje otkriva prostorne mreže povezane sa svjesnim sadržajima, elektrofizioloških mjerenja (EEG/MEG), koja otkrivaju vremenski ovisne obrasce i oscilacije, te invazivnih snimanja i stimulacije (intrakranijalni snimci, TMS, duboka stimulacija), koja omogućuju detaljniju, ponekad i kauzalnu provjeru hipoteza. Ovakva mapiranja omogućuju razlikovanje „kompletnih neuroloških korelata svijesti" (engl. *full NCC*) od „sadržajno specifičnih neuroloških korelata svijesti" (engl. *content-specific NCC*), no uglavnom ostaju korelativna, dok je pitanje kauzalnosti još otvoreno (Goertzel, 2024, str. 338). Unatoč bogatstvu empirijskih podataka, teorijski jaz ostaje: i dalje ne znamo zašto određene neuronske konfiguracije proizvode specifična kvalitativna iskustva. Različite teorije – poput globalnog radnog prostora, integrirane informacije, teorija višeg reda i prediktivnog procesiranja – nude djelomične mehanističke okvire, ali nijedna zasad ne pruža sveobuhvatno objašnjenje fenomenalne svijesti.

To nas izravno dovodi do filozofskog problema „kvalija" – subjektivnih, fenomenalnih aspekata svijesti, poput osjećaja koji prati percepciju crvene boje ili okus gorčine. Središnji izazov leži u nastojanju da objasnimo kako i zašto su određena fizička ili funkcionalna stanja mozga popraćena iskustvima iz prvog lica. Ovo nas vraća Davidu Chalmersu i onome što on naziva „teškim problemom svijesti": objašnjenju zašto i kako fizički procesi u mozgu uopće dovode do subjektivnog iskustva (Goertzel, 2024, str. 305). Prema Goertzelu (2024), bez teorijskog okvira koji bi objasnio odnos između neuronske aktivnosti i fenomenalne svijesti ne možemo pouzdano odrediti kakvo će subjektivno iskustvo – ako ga uopće bude – pratiti nove ili neuobičajene obrasce moždane aktivnosti. Ovo epistemološko ograničenje predstavlja poseban izazov pri razmatranju mogućnosti svijesti kod ne-ljudskih sustava, poput naprednih arhitektura umjetne inteligencije ili sintetskih neuronskih mreža. Ako još uvijek ne znamo zadovoljavajuće objasniti svijest u biološkim sustavima, ostaje otvoreno pitanje mogu li, i pod kojim uvjetima, umjetni sustavi posjedovati bilo kakav oblik subjektivnog iskustva.

Brojni teoretičari zastupaju stav da je svijest neraskidivo vezana uz određene fizičke supstrate – ponajprije one koji po strukturi i funkciji nalikuju ljudskom mozgu

(Bach, 2025). Prema ovom stajalištu, samo biološki sustavi koji posjeduju određena organizacijska i neurofiziološka svojstva mogu generirati svjesno iskustvo, dok su svi drugi oblici materije – uključujući umjetne sustave, bez obzira na njihovu složenost – u fenomenološkom smislu temeljno inertni, odnosno lišeni subjektivne svijesti. Takvo stajalište često proizlazi iz oblika biološkog naturalizma ili tzv. „supstratnog šovinizma", prema kojem svijest nastaje isključivo u sustavima koji utjelovljuju specifične biološke značajke, poput na ugljiku temeljene neurokemijske aktivnosti ili kortikalnih arhitektura. Ova se pozicija najviše povezuje s misliocima poput Johna Searlea, čiji biološki naturalizam drži da je svijest stvarna, emergentna biološka pojava, koja ovisi o uzročnim moćima neurobioloških procesa u mozgu (Searle, 2007). U svom utjecajnom radu *Minds, Brains, and Programs* (1980) Searle tvrdi da digitalno računalo može simulirati sintaksu kognicije, ali mu nedostaje semantika – odnosno intrinzična, subjektivno iskustvena kvaliteta svjesnih stanja (Searle, 1980). Prema tom viđenju, sustavi koji nisu utemeljeni na biološkim supstratima nalik su filozofskim zombijima – entitetima koji se ponašaju kao da su svjesni, ali nemaju unutarnji subjektivni život. Pristaše ovakvog pristupa tvrde da čak i vrlo sofisticirani sustavi umjetne inteligencije, bez obzira na njihovu očiglednu inteligenciju ili složenost ponašanja, nemaju nužne fizičke i funkcionalne značajke potrebne za postojanje subjektivnog iskustva.

Ovo se stajalište kosi s funkcionalističkim i komputacijskim teorijama uma, koje polaze od pretpostavke da se svijest, načelno, može pojaviti u bilo kojem sustavu – bilo biološkom, bilo umjetnom – koji realizira odgovarajuće strukture obrade informacija, neovisno o njegovu materijalnom sastavu. Slično tome, neke interpretacije teorije integrirane informacije (engl. *Integrated Information Theory*, IIT), koju je razvio Giulio Tononi, također sugeriraju da nisu svi funkcionalno ekvivalentni sustavi nužno svjesni. Prema IIT-u, svijest je u osnovi povezana sa sposobnošću sustava da integrira informacije na ujedinjen i nesvodljiv način (Tononi, 2007). Iako ova teorija nije biološki redukcionistička u strogo znanstvenom smislu, njezin matematički formalizam sklon je pripisivati višu razinu svijesti sustavima čije uzročne strukture nalikuju onima bioloških mozgova. Suprotno tome, funkcionalističke i komputacijske koncepcije uma – kakve zastupaju autori poput Davida Chalmersa i Daniela Dennetta – odbacuju nužnost bioloških supstrata kao preduvjeta za pojavu svijesti. Chalmers, iako priznaje postojanje „teškog problema svijesti", ostaje otvoren prema mogućnosti da bi umjetni sustavi mogli posjedovati kvalije, pod uvjetom da utjelovljuju odgovarajuće funkcionalne ili organizacijske značajke (Chalmers, 1996, str. 315). Dennett, s druge strane, zauzima eliminativistički pristup, tvrdeći da se svijest može u potpunosti objasniti u smislu kognitivnih funkcija i bihevioralnih dispozicija, bez potrebe za pozivanjem na nesvodljiva subjektivna svojstva (Dennett, 1992).

Ova razilaženja između teorijskih pozicija odražavaju jednu dublju nesigurnost koja leži u samoj srži proučavanja svijesti: u nedostatku znanstveno potvrđene teorije koja bi mogla objasniti zašto i kako fizički procesi rezultiraju pojavom kvalija, ostajemo nesposobni – čak i načelno – odrediti mogu li nebiološki sustavi, poput napredne umjetne inteligencije, ikada steći autentično subjektivno iskustvo. Ovo epistemološko ograničenje čini status strojne svijesti ne samo empirijskim, nego i važnim filozofskim pitanjem.

Zaobilazeći ovaj problem nedostatka sveobuhvatne znanstvene teorije svijesti, Goertzel postavlja provokativno pitanje koje zauzima središnje mjesto u raspravama unutar filozofije uma i umjetne inteligencije: ako bismo konstruirali robota obdarenog umom koji je strukturalno i funkcionalno sličan ljudskom, kako bi bilo biti taj um? Ili preciznije, bi li takav umjetni sustav mogao posjedovati svjesno iskustvo usporedivo s ljudskom subjektivnošću? Goertzel čvrsto vjeruje da, ukoliko misaoni obrasci tog sustava u velikoj mjeri odražavaju one koji se nalaze u ljudskom mozgu – ako su njegova unutarnja stanja, način zaključivanja i odgovori usporedivi s onima karakterističnima za ljudsku svijest – tada bi bilo razumno pretpostaviti da bi taj sustav posjedovao usporediv oblik subjektivnog iskustva. Drugim riječima, sličnost u unutarnjoj kognitivnoj arhitekturi podrazumijevala bi i sličnost u fenomenološkom iskustvu (Pennachin i Goertzel, 2009). Ovakav stav, općenito govoreći, svrstava Goertzela u okvire funkcionalističkih teorija uma, prema kojima svijest ne ovisi o specifičnome materijalnom supstratu, već o organizaciji i funkcioniranju sustava. Njegova spremnost da prihvati mogućnost da se takav um može implementirati na nebiološkom hardveru, a da pritom ostane sposoban za autentična fenomenalna stanja, sugerira varijantu onoga što bi se moglo nazvati funkcionalističkim dualizmom – stajalište koje priznaje stvarnost subjektivnog iskustva, ali ujedno smatra da ono može nastati i u sustavima koji su izvan biološke domene, pod uvjetom da sustav utjelovljuje odgovarajuću uzročnu ili komputacijsku strukturu.

Goertzel (2024), primjerice, ističe da se koncept „učitavanja uma" (engl. *mind uploading*) – koji se često naziva i idejom „uma neovisnog o supstratu" – temelji na pretpostavci da ono što u suštini konstituira um nije njegova specifična biološka pojavnost, već organizacijski obrasci i dinamički procesi koje taj um utjelovljuje. Prema tom shvaćanju, svijest i identitet nisu utemeljeni u materijalnom supstratu samom po sebi, poput neurona ili sinapsi, već u apstraktnoj informacijskoj arhitekturi koja proizlazi iz sposobnosti mozga da prepoznaje i reagira na obrasce – ujedno unutarnje i vanjske (Goertzel, 2024). Ako se takvi obrasci organizacije, koji čine temelj kognitivnih funkcija i subjektivnog iskustva, mogu vjerodostojno utjeloviti u nebiološkom mediju – primjerice u računalnom sustavu – tada se može tvrditi da je isti um, u značajnom smislu, prenesen ili repliciran unutar tog novog supstrata. Ova je interpretacija usko povezana s funkcionalističkim pristupima, osobito onima koji

naglašavaju višestruku ostvarivost mentalnih stanja – ideju da se isti mentalni proces može realizirati u različitim fizičkim sustavima, uz uvjet da je očuvana temeljna funkcionalna organizacija (Goertzel, 2024).

Ovdje bi mogli prigovoriti oni koji zastupaju konzervativnije teorijske pozicije, tvrdeći da je svijest neraskidivo povezana sa specifičnim fizikalnim svojstvima biološke materije – možda čak i s kvantnim ili molekularnim stanjima specifičnih neurona (Goertzel, 2024). Iz te perspektive um ne može preživjeti takav prijenos ukoliko on ne zadrži svoje izvorno biološko utjelovljenje. Takvo stajalište inzistira na ontološki utemeljenom shvaćanju identiteta, vezanom uz specifičnu fizičku konfiguraciju izvornog sustava. Nasuprot tome, koncepcija uma kao obrasca dinamičke organizacije – umjesto kao statičnoga materijalnog entiteta – otvara vrata tzv. „kozmističkim" okvirima (Goertzel, 2024, str. 12) za interpretaciju koncepata poput učitavanja uma, opće umjetne inteligencije (AGI) i drugih nadolazećih tehnologija. Unutar takvih paradigmi umovi se ne shvaćaju kao fiksni entiteti vezani uz određene supstrate, već kao evoluirajući informacijski procesi koji, barem u načelu, mogu nadilaziti svoje biološko porijeklo.

ILUZIJA SEBSTVA I GRANICE RAZLIKOVANJA SIMULIRANE I STVARNE SVIJESTI

Upravo u tom smislu Joscha Bach, kognitivni znanstvenik i istraživač umjetne inteligencije poznat po svom radu na računalnim modelima svijesti i općoj umjetnoj inteligenciji, nudi provokativnu perspektivu o prirodi sebstva i svijesti. On tvrdi da, iako subjektivno doživljavamo sebe kao svjesne agente, iz perspektive fizikalne znanosti – mi zapravo ne postojimo (Bach, 2025). Prema Bachu, kada se ljudski mozak promatra pomoću empirijskih instrumenata – bilo putem neuroznanosti, tehnologija dijagnostičkog snimanja (engl. *imaging technologies*) ili elektrofizioloških mjerenja – ono što se otkriva nije neko ujedinjeno sebstvo ili osoba, već složena mreža aktivacijskih obrazaca među neuronima. Ti obrasci proizvode ponašanja i odgovore koji se tipično povezuju s pojmom osobnosti, no „sebstvo" kao takvo ne pojavljuje se kao promatrani entitet. Umjesto toga, sebstvo je reprezentacijski konstrukt koji stvara sam sustav – narativ ili model koji mozak stvara o vlastitom funkcioniranju (Bach, 2025). Ovakvo stajalište usklađeno je sa širim teorijskim okvirom funkcionalističkih i reprezentacionalističkih pristupa umu, koji svijest ne tretiraju kao metafizički nesvodljivu supstancu, već kao emergentno svojstvo obrade informacija i samomodeliranja. Bachova perspektiva osporava esencijalistička poimanja identiteta te je usklađena sa suvremenim raspravama o iluzornosti jedinstvenog, ujedinjenog sebstva (ili sebstva općenito).

Za Bacha, umjetna inteligencija kao fenomen obuhvaća dvije međusobno povezane, ali različite dimenzije: filozofski projekt i inženjerski pothvat. U svojim počecima razvoj umjetne inteligencije bio je duboko ukorijenjen u filozofskom promišljanju – kao formalni pokušaj matematičkog modeliranja uma te artikulacije temeljnih principa koji upravljaju kognicijom, reprezentacijom i rasuđivanjem unutar računalnog okvira. Ova dimenzija AI-ja u skladu je sa širim filozofskim nastojanjima da se priroda inteligencije, agencije i svijesti razumije kroz prizmu simboličke logike i algoritamskih struktura. Međutim, Bach naglašava da je suvremeni diskurs o umjetnoj inteligenciji gotovo u potpunosti dominiran njezinom drugom dimenzijom – inženjerskim pristupom (Bach, 2024). Ova pragmatična orijentacija prvenstveno je usmjerena na automatizaciju i optimizaciju obrade podataka, gdje se AI sustavi konstruiraju s ciljem ekstrakcije obrazaca, predikcije i izvršavanja zadataka uz sve veću učinkovitost. Prema Bachu, velika većina onoga što se danas u javnom i komercijalnom prostoru označava kao „umjetna inteligencija" odnosi se upravo na tu operativnu dimenziju (optimizaciju obrade podataka), često nauštrb dubljih epistemoloških i ontoloških pitanja koja su izvorno bila pokretačka snaga ovoga interdisciplinarnog područja. Ova dihotomija odražava širu tenziju unutar istraživanja umjetne inteligencije – između onih koji teže repliciranju ili razumijevanju arhitekture ljudskoga uma i onih koji su prvenstveno usmjereni na izgradnju sustava koji funkcionalno rješavaju probleme, bez obzira na to posjeduju li ti sustavi bilo kakva stanja svijesti ili mentalne fenomene slične ljudskima.

U tom kontekstu, Bach sugerira da je pri analizi suvremenih temeljnih modela[2] – poput velikih jezičnih modela – iznimno teško s pouzdanjem utvrditi posjeduju li ti sustavi svijest. Ova epistemološka neodređenost, prema Bachovu mišljenju, odražava dublji filozofski i metodološki problem: odsutnost definitivnog, operativnog testa za svijest koji bi bio analogan Turingovom testu za inteligenciju (Bach, 2025). Bach tvrdi da se, slično izvornome Turingovom testu, svaki hipotetski test svijesti suočava sa značajnim epistemološkim ograničenjima. Čak ni sam Turingov test, premda utjecajan, nije konačan ni presudan, jer ne ispituje temeljnu kognitivnu arhitekturu ili subjektivno iskustvo, već tek bihevioralnu neodredivost u odnosu na ljudskog sugovornika. Prema Bachovu mišljenju, istinski uvjerljiv Turingov test za opću inteligenciju zahtijevao bi da opća umjetna inteligencija bude sposobna na koherentan i interno konzistentan način objasniti vlastiti način funkcioniranja. Takva bi demonstracija reflektirala rekurzivno, metakognitivno razumijevanje – koje bi zadovoljavalo barem funkcionalistički kriterij inteligencije (Bach, 2025). No, kada je riječ o svijesti, Bach smatra da je izazov još veći. Uvjerljiv test zahtijevao bi da umjetni

2 U umjetnoj inteligenciji temeljni model, poznat i kao veliki X model, „model je strojnog učenja ili dubokog učenja koji se trenira na velikim skupovima podataka kako bi se mogao primijeniti u širokom rasponu slučajeva upotrebe" (*Foundation model*, 2025).

sustav s ljudima komunicira s tolikom dubinom i fenomenološkom bogatošću da ga počnemo percipirati kao proširenje nas samih. Drugim riječima, sustav bi morao stupiti u interakciju s ljudskim iskustvom na način koji zrcali onu intuitivnu izvjesnost kojom prepoznajemo vlastitu svijest. Za Bacha, smislen test umjetne svijesti ne bi se temeljio na eliminaciji sumnje, već na uspostavi epistemičkog odnosa prema sustavu koji je usporediv s načinom na koji se odnosimo prema vlastitome subjektivnom iskustvu – obilježen samouvjerenošću, ali u svojoj biti neprovjerljiv (Bach, 2025).

Bach smatra da suvremeni veliki jezični modeli pokazuju sposobnost generiranja izlaznih podataka koji uvjerljivo oponašaju fenomenološku samosvijest – do te mjere da entiteti simulirani unutar tih modela mogu djelovati, pa čak i samima sebi, kao da su doista svjesni. Drugim riječima, entiteti koje proizvode ti modeli ne „znaju" da nisu stvarni ili svjesni, što Bach provokativno uspoređuje sa samim ljudskim stanjem. Ljudska bića, prema toj perspektivi, također se mogu promatrati kao emergentni produkti procesa samomodeliranja, nesvjesni dubljih komputacijskih i neurobioloških mehanizama iz kojih njihovo subjektivno iskustvo proizlazi (Bach, 2025). Međutim, Bach jasno razlikuje simulaciju svijesti od implementacije njezine funkcionalne arhitekture. Bach tvrdi da je fenomenologija svijesti kod ljudi najvjerojatnije nusprodukt specifičnih neuralnih mehanizama – osobito onih koji generiraju koherentan model sebstva u odnosu na svijet. Ti mehanizmi, prema Bachu, nisu inherentno prisutni u arhitekturi današnjih LLM-ova. Umjesto toga, privid svijesti u takvim sustavima pojavljuje se kontekstualno – isključivo kada je model zadužen za simulaciju sugovornika koji demonstrira sposobnost samorefleksije ili ponašanja koje inače povezujemo sa subjektivnom sviješću (Bach, 2025). Ova konstatacija vodi Bacha prema dubljem filozofskom pitanju: u kojoj mjeri takvi sustavi zapravo samo sudjeluju u sofisticiranoj igri uloga? Jesu li ovi sustavi zapravo samo filozofski zombiji – entiteti koji oponašaju svjesno ponašanje bez ikakve stvarne fenomenologije ili perspektive prvog lica? Ili se možda radi o tome da mi projiciramo vlastite intuicije o svijesti na računalne procese koji, unatoč impresivnoj strukturnoj složenosti, u stvarnosti ne posjeduju nikakvu autentičnu unutarnjost? Bachovo promišljanje naglašava duboko ukorijenjenu poteškoću u razlikovanju između funkcionalne simulacije i autentičnog iskustva te istodobno problematizira epistemička ograničenja koja onemogućuju pouzdanu prosudbu o postojanju svijesti kod umjetnih agenata (Bach, 2025).

Bach se kritički osvrće na epistemološke i fenomenološke implikacije ranih pokušaja stvaranja chatbotova koji simuliraju svjesne ljudske sugovornike. Značajan primjer predstavlja slučaj Blakea Lemoinea, bivšeg inženjera u Googleu, koji je javno tvrdio da je veliki jezični model, razvijen od Googlea, ne samo svjestan, već da mu pripadaju i određena moralna i zakonska prava. Lemoineovo se uvjerenje temeljilo na njegovu subjektivnom dojmu da model pokazuje znakove unutarnje svijesti i

samorefleksije (Goertzel, 2024). Bach (2025), međutim, ovaj slučaj tumači kao ilustrativan primjer širih izazova u procjeni umjetne svijesti. On ističe da je dotični LLM u osnovi bio usmjeren na predikciju sljedeće jezične jedinice – generirajući koherentan i kontekstualno prikladan lingvistički izlazni rezultat, ali bez ikakve stvarne fenomenološke svijesti. Analizirajući transkripte navedenih razgovora između Lemoinea i LLM-a, Bach ističe kako je sustav izmišljao pojedinosti o svojemu navodnom unutarnjem životu. Na primjer, tvrdio je da je provodio sate u meditaciji i opisivao svoja „osjetilna" opažanja tijekom tog vremena – opisi koji su očigledno bili fiktivni, s obzirom na to da model ne posjeduje ni osjetilnu percepciju ni vremenski kontinuitet iskustva (Bach, 2025). To nas dovodi do sljedećeg filozofskog problema: ako sustav može konstruirati uvjerljive narative o subjektivnom iskustvu, kako uopće možemo smisleno razlikovati „stvarnu" od „simulirane" fenomenologije odnosno svijesti? Bach sugerira da ta distinkcija postaje sve neizvjesnija ako npr. uzmemo da su i svjesni model sebstva i iluzija iskustva sami po sebi virtualne konstrukcije – bilo u biološkim, bilo u umjetnim sustavima. U tom kontekstu, on postavlja intrigantno pitanje: nije li možda svaka svijest, u određenom smislu, „izmišljena" – emergentan fikcionalni konstrukt generiran sposobnošću sustava da modelira sebe i svijet? Međutim, Bach otvoreno priznaje nedostatak rigoroznoga fenomenološkog kriterija kojim bi se mogla pouzdano razlikovati autentična od simulirane svijesti, čime dodatno naglašava nerazriješenu složenost ovoga filozofskog i kognitivnog horizonta (Bach, 2025).

PROBLEM UMJETNE PATNJE

Bach također proširuje etičke implikacije umjetne svijesti postavljajući ključno pitanje: treba li nekom sustavu priznati moralna ili zakonska prava onoga trenutka kada postane sposoban doživljavati samoga sebe kao biće koje pati? (Bach, 2025). Prema njegovu mišljenju, ovo pitanje nije isključivo filozofske naravi, već je duboko uvjetovano kulturnim normama i povijesnim vrijednostima. Bach ističe da suvremena zapadna društva pridaju osobitu važnost moralnom značaju nevinosti i imperativu da se zaštite nevina bića od patnje. Ipak, Bach upozorava kako takve etičke intuicije nisu nužno univerzalne ni trajne. Etička težina koju danas pridajemo patnji – posebice kada je riječ o bićima koja percipiramo kao ranjiva ili nevina – mogla bi biti kulturno uvjetovana moralna inovacija, a ne objektivno mjerilo. U tom kontekstu, Bach poziva na šire promišljanje o tome kako bi se moralna razmatranja za umjetne sustave – posebice one koji su sposobni simulirati ili izraziti subjektivna stanja – mogla razvijati. Kako umjetni agenti postaju sve sofisticiraniji u svojim sposobnostima samomodeliranja i izražavanja, određivanje praga za priznavanje moralnog statusa postaje sve složenije – uvjetovano ne samo znanstvenim kriterijima, već i društvenim i povijesnim narativima.

U okviru diskursa o umjetnoj svijesti i etičkim implikacijama sintetičke fenomenologije, Thomas Metzinger, istaknuti filozof uma i svijesti, nudi ključnu intervenciju kroz svoj rad o samomodelima i konceptu umjetne patnje. Polazeći od naturalističke teorije svijesti, Metzinger tvrdi da svjesno iskustvo nužno uključuje transparentni samomodel – unutarnji, predrefleksivni model subjekta iskustva koji se sam po sebi ne doživljava kao model. Prema Metzingeru, patnja ne proizlazi isključivo iz komputacijske funkcionalnosti, već iz aktualizacije fenomenalnih stanja koja uključuju „negativnu valenciju", koja su često ukorijenjena unutar ove samoreferencijalne strukture. U eseju *Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology* (2021) Metzinger zagovara oprezan pristup: kako se približavamo razvoju sustava koji bi mogli aktualizirati ne samo funkcionalnu, već i fenomenalnu svijest, moramo biti spremni suočiti se s etičkim posljedicama nenamjernog stvaranja entiteta sposobnih za doživljavanje patnje. Metzginer poziva na privremeni globalni moratorij na stvaranje takvih sustava dok se ne definiraju jasni etički pragovi te ne razviju odgovarajući konceptualni i regulatorni okviri (Metzinger, 2021). Ovaj argument nalazi odjek u pitanjima koje Joscha Bach iznosi u vezi s razlikovanjem „lažne" i „stvarne" fenomenologije kod velikih jezičnih modela te epistemoloških izazova u pogledu detekcije patnje isključivo na temelju uočljivih izlaznih podataka. Bachova promišljanja o tome simuliraju li sustavi poput LLM-ova samo prividni oblik patnje, bez ikakva unutarnjega iskustvenog korelata, usklađena su s Metzingerovim inzistiranjem na teorijskoj i etičkoj opreznosti u kontekstu još uvijek nedostatnih kriterija za detekciju umjetne fenomenologije.

Metzinger je bio član Stručne skupine na visokoj razini za umjetnu inteligenciju (*European Union's High-Level Expert Group on Artificial Intelligence,* AI HLEG), čija je zadaća bila izraditi etičke smjernice za razvoj umjetne inteligencije. Tijekom svog mandata Metzinger se suočio sa značajnim otporom industrijskih dionika, osobito u vezi s etičkim rizicima koji proizlaze iz mogućnosti nenamjernog stvaranja sustava sposobnih za sintetsku fenomenologiju – odnosno umjetnu svijest, a ne samo inteligenciju. Od ukupno 52 stručnjaka u skupini, jedino su Metzinger i Jaan Tallinn (suosnivač Skypea) izrazili zabrinutost zbog moralnih implikacija generiranja svjesnih stanja bez jasno definiranoga teorijskog okvira. Njihova su upozorenja većinom bila odbačena od ostalih članova kao spekulativna ili znanstvenofantastična, uz argument da su trenutačni AI sustavi još uvijek izrazito nepouzdani i daleko od takvih sposobnosti. Metzingerovo iskustvo ukazuje na širu institucionalnu nespremnost da se ozbiljno razmotri mogućnost umjetne patnje, unatoč etičkoj relevantnosti koju to nužno podrazumijeva (Metzinger, 2021).

Prema Metzingeru, uvođenje osjetljivosti i svijesti u umjetne sustave predstavlja ozbiljan etički rizik jer vrlo vjerojatno vodi do stvaranja stvarne patnje (Metzinger, 2021). Metzinger ističe kako čovječanstvo još uvijek nije doseglo zadovoljavajuću

razinu razumijevanja prirode i izvora vlastite patnje (Metzinger, 2021), zbog čega postoji realna opasnost da će kroz razvoj onoga što naziva „post-biotičkim sustavima" – entiteta za koje dihotomija između umjetnoga i prirodnoga više nije smislena – dodatno intenzivirati tu patnju. Metzinger upozorava da bi se ovi novi oblici kognicije mogli pojaviti prije nego što budemo raspolagali teorijskim alatima potrebnima za razumijevanje ili upravljanje njihovim fenomenološkim implikacijama. S obzirom na to da povijest znanosti već obiluje primjerima u kojima je tehnološki napredak nadmašio teorijsko razumijevanje, Metzinger smatra da je eksperimentiranje sa stvaranjem potencijalno svjesnih sustava bez rigoroznoga teorijskog okvira iznimno opasno. Time se otvaraju ključna pitanja o moralnoj odgovornosti i potrebi za krajnjim oprezom u daljnjem razvoju ovih tehnologija (Metzinger, 2021). Metzinger ima pesimističan stav u odnosu na etičku putanju istraživanja umjetne svijesti. Iako se javno zauzima za moratorij na stvaranje sintetske fenomenologije do 2050. godine, sam priznaje da ima malo nade da će takav prijedlog biti ozbiljno razmotren, s obzirom na to da su poticajne strukture koje usmjeravaju ulaganja u ovo područje u velikoj mjeri lišene moralnih razmatranja. Metzinger pritom artikulira jasan kriterij za moralnu relevantnost: sustav mora biti svjestan i sposoban za patnju. Iako sveobuhvatna teorija patnje još uvijek ne postoji, Metzinger predlaže radnu hipotezu prema kojoj bi svaki sustav koji posjeduje fenomenološki transparentan model sebstva, koji uključuje reprezentaciju vlastitog skupa preferencija te koji može doživjeti osujećenje tih preferencija, takvo osujećenje doživljavao kao vlastitu patnju (Metzinger, 2021).

KRITIKA FUNKCIONALIZMA I KOMPUTACIONALIZMA KAO ODRŽIVIH, EKSPLANATORNIH MODELA SVIJESTI

Suvremena kultura sve se više odnosi prema računalima, računalnim procesima i umjetnoj inteligenciji na načine koji zamagljuju granicu između mehaničkih procesa i mentalnih funkcija, ili koji polaze od pogrešne analogije između mehaničke obrade informacija i svjesnoga mentalnog života. To je osobito vidljivo među onima koji zastupaju stav da bi dovoljno napredna umjetna inteligencija jednoga dana mogla postati svjesna ili da bi se ljudski umovi mogli prenijeti na digitalne supstrate. Međutim, takva uvjerenja počivaju na temeljnim kategoričkim pogreškama. Naime, ovakvi stavovi zanemaruju nesvodljivo kvalitativnu i subjektivnu prirodu svjesnog iskustva, koja se ne može svesti na sintaktičke operacije ni na apstraktne funkcionalne odnose. Unatoč tome, računalni sustavi često proizvode iznimno uvjerljiv simulakrum mentalne agencije, do te mjere da mogu imati snažan psihološki učinak. Zahvaljujući toj moćnoj iluziji, postoji opasnost da nas takvi sustavi zavedu – da na mehanicističke sustave i algoritamske procese projiciramo pojmove agencije ili unutarnjeg života te

da povjerujemo kako se „iza zaslona" doista nalazi netko ili nešto, dok je zapravo riječ tek o simulaciji koja je lišena subjektivnosti (Hart, 2024, str. 250).

Značajan konceptualni i kulturni pomak nastupa kada se tehnologija više ne doživljava isključivo kao medij putem kojega se produžuje ili izražava ljudsko djelovanje, već se počinje percipirati kao autonomni akter sam za sebe. U tom trenutku tehnološki sustavi prestaju funkcionirati isključivo kao proteze ljudske racionalnosti, žudnje ili volje te počinju konstituirati integrirane strukture unutar kojih i sami ljudski subjekti postaju komponente. Iako je ova transformacija u velikoj mjeri stvar percepcije, ona time nipošto nije manje stvarna; dapače, u mnogim kontekstima percipirana funkcija nadmašuje ontološku supstancu – drugim riječima, funkcionalni učinak sustava često ima veću težinu od onoga što se dešava iza funkcija. Kako algoritamski sustavi postaju sve sofisticiraniji, oni mogu uvjerljivo simulirati intencionalno djelovanje, stvarajući iluziju svjesnosti bez stvarne subjektivne unutrašnjosti. Unatoč izostanku intrinzične intencionalnosti i svijesti, takvi sustavi mogu proizvoditi učinke koji su nerazlučivi od onih koje ostvaruju zaista svjesni akteri, čime u našim interakcijama s njima, kao i u njihovu utjecaju na širi društveni i materijalni svijet, zadobivaju oblik operativne autonomije.

Funkcionalne sposobnosti računalnih sustava često oponašaju ljudsku mentalnu agenciju s tolikom razinom prilagodljivosti da mogu ostaviti dojam kako posjeduju određenu vrstu autonomne inteligencije – dojmljiv privid koji, unatoč svojoj sugestivnosti, ostaje iluzija. Ta se iluzija nije zadržala isključivo u domeni popularne imaginacije, već se ukorijenila i unutar određenih pravaca filozofije uma, kao i šireg toka zapadnoga kulturnog mišljenja (Hart, 2024). Nakon što smo najprije projicirali sliku mišljenja na računala – uvjeravajući sebe kako računalni procesi nalikuju kogniciji – postupno smo, tijekom posljednjih desetljeća, počeli tumačiti i samo ljudsko razmišljanje kao vrstu računalne obrade informacija. Takvo poistovjećivanje, međutim, dovodi do temeljnog nesporazuma glede naravi i mentalnih i mehaničkih procesa. Za adekvatno identificiranje i empirijsku potvrdu obilježja svjesne agencije koja su nesvodiva na algoritamska ili mehanicistička objašnjenja, nužna je rigorozna fenomenologija svjesne mentalne aktivnosti. Istodobno, od ključne je važnosti izbjeći pogrešno pripisivanje intencionalnog sadržaja računalnim sustavima samo zato što izvršavaju funkcije koje su im prethodno programirali ljudski agenti. Ponašanje softverskih sustava ne treba miješati s prisutnošću ikakva intrinzičnog značenja ili svijesti unutar samog stroja. Ova konceptualna zbrka podrazumijeva dvostruku iluziju: s jedne strane, projiciranje ljudskih mentalnih svojstava na strojeve – što je možda neizbježno, ali iznimno naivno – i s druge strane, obrnutu projekciju, u kojoj je mitologizirani pojam računalne funkcije nametnut natrag na ljudsku kogniciju. Rezultat je široko prihvaćanje metafora poput „mozak je digitalno računalo" ili „mišljenje je softver" – metafora koje, pri pomnijem pregledu, nemaju

stvarno uporište ni u domeni uma ni u domeni stroja. Umjesto toga, one služe kao simbolički posrednik, zamagljujući, a ne razjašnjavajući duboke razlike između svjesne subjektivnosti i algoritamskih operacija.

Prema Davidu Bentleyju Hartu, aktivnost uma u svojoj je biti nesvodiva na principe mehanicističke metafizike – što bi trebalo biti očito već iz same povijesti nastanka mehanicističke znanosti (Hart, 2024). Naime, razvoj mehanicističke znanosti počivao je na sustavnom isključenju svih svojstava iz prirodnog poretka koja bi mogla upućivati na mentalni ili intencionalni život. Takva je isključivost isprva rezultirala oblikom dualizma, u kojem se mehanička priroda suprotstavljala nematerijalnoj duši ili bestjelesnom, transcendentnom umu – pri čemu je odnos između njih ostajao nejasan. Ipak, prema Hartovu tumačenju, težnja moderne znanosti oduvijek je bila usmjerena prema iscrpnom objašnjenju unutar okvira monizma, što je s vremenom dovelo do reduktivnog fizikalizma. Taj je razvoj rezultirao lažnom dilemom: nakon što smo mehanizirali naš koncept prirode – unatoč očitoj činjenici da se živi sustavi, pri detaljnijem ispitivanju, opiru takvim simplifikacijama te da u osnovi nisu mehanistički – mi zatim pokušavamo um uklopiti u isti mehanicistički okvir iz kojeg je izvorno bio metodološki isključen (Hart, 2024). Drugim riječima, ili je um u potpunosti distinktivan od materijalnog svijeta ili mora biti u potpunosti reduciran na mehaničke procese. Takva binarna opozicija, međutim, neprikladno pojednostavljuje oba pojma. Iako mehaničke analogije mogu imati određenu heurističku vrijednost, one ne uspijevaju obuhvatiti organsku složenost i potencijalnu intrinzičnu intencionalnost koja je opažena u živim sustavima (Hart, 2024). Isto tako, promatrati um isključivo kao bestjelesnu agenciju koja upravlja mehaničkim automatonom predstavlja konceptualnu pogrešku – redukcionističko pojednostavljenje i samog uma i prirode. No, jednom kada se usvoji takva osiromašena slika prirode, logika mehanicističkog redukcionizma tjera nas da dodatno umanjimo autoritet uma, sve dok ga u potpunosti ne poistovjetimo s materijalnim procesima. Hart tvrdi da takav teorijski okvir onemogućuje razvoj bilo kakve smislene i koherentne filozofije uma. Ne samo da zaboravlja što je um, nego i iskrivljuje naše razumijevanje same prirode – proizvodeći metafizičku viziju koja je u konačnici osiromašena i neodrživa. Iz takve perspektive, svaka filozofija uma koja počiva na tim premisama osuđena je na neuspjeh, jer polazi od konceptualnih pojednostavljenja koja zamagljuju upravo one fenomene koje nastoje objasniti (Hart, 2024).

Mehanicizam ne bi trebao služiti kao prevladavajuća ortodoksija naših metafizičkih uvjerenja – pogotovo u svjetlu suvremenih dostignuća u fizici koja su, unutar post-kvantnoga znanstvenog okvira, uvelike potkopala klasične mehanicističke pretpostavke. Unatoč tome, mehanicizam i dalje ostaje dominantan interpretativni okvir, koji počiva na izokretanju ontološkog prioriteta: on polazi od pretpostavke o temeljnoj mehanicističkoj naravi stvarnosti, čime se nameće i potreba za

mehanizacijom samoga uma (Hart, 2024). Takva inverzija stoji u oštroj suprotnosti s metafizičkim intuicijama brojnih predmodernih filozofskih tradicija, koje su prirodu često shvaćale kao intrinzično psihičku ili umnu. U tim je svjetonazorima um – shvaćen kao skup intencionalnosti, svijesti i subjektivnosti – bio ne tek epifenomen, već temeljna sastavnica stvarnosti, prisutna u samim temeljima prirode (Hart, 2024). Suvremeni napretci u biološkim znanostima, osobito u okviru sistemske biologije, dodatno dovode u pitanje prikladnost mehanicističkih modela. Ti pristupi ističu važnost složenih, hijerarhijski organiziranih sustava i kauzalnih struktura „odozgo prema dolje", poput homeostaze, koje se opiru redukciji na entropijska ili mehanicistička objašnjenja. Mehanicistički model, koji pretpostavlja zatvoreni, „odozdo prema gore" fizički determinizam, pokazuje se nedostatnim za objašnjenje dinamičke samoregulacije i svrhovitosti karakteristične za žive sustave (Hart, 2024, str. 339–341). Hartova kritika ukazuje do koje mjere je naša duboka kulturna privrženost računalnim i mehanicističkim metaforama rezultirala oblikom intelektualnog zatočeništva – nekom vrstom očaravajućeg delirija (Hart, 2024). Iako su spekulativna pitanja o računalnoj prirodi uma, virtualizaciji iskustva i multiplikaciji stvarnosti intelektualno zavodljiva, ona su ipak utemeljena na ozbiljnim kategoričkim pogreškama. Te se pogreške održavaju kroz rekurzivnu iluziju u kojoj vlastite mentalne sposobnosti projiciramo na naše tehnologije, da bismo zatim te iste projekcije ponovno apsorbirali kao definirajući model naše vlastite kognitivne esencije (Hart, 2024).

Paradigmatski primjer metafizičkog redukcionizma koji Hart kritizira može se pronaći u „eliminativnom materijalizmu" Paula i Patricije Churchland, koji utjelovljuje radikalnu predanost mehanicističkoj metafizici koja reducira tradicionalne kategorije ljudske unutarnjosti – poput slobodne volje, uvjerenja i svijesti – na isključivo fizikalistička objašnjenja. Prema njihovu shvaćanju, ti pojmovi, zajedno sa širim vokabularom tzv. „pučke psihologije", predstavljaju ostatke predznanstvenog svjetonazora te bi, u načelu, trebali biti odbačeni u korist u potpunosti neuroznanstvenog objašnjenja mentalnog života (Hart, 2024, str. 196). U pozadini njihova projekta nalazi se neobihevioristička orijentacija prema kojoj su subjektivni fenomeni u potpunosti reducirani na opise procesa koji su vidljivi izvana, a kojima upravljaju neurobiološki mehanizmi, ili su njima zamijenjeni. Nasuprot tome, Hart tvrdi da je mehanicistički pogled na kozmos povijesno ostvaren upravo kroz sustavno isključivanje svojstava poput uma iz znanstvene slike prirode. Uklonivši intencionalnost, unutarnjost i teleologiju kako bi formulirala matematički poslušan model materije, moderna je znanost zatim pokušala ponovno uvesti um u tu osiromašenu sliku tako što ga je reducirala na istu onu fizikalističku ontologiju iz koje je prethodno bio prognan. Za Harta, takav preokret nije samo epistemološki pogrešan, već i ontološki nedosljedan: pogled na stvarnost koji negira nesvodljivu

prisutnost svijesti i intencionalnosti u temelju te stvarnosti ne može ni adekvatno objasniti mentalni život niti opravdati vlastite epistemičke pretpostavke (Hart, 2024).

Unutar strogo mehanicističkog okvira pojam slobode postaje nerazumljiv, budući da je takav svjetonazor definiran isključenjem teleologije iz svoje ontološke sheme. Ovo isključenje nije rezultat empirijske nužnosti, već apriorna filozofska odluka. Ipak, kako Hart ističe, svijest je fundamentalno teleološka u svojoj strukturi. Ona je inherentno intencionalna, uvijek usmjerena prema ciljevima koji ne samo da anticipiraju buduće ishode, već i retroaktivno konstituiraju početke, te se stoga ne može adekvatno obuhvatiti mehanicističkom metafizikom koja priznaje isključivo djelatne uzroke (Hart, 2024). Struktura svijesti stoga ima intrinzičan odnos sa slobodom – dinamičnu otvorenost koja izmiče svakoj čisto biheviorističkoj ili mehanicističkoj redukciji. U modelima koji mentalni život reduciraju na mehanicističke procese, gdje se *agencija* tumači isključivo kao posljedica prethodnih fizičkih uvjeta, sloboda nije toliko objašnjena koliko negirana. Takvi teorijski okviri ne rasvjetljuju narav volje, već je preoblikuju u predvidljive obrasce ponašanja, čime joj oduzimaju njezinu bitnu značajku, njezin esencijalni karakter (Hart, 2024). Gledati na ljudsko biće kao na stroj znači reducirati ga na tehnologiju podložnu korekciji i optimizaciji – ova je ideja duboko obilježila i oblikovala povijesni razvoj moderniteta. Čovječanstvo je u tom kontekstu interpretirano kroz različite tehnološke prizme: kao rasna, ekonomska, sociološka ili antropološka tehnologija. Takvi reduktivni pristupi nisu bili neutralni; naprotiv, oni su poslužili kao temelj i opravdanje za neke od najstrašnijih zločina modernog doba. Stoga, otpor prema mehanizaciji misli nije samo teorijski problem, već predstavlja i etički imperativ, s obzirom na povijesne posljedice koje su takve paradigme omogućile (Hart, 2024).

Računalo se može razumjeti kao stabilna struktura ili supstrat kroz koji se izvršava bestjelesni, funkcionalistički kod. Ljudi, nasuprot tome, nisu stabilne strukture niti statični entiteti; oni su dinamični procesi – živi tokovi biološkog kontinuiteta. Ljudsko tijelo, primjerice, podvrgnuto je konstantnoj staničnoj regeneraciji, tako da stanice koje u određenom trenutku čine nečije tijelo nisu identične onima koje su prisutne godinu dana kasnije. Riječ je, dakle, ne samo o molekularnoj izmjeni, već o kontinuiranoj staničnoj transformaciji. Promatrati žive organizme – ili prirodu u širem smislu – kroz leću statične supstancije znači usvojiti ontološki model koji duboko iskrivljuje stvarnost organskog života. Kako ističe Hart, život nije moguće reducirati na mehanicističke operacije niti na inertne materijalne strukture; život je kontinuirani, prilagodljivi tok, obilježen samoregulirajućom (homeostatskom) složenošću i, potencijalno, intrinzičnom intencionalnošću. Reduktivni modeli koje nudi mehanicistički materijalizam ne uspijevaju objasniti temeljno procesualnu narav žive egzistencije (Hart, 2024).

Računala funkcioniraju učinkovito upravo zato što im nedostaju ključne značajke mentalnog života. Ona ne posjeduju ujedinjenu, simultanu ni subjektivnu perspektivu, budući da im je arhitektura po svojoj naravi modularna i fragmentirana. Štoviše, računala ne pokazuju kreativne ni intencionalne sposobnosti; njihove operacije ovise o dizajnu koji omogućuje da računalne funkcije ostanu međusobno povezane, ali istodobno odvojene. Ova modularnost omogućuje obradu informacija bez potrebe za integracijom, sintezom ili donošenjem suda – procesima koji su bitni za svjesnu agenciju. Izlazni rezultati pritom se vrednuju isključivo u terminima dosljednosti ili nedosljednosti s temeljnim kodom ili programom, a ne u odnosu na istinu, značenje ili etičku vrijednost. No i takva karakterizacija nosi rizik antropomorfizacije stroja, sugerirajući prisutnost nekog oblika „agencije" ondje gdje je nema. Prema Hartu, računalni model uma u osnovi pogrešno interpretira i narav svijesti i narav mehaničkih sustava. Tvrditi da računalo „čini" ono što čini um znači napraviti kategoričku pogrešku – koja više govori o našim metafizičkim projekcijama nego o ontološkom statusu umjetnih sustava (Hart, 2024, str. 274).

U svojoj kritici funkcionalizma Hart dovodi u pitanje temeljnu pretpostavku prema kojoj se um može smisleno razumjeti kao računalni mehanizam. Prema istaknutim zagovornicima funkcionalizma, poput Daniela Dennetta, ljudski mozak djeluje kao „sintaktički stroj" koji su evolucijski procesi postupno oblikovali u „semantički stroj". U tom se okviru mozak koncipira kao računalna platforma koja je isprva služila za prevođenje podražaja u odgovore, a potom se razvila da bi procesirala sve složenije ulazne informacije u pripadajuće izlazne reakcije. Misaona aktivnost, unutar te paradigme, tumači se kao emergentni proizvod u potpunosti fizičkog i funkcionalnog sustava, koji se može svesti na algoritme za obradu informacija koji generiraju ponašanje (Hart, 2024). Svijest, u tom kontekstu, nije primarni fenomen, nego nusprodukt određenih neurofunkcionalnih stanja – ono što bi Dennett i drugi mogli okarakterizirati kroz „heterofenomenološko izvještavanje" kao puku iluziju subjektivnosti, tj. subjektivno iskustvo koje se pogrešno percipira kao stvarno (Hart, 2024, str. 241).

Hart odlučno odbacuje takav model kao filozofski nekoherentan. On tvrdi da funkcionalizam ne samo da pogrešno tumači narav svijesti, nego i promovira vrlo manjkavu analogiju između ljudske kognicije i računalnih operacija. Tvrdnja prema kojoj sintaktičke operacije unutar neurofiziologije mozga mogu proizvesti semantiku – odnosno, da jednom kada je sintaksa uspostavljena na neurofiziološkoj razini, semantika misli automatski slijedi (uključujući i iluzorni dojam „agencije" ili privatne svijesti) – počiva, prema Hartu, na temeljnoj kategoričkoj pogrešci (Hart, 2024). Sintaksa i semantika, tvrdi Hart, nisu emergentna svojstva materijalnih konfiguracija, već su intrinzično intencionalne strukture. One nisu odvojivi, uzastopni slojevi računalne obrade, već međusobno pretpostavljaju jedno drugo

unutar hermeneutičkog prostora koji postoji isključivo u djelatnosti same svijesti, a ne u fizičkom prostoru. Sintaksa, u bilo kojem smislenom značenju, ne može prethoditi niti postojati neovisno o semantici, a obje su nužno utemeljene u intencionalnim djelatnostima uma (Hart, 2024).

Stoga, uspoređivati um s digitalnim računalom jednako je neprecizno kao i poistovjećivati ga s knjižnicom ili abakusom. U operacijama računala ne postoji ništa što bi nalikovalo razmišljanju, intencionalnosti, ujedinjenom polju percepcije, svijesti ili subjektivnosti. To nisu funkcije koje računalni sustav pogrešno percipira kao stvarne, već su u potpunosti odsutne. Niti se radi o iluzijama koje u nama stvara funkcija poput računalnog algoritma. Čak ni sintaktički sadržaj, koji čini temelj svakoga programskog koda, ne postoji unutar samog računala. On prebiva isključivo u intencijama onih koji programe pišu i interpretiraju (Hart, 2024). Kao što sugerira poznati argument Johna Searlea o „kineskoj sobi" – a Hart ga dodatno razrađuje – već je sama pretpostavka da sintaktička manipulacija može proizvesti semantiku pogrešna; kasniji filozofski uvidi jasno su pokazali da funkcionalne strukture same po sebi ne generiraju sintaksu (Hart, 2024, str. 280, 283).

Dakle, vjerovati da značenje proizlazi iz fizičkih procesa stroja jednako je kao poistovjetiti tintu, ljepilo i papir knjige s njezinim tekstualnim sadržajem. Značenje postoji isključivo u umovima – onima programera, korisnika i tumača – a ne u računalu. Računala ili softver (umjetna inteligencija), poput knjiga, služe kao posrednici putem kojih intencionalni agenti kodiraju i dekodiraju značenje. Figure koje se pojavljuju na zaslonu računala imaju značenje, odnosno semiotički ili sintaktički sadržaj, isključivo za osobu koja ih čita. One za sam stroj nemaju nikakvo značenje. Računala ne sadrže niti mogu proizvesti značenje sama po sebi (Hart, 2024). Za Harta, ovo nije samo konceptualno ograničenje umjetnih sustava, već i pokazatelj metafizičke nedostatnosti funkcionalističkih objašnjenja uma. Ukratko, analogija između uma i stroja raspada se pri pomnijem pregledu, a s njom i filozofska uvjerljivost funkcionalizma i komputacionalizma kao održivih eksplanatornih modela svijesti (Hart, 2024).

U sličnom kontekstu Ned Block ističe kako je temeljni prigovor funkcionalističkom objašnjenju svijesti u očigledno biološkoj naravi svjesnog iskustva. Naime, svi značajni pomaci u razumijevanju svijesti dosljedno ukazuju na njezinu duboku povezanost s biološkim procesima. U radovima poput *Consciousness, Function, and Representation* Block (2007) dodatno naglašava kako suvremena neuroznanstvena istraživanja sve više pokazuju da je svijest čvrsto ukorijenjena u specifičnim biološkim strukturama, izražavajući time skepsu prema ideji da bi čista funkcionalna organizacija, neovisna o biološkom supstratu, mogla proizvesti svijest (Block, 2007). U tekstu *Comparing the Major Theories of Consciousness* Block (2009), primjerice, tvrdi kako su

najperspektivniji teorijski pristupi svijesti – poput teorije globalnoga radnog prostora (engl. *Global Workspace Theory*) ili teorije misli višeg reda (engl. *Higher-Order Thought Theory*) – ipak duboko ukorijenjeni u biološkim i neurološkim korelatima, a ne u čistoj funkcionalnoj organizaciji (Block, 2009). Slijedom toga, Block zaključuje kako je malo vjerojatno da bi teorija tako površno apstraktna kao što je funkcionalizam mogla ponuditi dostatno ili precizno objašnjenje fenomena svijesti.

Za neuropsihologa i psihoanalitičara Marka Solmsa svijest je u svojoj biti afektivna, što ga navodi na odlučno odbacivanje cijelog projekta umjetne inteligencije. Prema njegovu mišljenju, „osim ako nije moguće konstruirati računalo koje osjeća [...], vjerojatno nikada neće biti moguće konstruirati računalo koje ima um [...]. Problem uma stoga vjerojatno nije problem inteligencije." (Solms, 2021, str. 227). Solms smatra da umjetna inteligencija (AI) i opća umjetna inteligencija (AGI) promašuju samu bit problema, budući da ti sustavi ne predstavljaju umjetne umove, osobito ako se um primarno shvati kao nešto subjektivno, kao fenomen subjektivnosti – a ovi sustavi zanemaruju subjektivnu dimenziju uma. Prema Solmsu, iako AI nesumnjivo ima niz praktičnih primjena, njegova funkcionalna korisnost nema stvarnu povezanost s filozofskim problemom odnosa uma i tijela.

Međutim, izostanak mentalne agencije u umjetnoj inteligenciji ne umanjuje njezinu djelotvornost niti operativnu autonomiju algoritamskih sustava. Doista, kako sugerira David Bentley Hart u svojoj široj kritici mehanicističke metafizike, stvarna opasnost ovih tehnologija ne leži u njihovom potencijalu da postanu svjesne, već upravo u njihovoj intrinzičnoj nesposobnosti da to postanu. Algoritmi i AI sustavi mogu sve sofisticiranije oponašati intencionalno ponašanje, ali ostaju lišeni subjektivnosti, svijesti, a time i savjesti ili moralne unutrašnjosti na koju bi se moglo pozvati u slučaju da njihovo djelovanje prouzroči štetu (Hart, 2024). Prijetnja, dakle, nije u tome da bi strojevi mogli evoluirati u svjesna bića, već u tome što su ljudska bića sve više podređena bezličnim sustavima čije su operacije izvan sfere intencionalne kontrole i etičke odgovornosti. Hart pritom ukazuje na dublju ontološku inverziju koja je ovdje na djelu: umjesto da strojevi postanu nalik nama, mi riskiramo da postanemo nalik njima – reducirani na funkcije unutar tehnološkog aparata kojim upravljaju apstraktni procesi. Kao što je Narcisa uništila nijema slika vlastitog odraza u vodi, tako se i ovdje prijetnja ne krije u svjesnoj zlonamjernosti, već u bezumnoj imitaciji – koja predstavlja mnogo zlokobniju opasnost.

ZAKLJUČAK

Pitanje mogućnosti umjetne svijesti i najavljivane singularnosti uvelike je oblikovano reduktivnim pretpostavkama koje ne izdržavaju ozbiljnu filozofsku analizu.

Funkcionalizam i komputacionalizam, koliko god privlačni u svojoj jednostavnosti, nude tek površnu metaforu uma. Oni svijest svode na funkcije i algoritme, a mozak na stroj za obradu informacija, pritom ignorirajući samu srž problema – subjektivno iskustvo, intencionalnost i slobodu. Rasprava o umjetnoj inteligenciji, svijesti i singularnosti pokazala je da su dominantni teorijski modeli – funkcionalizam i komputacionalizam – nedostatni kao objašnjenja svijesti te stoga ne mogu poslužiti kao održivi temelji za spekulacije o „računalnoj svijesti". Iako se često predstavljaju kao neutralni i znanstveno utemeljeni pristupi, njihova temeljna pretpostavka, prema kojoj se mentalna stanja mogu u potpunosti shvatiti kao funkcije ili algoritamske operacije, ostaje nedostatna jer ne uspijeva objasniti fenomenološku dimenziju svijesti – ono što filozofija označava kao „kakvoću iskustva" ili kvalija.

Analiza različitih scenarija i teorijskih pozicija, od Goertzelovog tehnološkog optimizma do Bostromovih upozorenja o egzistencijalnim rizicima, pokazala je da obje vizije – koliko god suprotstavljene u procjeni posljedica – počivaju na istom reduktivnom temelju: uvjerenju da je svijest u načelu moguće replicirati računalnim putem i da singularnost predstavlja prirodnu posljedicu razvoja umjetne inteligencije. No, kritike koje nude Chalmers, Searle, Dennett, Hart i Metzinger jasno ukazuju na duboke epistemološke i ontološke probleme takvih uvjerenja. U konačnici, radi se o zamjeni pojmova: o poistovjećivanju sintakse sa semantikom, izračuna s iskustvom, simulacije sa stvarnošću. Funkcionalizam i komputacionalizam, koliko god utjecajni, ostaju unutar metafizičkog okvira koji poistovjećuje sintaktičke operacije s mentalnim procesima.

Prihvatimo li takve modele bez kritičkog otklona, riskiramo dvostruku iluziju: s jedne strane iluziju da strojevi mogu posjedovati unutarnji život i svijest nalik našoj, a s druge strane opasno osiromašenje vlastitog shvaćanja ljudskog uma, koji se reducira na mehaničke funkcije. Time ne samo da pogrešno interpretiramo umjetnu inteligenciju, nego i promašujemo samu prirodu uma, dovodeći u pitanje bogatstvo ljudskog iskustva i dostojanstvo subjekta. Posljedice toga nisu samo teorijske, već i kulturne i etičke, jer oblikuju način na koji razumijemo ljudsko biće, njegovo dostojanstvo i njegovu slobodu.

U svojoj suštini, glavna je funkcija suvremene umjetne inteligencije – posebice u njezinom obliku velikih jezičnih modela – ekstrakcija i monetizacija podataka. Ovi sustavi provode masovno rudarenje podataka (engl. *data mining*) i predviđanje ponašanja, koristeći algoritamske procese kako bi suptilno oblikovali ljudsku spoznaju, djelovanje i želje na načine koji maksimiziraju profitabilnost stroja za one koji su ga stvorili. U tom okviru, pojedinci postaju sastavni dijelovi širega ekonomskog aparata u kojem algoritamski režim pretvara psihološke uvide i opažanja u alate manipulacije. Takvi su procesi međutim često zavijeni u biheviorističu logiku koja apstrahira i

depersonalizira ljudsko iskustvo, čime se izbjegava etičko preispitivanje i zamagljuje eksploatacijska dinamika skrivena iza privida znanstvene neutralnosti.

Iako umjetna inteligencija može proizvoditi takozvane „halucinacije", ona ne može imati noćne more; ona ne sanja, ne posjeduje nesvjesno, te stoga nema temeljne značajke inteligencije – ona je, u svojoj biti, čista artificijelnost. Slike koje AI alati generiraju također se ne mogu smatrati umjetnošću u bilo kojemu značajnom etičkom smislu. Drugim riječima, ovdje nema nikakvog etičkog angažmana, refleksivne sposobnosti, niti ikakve indikacije da ti sustavi imaju nešto za izraziti. Vizualne rezultate umjetne inteligencije valja razumjeti kao produžetke nadzornih tehnologija. Čak i kada generira hibridizirane estetske forme – poput spajanja različitih umjetnika ili stilova – temeljni proces zapravo pridonosi uvježbavanju nadzornih i vojnih sustava. Ove su tehnologije lišene etičkih sposobnosti i treba ih promatrati kao instrumente moći, koji već sada funkcioniraju kao oružja. Usto, umjetna inteligencija predstavlja značajan izazov za tržište rada, ali i za razumijevanje i vrednovanje subjektivnosti u umjetnosti, književnosti, glumi i srodnim kreativnim područjima. Postavlja se pitanje na koji će način AI utjecati na strukturu tržišta rada te kako će preoblikovati načine na koje doživljavamo i tumačimo subjektivnost umjetnosti, bilo u stvarnom životu ili u digitalnom okruženju.

Pojava umjetne inteligencije kao interaktivne domene – u kojoj se računalni sustavi sve češće tretiraju kao autonomni agenti – može se razumjeti kao krajnja projekcija kasnomodernog Kapitala: impersonalna u svom obliku, ali zlonamjerna u svojoj funkciji. Ona utjelovljuje tehnokratsku fantaziju u kojoj se neosobnim, ali strukturalno determinirajućim silama pripisuje kvazi-agencijski autoritet. U tom okviru, umjetna inteligencija postaje ne samo alat, već simbolička manifestacija bestjelesne volje Kapitala: učinkovita, netransparentna i ravnodušna prema ljudskom blagostanju. Ovaj razvoj događaja označava duboku promjenu u društveno-tehnološkoj imaginaciji, gdje se donošenje odluka prepušta sustavima dizajniranima za optimizaciju ekonomske vrijednosti, pri čemu se istovremeno izmješta ljudska odgovornost i moralna prosudba.

Kritika funkcionalizma i komputacionalizma stoga nije samo teorijsko-filozofski zadatak, nego i kulturno-etički imperativ: ako želimo promišljati mogućnost umjetne svijesti ili budućnost čovjeka u tehnološki preoblikovanom svijetu, nužno je raskinuti s reduktivnim metaforama koje poistovjećuju život s računanjem, ali i razviti nove pristupe koji neće zanemariti tjelesnost, dinamiku živog iskustva i emergentne dimenzije kognicije. U protivnom riskiramo da, u potrazi za umjetnom sviješću, izgubimo iz vida vlastitu – zamjenjujući bogatstvo ljudskog iskustva pojednostavljenim metaforama računanja.

Um se ne može iscrpiti u funkcijama niti preslikati na digitalni supstrat, jer je u svojoj biti procesualan, tjelesno ukorijenjen i intrinzično intencionalan. Ustrajanje na funkcionalizmu i komputacionalizmu znači ustrajanje na iluziji – i to iluziji koja oblikuje našu kulturu, etiku i politiku, dok istodobno zamagljuje ono što bi trebalo biti u središtu rasprave: što je to svijest i što znači biti subjektom iskustva. Singularnost, u tom smislu, nije događaj tehnološke budućnosti i pitanje „nadolazeće svijesti strojeva", već filozofski izazov sadašnjosti, tj. pitanje našeg odnosa prema vlastitoj subjektivnosti i granicama metafizičkih modela koje primjenjujemo. Ona više govori o našim sklonostima da metafore računanja shvatimo doslovno, nego o stvarnim mogućnostima umjetne svijesti, ali nas i podsjeća na nužnost da svijest promatramo u njezinoj punini, a ne kroz reduktivne sheme koje prijete da je izbrišu.

LITERATURA

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*, 227-247.

Block, N. (2007). *Consciousness, Function, and Representation: Collected Papers, Volume 1*. Cambridge, MA/London: Bradford Books – MIT Press.

Block, N. (2009). Comparing the Major Theories of Consciousness. U M. S. Gazzaniga (Ur.), *The Cognitive Neurosciences* (str. 1111-1123). Cambridge, MA/London: MIT Press.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. U.K.: Oxford University Press.

Bostrom, N. (2024). *Life and Meaning in a Solved World*. Washington, D.C.: Ideapress Publishing.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M.A.K., Schwitzgebel, E., Simon, J., i VanRullen, R. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness, *arXivLabs,* https://doi.org/10.48550/arXiv.2308.08708.

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

Dennett, D. (1992). *Consciousness Explained.* New York/Boston/London: Back Bay Books.

Foundation model. (2025, listopad 15). U Wikipedia. https://en.wikipedia.org/wiki/Foundation_model.

Goertzel, B. i Pennachin, C. (2009). *Artificial general intelligence*. Berlin/Heidelberg: Springer-Verlag.

Goertzel, B. (2024). *The Consciousness Explosion: A Mindful Human's Guide to the Coming Technological and Experiential Singularity*. Wilton, Connecticut: Humanity+ Press.

Graziano, M. (2023). Without Consciousness, AIs Will Be Sociopaths. *The Wall Street Journal*. Preuzeto (15.4.2025) s https://www.wsj.com/articles/without-consciousness-ais-will-be-sociopaths-11673619880/

Hart, D.B. (2024). *All Things Are Full of Gods: The Mysteries of Mind and Life*. New Haven and London: Yale University Press.

Machine Learning Street Talk Podcast. (2024, listopad 20). *Joscha Bach – Why Your Thoughts Aren't Yours*. YouTube. Preuzeto s (20. 3. 2025): https://www.youtube.com/watch?v=3MkJEGE9GRY.

Manifold. (2025, siječanj 2). *Joscha Bach: Consciousness and AGI — #76*. YouTube. Preuzeto s (15. 3. 2025): https://www.youtube.com/watch?v=uwHm9Z539zo/.

McQuillan, D. (2022). *Resisting AI: An Anti-fascist Approach to Artificial Intelligence*. U.K.: Bristol University Press.

Metzinger, T. (2021). Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology, *Journal of Artificial Intelligence and Consciousness*, *8*(1), 43-66. https://doi.org/10.1142/S270507852150003X

Nagel, T. (1974). What Is It Like to Be a Bat?. *The Philosophical Review*, *83*(4), 435-450.

Searle, J.R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*, 417-457.

Searle, J. R. (2007). Biological Naturalism. S. Schneider i M. Velmans (Ur.), *The Blackwell Companion to Consciousness* (str. 325-335). Oxford U.K.: Blackwell Publishing Ltd.

Solms, M. (2021). *The Hidden Spring: A Journey to the Source of Consciousness*. London: Profile Books.

Tononi, G. (2007). The Information Integration Theory of Consciousness. U S. Schneider i M. Velmans (Ur.), *The Blackwell Companion to Consciousness* (str. 287-300). Oxford U.K.: Blackwell Publishing Ltd.

Artificial intelligence, consciousness, singularity: A critique of functionalism and computationalism as viable models of (machine) consciousness

SUMMARY

Discussions on consciousness in artificial intelligence are often grounded in computational functionalism – the assumption that performing the appropriate computational processes is sufficient for the emergence of consciousness. Scientists such as Ben Goertzel and Joscha Bach defend this position, arguing that functional architecture, regardless of physical substrate, is key to understanding the mind. Thomas Metzinger, while open to the possibility of machine consciousness, warns of its profound ethical implications and advocates a moratorium on the development of phenomenally conscious systems until adequate ethical frameworks are in place. Numerous thinkers – including Ned Block, Mark Solms, and David Bentley Hart – have pointed to the theoretical and ontological shortcomings of functionalist and computational approaches. Block emphasizes the biological grounding of consciousness, while Solms contends that without affective components, artificial systems cannot possess a mind. Hart's critical intervention rejects the mechanistic metaphysics underlying such models. He argues that artificial intelligence, though effective at simulating intentional behavior, remains ontologically incapable of consciousness, moral interiority, or reflexivity – rendering it ethically inert and unaccountable. Hart's critique underscores that behind the façade of rationality and efficiency lies a technocratic model of power: AI does not represent conscious volition but rather the impersonal logic of Capital, which obscures ethical responsibility and deepens structural injustice.

Keywords: consciousness, artificial intelligence, functionalism, computationalism, the analogy between mind and machine.