

Izvirni znanstveni članek/Article (1.01)

Bogoslovni vestnik/Theological Quarterly 84 (2024) 4, 797—806

Besedilo prejeto/Received:09/2024; sprejeto/Accepted:10/2024

UDK/UDC: 111:004.89

DOI: 10.34291/BV2024/04/Gregorčič

© 2024 Gregorčič, CC BY 4.0

Rok Gregorčič

Razmerje med človeško in umetno

inteligenco v luči Habermasove teorije

The Relation between Human and Artificial Intelligence in Light of Habermas's Theory

Povzetek: Prispevek namerava raziskati razmerje med človeško in umetno inteligenco v luči teorije komunikativne racionalnosti. V prvem delu analize so predstavljena Habermasova teoretična izhodišča. Izpostavljena je predvsem Habermasova teoretizacija racionalnosti ter premik od monološke h komunikativni racionalnosti. V nadaljevanju se osredotočamo na pojav umetne inteligence (UI) ter na razlikovanje med induktivnimi in deduktivnimi sistemi umetne inteligence. V zvezi s tem osvetljujemo tudi zgodovinsko dogajanje v pogledu na prvi in drugi val raziskovanja UI v preteklem stoletju. Obravnavamo tezo, da je prehod od simbolne UI h konekcionizmu mogoče interpretirati v navezavi na Habermasov prehod od monološke h komunikativni racionalnosti. V zadnjem delu prispevek prikazuje, kako bi Habermasovo teorijo v odnosu med človeško in umetno inteligenco aplicirali še drugače. Pričujoča analiza s pomočjo Habermasove kritike liberalne evgenike oblikuje stališče, ki lahko pomaga pri trajni razmejitvi med človekom in UI.

Ključne besede: Jürgen Habermas, komunikativna racionalnost, umetna inteligenca, simbolizem, konekcionizem, genetska manipulacija

Abstract: The aim of this article is to investigate the relation between human and artificial intelligence in light of the theory of communicative rationality. In the first part, Habermas's theoretical starting points are presented. Primarily, the transition from a monological to a communicative form of rationality is emphasized. In the next step, the focal point is moved toward the distinction between inductive and deductive systems of AI. In relation to this, a historical examination of the first and second wave of AI research in the previous century is also presented. A thesis that interprets the transition from symbolic AI to connectionism in correlation to Habermas's transition from monological to communicative rationality is thoroughly analyzed. Towards the end of the article, another way of applying Habermas's theory to the relation between human and artificial intelligence is presented. Deriving from Habermas's criticism of liberal eugenics, a well-argued position emerges that could help us distinguish between humans and AI in the long run.

Keywords: Jürgen Habermas, communicative rationality, artificial intelligence, symbolism, connectionism, genetic manipulation

1. Uvod

Jürgen Habermas velja za enega najprodornejših sodobnih mislecev.¹ Njegova družbena in filozofska misel je znana po vsem svetu. Zasnoval je teorijo komunikativne racionalnosti, ki predstavlja pomembno teoretično podlago za sodobne oblike družbenega sobivanja, s tem pa nudi filozofsko ogrodje za nekatere aktualne demokratične oblike političnih praks. Habermas v sodobnem svetu nedvomno spada med najuglednejše filozofske avtoritete. Odmevnost njegove misli je po eni strani posledica njegove občutljivosti za dialoško vključevanje vseh, po drugi strani pa zavzemanja za racionalnost in univerzalnost. Lahko bi rekli, da Habermasova misel uteleša poskus, kako vrlino racionalnosti, izhajajočo iz razsvetljenstva, uglasti v sozvočje z vrlino komunikacije, ki je posebno mesto dobila v filozofski postmodernosti. V tej luči lahko razumemo tudi koncept komunikativnega delovanja, ki predstavlja temelj Habermasove teorije (Habermas 1984). Gre za obliko družbenega delovanja, pri kateri se racionalnost uresničuje na komunikativen način.

Za Habermasovo teorijo racionalnosti je značilen odmik od monoloških modelov racionalnosti, ki so temeljili bodisi na zavesti posameznika bodisi na strogi znanstveni dokazljivosti. Premik, ki ga v odnosu do razumevanja vidimo pri Habermasu, bi na poseben način lahko primerjali s premikom v raziskovanju umetne inteligence (UI): strokovnjaki in raziskovalci, ki se ukvarjajo z UI, so večinsko opustili proučevanje deduktivno zasnovanih modelov UI in se usmerili k razvoju induktivno zasnovanih modelov UI. Zastavlja se vprašanje, kako lahko premik v dojemanju racionalnosti, kot je opazen pri Habermasu, primerjamo s premikom na področju različnih modelov UI. Še bolj pereče vprašanje pa je, kako nam lahko Habermasova teorija pomaga pri iskanju oz. dojemanju pravega razmerja med človeško in umetno inteligenco.

2. Habermasov odmik od monološke racionalnosti in pozitivizma

Habermas je teorijo racionalnosti zasnoval na podlagi analize pomembnih teorij racionalnosti iz preteklosti. Obširni del svojega znanstvenega prispevka je oprl na pregledno obravnavo starih teorij racionalnost, pri vsaki izpostavil pomanjkljivosti in nemalokrat dodal svoj lastni predlog izboljšav (Habermas 1984, 1–7). Ta vzorec je posebej izrazit pri Habermasovi obravnavi Kanta. Kantova misel po mnenju mnogih v Habermasovi teoretizaciji velja za vodilen vpliv (Hovdelien 2011, 111).

¹ Prispevek je nastal v okviru raziskovalnega programa P6-0269 „Religija, etika, edukacija in izzivi sodobne družbe“, ki ga sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS).

Habermas je menil, da Kantova filozofija pri obrambi racionalnosti v mnogočem sicer predstavlja pomemben korak, vendar se v sodobnem svetu v soočenju s pluralno družbeno realnostjo ne izkaže zadosti učinkovito (Gregorčič 2023, 912). Zardi tega Habermas do Kanta izkazuje neke vrste ambivalenten odnos: po eni strani se na njegove ideje vedno znova sklicuje, po drugi pa v mnogih potezah Kantovo mišljenje močno presega ali pa se od njega vsaj zmerno distancira. Očiten primer najdemo v delu *Resnica in justifikacija*. Habermas je tu predstavil štiri domnevne ustreznice med lastno teorijo formalnega pragmatizma in Kantovo transcendentno filozofijo – idejo enotnosti sveta, postulat svobode, idejo vsepresegajočega uma in avtoriteto justifikacije (Habermas 2003, 87). Iz zapisanega je vidno, da Habermas v širšem smislu vsakega od teh pojmov v svojo teorijo vključuje, v ožjem smislu pa vsakega rekonstruira in vsebinsko prilagaja komunikativni paradigmi. Sam je svoj odnos do Kantove filozofije izrazil s pojmom detranscendentalizacija (175). Izraz po Habermasovih besedah pomeni »vključitev umnih subjektov v kontekst življenjskega sveta in po drugi strani v prepletenost znanja z govorom in delovanjem« (88). Na podoben način kot pri Kantu je Habermas utemeljil rekonstrukcijo mnogih drugih filozofskih in socioloških teorij. Lahko bi rekli, da je Habermas pretekle avtoritete s pridom izkoristil tako, da je v svojo teorijo vključil nekatere njihove ideje. Hkrati pa je pri vsaki izpostavil pomanjkljivosti, ki jih v svoji teoriji poskušal odpraviti. Tako je želel vzpostaviti prenovljeno teorijo družbene racionalnosti, ki jo odlikuje postmetafizični in komunikativni značaj.

Za Habermasovo teorijo racionalnosti je značilen premik h komunikativni paradigmi, kar pomeni, da pglavitna vsebina komunikativne racionalnosti postane proces komunikativnega delovanja, imenovan tudi deliberacija (premišljujoče odločanje): proces izmenjave argumentov med posamezniki, ki se udeležujejo diskurza o določeni skupni problematiki (1984, 27). Udeleženci med seboj vzpostavijo veljavnostne zahteve, ki postanejo temelj njihovega medsebojnega pripoznanja ter nujna podlaga za možnost sporazumne razrešitve zadeve (22). Veljavnostne zahteve pomenijo, da posameznik v diskurzu sodeluje s pričakovanjem, da bodo drugi udeleženci diskurza (tudi v primeru nestrinjanja) njegove argumente priznali kot veljavne. Ta proces je za Habermasa zelo pomemben zato, ker naj bi se skupaj zbrali ljudje najrazličnejših stališč, se udeležili diskurza in v sporazumu poskušali doseči skupno rešitev (Geiger 2009, 23). Glede na pluralnost družbenih nazorov, ki smo jim danes priča, se Habermasu to zdi edina resnično univerzalna pot reševanja družbenih problemov – pot, ki jo ljudje različnejših nazorov lahko sprejmejo kot način interakcije v sekularni družbi. Habermasu se zdi neustrezno, da bi družbene probleme reševali bodisi po načelu filozofije zavesti (znotraj posameznikovih umskih zmožnosti) bodisi po načelu znanstvenega pozitivizma (strogo vezano na znanstveno metodo). V Habermasovem dojemanju je racionalnost nekaj, kar posameznikov um presega, hkrati pa presega tudi ozkost znanstvenih dokazov. Ob vsem tem je Habermas zavezan ideji praktičnega razuma, ki se loteva konkretne družbene problematike. Pri praktičnem razumu gre za to, da razumevanja ne sme diktirati teorija, ampak mora biti plod skupnega procesa celotne družbe. Racionalnost je vezana na posameznike in njihovo diskurzivno dejavnost, hkrati pa slehernega po-

sameznika presega (Habermas 2002, 82). Stare teorije racionalnosti so medtem trdile, da je razumevanje pridobljeno na podlagi danih fiksnih izhodišč oz. temeljev – od tod izvira izraz fundacionalizem. Habermasova teorija je zastavljena kot alternativa fundacionalizmu (Higgitt 2011, 94). Predvideva namreč skupni univerzalni proces intersubjektivne interakcije – temu sledi dosega sporazuma, ki bo vsa morebitna izhodišča metodično in vsebinsko presegel.

Ker Habermasova teorija močno poudarja živeto prakso, se pojavljajo mnoga vprašanja, kako to teorijo aplicirati na spremenjene razmere digitalne sodobnosti. Mnogi interpreti pravijo, da kriteriji, po katerih Habermas presoja komunikacijo in javnost, zaradi radikalno spremenjene informacijske realnosti niso več primereni; gotovo lahko opazimo, da sodobne oblike digitalnega povezovanja – denimo na novih družbenih omrežjih – Habermasovih standardov univerzalne komunikacije ne izpolnjujejo. Marsikdo zato meni, da bi moral Habermas, da bi lahko ponudil večjo možnost za vključevanje novih oblik, svoje kriterije prilagoditi (Khan, Gilani in Nawaz 2012, 47). Nekateri celo namigujejo, da bi se moral Habermasov teoretični model radikalno preoblikovati – ker da je v aktualnem stanju od sodobnih pojavov, kot je internet, vedno bolj oddaljen (Salikov [s. a.], 92). Mnogi drugi interpreti pa vztrajajo, da se morajo Habermasovi standardi pri konkretni evalvaciji sodobnih tehnoloških pridobitev ohraniti in aplicirati. Teoretizacija univerzalne pragmatike naj bi bila torej primerna in potrebna. Uporabiti bi jo morali predvsem v procesu preverbe, ali vse te nove oblike komunikacije demokratični družbi res koristijo (Ritzi [s. a.], 59). Habermasova načela se npr. lahko izkažejo kot koristna pri presoji različnih oblik medijev (Fisher 2021, 12). Kot kaže, smo priča pereči dilemi: na kakšen način – če to sploh zmore – naj se teorija komunikativne racionalnosti sooči s tehnološkimi in družbenimi novostmi moderne digitalne dobe?

3. Oblike umetne inteligence v luči Habermasove teorije

Ena od možnih točk soočenja med Habermasovo teorijo in digitalno modernostjo je vprašanje umetne inteligence. Iz nekaterih pomembnih vidikov komunikativne teorije namreč lahko izpeljemo vsebinski odziv, ki ga ta v povezavi z umetno inteligenco ponuja. Lahko bi rekli, da tehnološki projekt umetne inteligence, ki smo mu priča v sodobnosti, oscilira okoli želje, da bi izdelali program, ki bi čim bolj posnemal pravo človeško inteligenco. Treba je povedati, da se informacijska tehnologija pri uresničevanju te želje poslužuje različnih načinov. Poznamo dva osnovna načina za razvijanje umetne inteligence. Deduktivni način si inteligenco prizadeva vzpostaviti na podlagi vnaprej danih premis (Takefuji in Yamada 2019). Premise oz. pravila tvorijo strukturo, ki se je model UI strogo drži in zato za učenje novih vzorcev ni odprt. Na tej točki – torej pri vprašanju možnosti učenja – se deduktivni modeli UI od induktivnih najbolj razlikujejo. Induktivni modeli imajo namreč vsaj minimalno zmožnost, da se v interakciji s svojim okoljem naučijo česa novega (Nikitenko 2005, 387). Zaradi tega induktivno UI pogosto povezujemo z izrazom strojno učenje (Bognar 2021, a5). Induktivni sistemi ne temeljijo več na

rigoroznih premisah (in niso omejeni z ozko strukturo pravil), pač pa na novih interakcijah, ki jih ustvarjajo v novih okoljih, v katerih se znajdejo. Ker je ustvarjanje novih povezav njihova ključna karakteristika, se je induktivno zasnovane UI oprijel izraz konekcionizem. Na drugi strani pa se deduktivno zasnovano UI pogosto označuje z izrazom simbolizem, saj temelji na izhodiščnih simbolih ali premisah, ki vzpostavljajo rigorozen okvir njenega delovanja (Steging 2018, 1).

Delovanje deduktivnih modelov UI je ljudem dosti lažje razumljivo kot delovanje induktivnih (9). Tudi ob pogledu v preteklost lahko opazimo, da je bil prvi val tehnološkega razvoja UI, ki se je začel v sredini 50. in končal v poznih 70. letih prejšnjega stoletja, osredotočen skoraj izključno na deduktivne modele (Higgitt 2011, 87). Razlog je, da koncept simbolizma temelji na enostavnejšem dojemanju inteligence: delovanje uma je skoraj izenačeno z delovanjem računalnika (88). Šlo je v bistvu za to, da so človekovo mišljenje želeli razložiti kot mehansko razporeditev, sestavljeno iz končnih struktur. Kmalu je postalo jasno, da s takšnim pristopom dovolj učinkovitega približka človeški inteligenci ni mogoče izdelati. Prvi val raziskovanja UI je ugasnil, nastopil pa je drugi val, pri katerem je v ospredje stopil koncept statističnega učenja. Novi sistemi UI niso več delovali tako, da bi jih bilo treba z novimi pravili pripraviti prav za vsako novo izkušnjo, ki bi jih lahko doletela (Steging 2018, 1). Namesto tega je učenje potekalo na temelju velike količine podatkov; ni več veljala predpostavka, da je delovanje uma podobno delovanju računalnika. Konekcionizem je inteligenco namreč dojemal kot rezultat povezanosti dinamičnega vzorca omrežij – umevanje se dojemalo kot »nekaj fluidnega, kar vznikne znotraj kompleksnega reda« (Higgitt 2011, 89). Treba pa je poudariti, da je zaradi omenjene fluidnosti proces odločanja induktivne UI za laičnega opazovalca bistveno težje sledljiv ali razložljiv (Steging 2018, 1).

Ob pogledu na premik ob prehodu iz prvega v drugi val razvoja UI se na prvi pogled zdi, da se vse skupaj dobro ujema s Habermasovo teorijo racionalnosti. Kot kaže, je Habermasov premik od monološke h komunikativni racionalnosti našel svoj odmev v načinu razvijanja sodobne UI. Takšno mnenje zagovarja sociolog Ryan Higgitt, ki trdi, da konekcionizem in komunikativno racionalnost povezuje skupna namera: uresničenje človeškosti (Higgitt 2011, 96). Habermas to želi doseči skozi komunikativno delovanje, ki naj bi pripeljalo do univerzalnih moralnih norm (Habermas 2008, 272). Raziskovalci umetne inteligence pa pri svojem delu želijo razviti takšne modele, ki bodo čim bolj posnemali pristno človeško mišljenje in vedênje. Kot pravi Higgitt, oba premika – tako premik v teoriji racionalnosti kot v razvoju UI – odražata večjo ustreznost svetu, ki nas obdaja (Higgitt 2011, 97). Stična točka med konekcionizmom in Habermasovo teorijo je namreč priznanje fluidnosti (89). Ideja fluidnosti pri Habermasu temelji predvsem na nasprotovanju prevelikemu poudarjanju znanstvene metode (87). Na področju raziskav UI pa je fluidnost predstavljena kot izhod iz miselnosti, ki je človeški um enačila z delovanjem računalnika. Potemtakem bi lahko rekli, da gre tudi na področju UI za poskus zoperstavitve fundacionalizmu – za oblikovanje holističnega in zato tudi dinamičnega sistema mišljenja (84). Higgitt ugotavlja, da so sodobnejši modeli UI bolj v skladu s Habermasovim sistemom mišljenja (97). Še več: zdi se, da to ugotovitev

predstavlja kot razlog za pozitivno ovrednotenje drugega vala UI in komputacijske teorije uma (84). V zaključku Higgitt sicer poudarja, da bo najboljši razsodnik o novih sistemih UI dejanska praksa njihove uporabe (97). Vendarle pa njegovo pisanje vzbuja občutek, da domnevna povezava med komunikativno teorijo in konekcionizmom slednjemu znatno povečuje veljavo.

4. Princip nevtralnosti naključja v pogledu na človeškost

Interpretacija Higgita želi nakazati skladnost med konekcionizmom in teorijo komunikativne racionalnosti. Sam pa bi na tem mestu predlagal, da Habermasovo teorijo beremo z drugačnim poudarkom. Četudi se ujemanje med lingvističnim obratom in drugim valom UI na prvi pogled zdi prisotno (88), bi bilo treba zadevo raziskati podrobneje. Predvsem je treba večjo pozornost nameniti nekaterim Habermasovim prispevkom, ki bi lahko jasneje pokazali, kako se njegova teorija lahko odzove na pojav umetne inteligence. Tu se kot prvi ponuja prispevek z naslovom *Prihodnost človeške narave*, v katerem je Habermas prednostno obravnaval moralno dilemo genetske manipulacije (Habermas 2005). Razprava je pomembna, ker jo do neke mere lahko apliciramo tudi na problematiko umetne inteligence.

Habermas se v omenjeni razpravi postavlja na stališče, da je manipuliranje genetskega materiala pri človeku z vidika etike diskurza docela nesprejemljivo in zavržno dejanje (32). Etika diskurza namreč predpostavlja, da se diskurz oblikuje med posamezniki različnih stališč – in pravilno etično razumevanje v skupnem prizadevanju (19). Habermas pokaže, da nas k temu napeljuje že Kantov kategorični imperativ, ki v svojem bistvu zahteva prehod od ‚jaz‘ k ‚mi‘ (63). Gre torej za uveljavitev perspektive širšega diskurza, ki pa za svoj obstoj predvideva tudi določene pogoje. Ključen pogoj je gotovo to, da strinjanje ali nestrinjanje, ki ga izražajo posamezniki, resnično šteje. Moramo se zavedati, da odobravajoči ‚da‘ in odklanjajoči ‚ne‘ lahko štejeta samo v primeru, da posameznik, ki ju izrazi, za svojimi nameni resnično stoji (65). To pomeni, da se v posameznikovo notranje naziranje predhodno ne sme vmešavati nihče drug. V diskurzu nihče ne sme govoriti namesto nas – niti ne smemo dobiti občutka, da skozi naša stališča in argumente govori nekdo drug. Habermasovo komunikativno delovanje v svojem bistvu namreč predpostavlja nekakšno nevtralnost, ki posameznikom, vključenim v diskurz, zagotavlja enaka izhodišča. To nevtralnost Habermas vidi predvsem v nedotakljivosti odnosov medsebojnega priznanja (42–43). Prav medsebojno priznanje je tisto, ki iz človeka naredi osebo (43). V prepoznavanju pa ključno vlogo igrata ravno posameznikove nepopolnosti:

»Ker se človek v biološkem smislu rodi nepopoln in ostaja vse življenje odvisen od pomoči, pozornosti in priznavanja svojega družbenega okolja, postane nepopolnost individuacije v zaporedju DNK razvidna v trenutku, ko se začne proces družbene individuacije. Individuacija življenjske zgodbe poteka skozi podružbljanje.« (43–44)

Habermas se zato zavzema za koncept ranljivosti. Ranljivost, ki izhaja že iz posameznikovih telesnih in duševnih predispozicij, postane vstopna točka za družbeno integracijo, ki človeka dokončno oblikuje v osebo. Na tem mestu ponovno pride do izraza Habermasovo pojmovanje racionalnosti: racionalnosti pri človeku ne smemo iskati v tehnologizaciji njegove genetske strukture, temveč v integraciji človeka (z njegovo ranljivo genetsko strukturo vred) v širšo družbo. Človek namreč postane oseba z razumom šele takrat, ko je deležen pripoznanja (44). Pripoznanju nasproten pa je odnos polaščanja in programiranja. Programiranje pri starših ali znanstvenikih nastopi takrat, ko ne vstopajo v življenjsko zgodbo, kjer je avtor otrok sam (59). To pomeni, da ne dopuščajo njegovega avtonomnega in spontanega telesnega in duševnega razvoja. Posledica tega je, da programirani nikoli ne bo čutil, da je na enaki ravni kot programer (82–83). Nasproti liku programerja, ki se želi polastiti avtonomije drugega, Habermas postavlja lik zdravnika. Zdravnik v globljem pomenu besede je tisti, ki upošteva, da je vsak človek vir avtentičnih zahtev (63). Zdravnik sledi t. i. smotrnostni formuli, da je vsaka oseba obravnavana kot smoter že sama po sebi. Zaveda se, da drug drugemu dolgujemo spoštovanje pravice do oblikovanja lastne usode (Anderson 2005, 817). Lik zdravnika v Habermasovem dojemanju instrumentalizaciji povsem nasprotuje.

Prav predpostavka univerzalnega diskurza bi bila z uvedbo genetske manipulacije neobhodno ogrožena. Osebo, deležno tehnološko modificiranega genoma, bi spremljal občutek, da je njen DNK in s tem njeno osebnost priredil nekdo drug. Habermas se pri tem konkretnem vprašanju trdno zavzema za pozicijo mehkega genetskega determinizma (Morar 2015, 107). Če bi prišlo do manipulacije z genomom, bi se po Habermasovem mnenju identifikacija z lastnim telesom porušila, hkrati pa tudi identifikacija z lastnim mišljenjem, stališči in izjavami. Habermas trdno verjame, da bi takšna situacija možnost univerzalnega diskurza odločilno ogrozila. Zaradi tega neomajno vztraja pri stališču, da je nujno varovati kontinuiteto sebstva in zavest naravnega naključja (Habermas 2005, 68). Če pride do genetske fiksacije, pomeni, da se je tok naravnega naključja na neki točki življenja prekinil (70). Habermas meni, da to lahko povzroči grozovite posledice tako za osebno zavest avtonomije kot tudi za medosebne odnose (71). Iz tega primera se zopet vidi, da mora biti vsakršno tehnološko iskanje bližnjic podrejeno vrednoti avtonomije.

5. Utemeljena različnost

Kot vidimo, je eden izmed temeljnih predpogojev posameznikove osebne integritete naključje njegovih genetskih danosti. Habermas trdi, da je ravno naključnost tisti dejavnik, ki omogoča, da človek svoje življenje prevzema odgovorno (20). Naključje se kaže kot nujna predpostavka, da je človek lahko resnično sam svoj – in da posledično v medosebnih odnosih lahko nastopa enakopravno (21). Ključno je, da posameznik v svojem družbenem delovanju nastopa z lastnimi argumenti, za katerimi resnično stoji. Izhodiščno možnost za to mu daje prav naključnost njegove naravne genetske določenosti (20). Le če je genetski zapis res plod naključne

genetske geneze, je mogoče iz te naravne predpodlage izhajati odgovorno – in jo zares prevzeti kot lastno identiteto. Brž ko je vsejan dvom, da je s posameznikovo genetsko opremljenostjo manipuliral nekdo drug, nastane nova oblika medsebojnih odnosov, ki moralno občutje močno prizadene (21) – gre torej za dilemo, ki zadeva tudi etično samorazumevanje človeka (22). Poseg v globino organskih zasnov namreč v posamezniku lahko porodi občutek, da je sredi družbe, ki temelji na medsebojnem pripoznanju, tujek (21) – in iz istega razloga je prizadet tudi občutek za osebno odgovornost, ki je temelj moralnega življenja v družbi. Habermas dodaja še, da je človeku, odraščajočemu v običajnih razmerah, omogočen refleksivni odnos do tega, kar je: refleksivni odnos do procesa lastnega oblikovanja. V primeru genetske manipulacije pa bi bila zmožnost refleksije znatno okrnjena (22).

Če se vrnemo k vprašanju umetne inteligence, lahko vidimo, da Habermasova teorija poudarja predvsem razliko med delovanjem človeka in delovanjem UI. Če torej Higgitt osvetljuje zlasti vidike, v katerih se med komunikativno racionalnostjo in konekcionistično UI kaže podobnost, pa sam zatrjujem, da Habermasova teorija ob celostnem branju do pojava umetne inteligence razodeva drugačen odnos. Pri obravnavi genetske manipulacije namreč velik pomen pripisuje naključju, moralni odgovornosti in zmožnosti refleksije (20–23) – in treba je priznati, da pri UI ne moremo najti nobene od teh stvari. Tudi najsodobnejši sistemi umetne inteligence se niso razvili iz naključnih danosti, niso razvili avtonomnega moralnega čuta, prav tako niso razvili zmožnosti za refleksijo o procesu lastnega nastanka. Res je sicer, da konekcionizem zasleduje idejo, da se UI uči iz ustvarjanja novih povezav v novih situacijah. Toda vse to UI lahko počne zgolj iz vnaprej programiranih danosti, ki pa nikakor niso naključne. Tako deduktivna kot tudi induktivna UI delujeta na izhodiščih, ki smo jih vanju vstavili ljudje. Tako moramo pri vseh oblikah umetne inteligence princip nevtralnosti naključja odmisлити. Z odsotnostjo naključja je pri UI povezana tudi nezmožnost prve refleksije. Ta vidik posebej izpostavlja moralni filozof Paolo Monti, ki status moralnega subjekta povezuje z reflektivnim aspektom (Monti 2024, 64). Modelov UI torej nikakor ne moremo dojemati, kot da so odgovorni moralni subjekti. Tudi med najnaprednejšimi sistemi UI – Monti ima v mislih predvsem velike jezikovne modele – statusa moralnega subjekta ne moremo pripisati nobenemu. Dodaten razlog za to je tudi dejstvo, da noben model UI ne izhaja iz lastnega življenjskega sveta (72–74). Strahovnik k vsemu temu dodaja še krščanski vidik ustvarjenosti človeka po Božji podobi: za UI ne moremo trditi, da je ustvarjena po Božji podobi, kot to velja za človeka – zato tudi nima enakega statusa (Strahovnik 2023, 852). Z vidika komunikativne racionalnosti je mogoče naštet še več lastnosti, po katerih umetna inteligenca ostaja od človeške izrazito različna: sodobne oblike umetne inteligence niso zares zavezane resnici, ampak jih vodi ekonomski interes; ne upoštevajo Habermasovega načela pravosti, ker njihov razvoj temelji na digitalnem izkoriščanju revnih; vodijo v smer strateškega delovanja v smislu individualizma in utilitarizma (Westerstrand, Westerstrand in Koskinen 2024). Zato se poraja vprašanje, ali umetno inteligenco sploh lahko imamo za pravo inteligenco (Žalec 2023, 813). Odgovor na to vprašanje je odvisen od opredelitve pojma ‚inteligenca‘. Vendarle pa vidimo, da je pri

trajnem razlikovanju med človeško in umetno inteligenco prav Habermasova teorija lahko v veliko pomoč. Pomaga nam, da odločilne vidike, v katerih umetna inteligenca ne dosega človeške, odkrijemo in učinkovito obrazložimo.

6. Zaključek

Teorija komunikativne racionalnosti v odnosu do umetne inteligence poudarja zlasti različnost med človeško in umetno inteligenco. Med Habermasovimi interpreti sicer obstajajo poskusi, ki prehod od deduktivne k induktivni UI razlagajo po analogiji s prehodom od fundacionalizma h komunikativni racionalnosti. Toda sam menim, da poglobljeno branje Habermasa odkriva zlasti vidik razlike. Menim, da se Habermasova teorija v času, ko se porajajo različna ugibanja o tem, ali bi umetna inteligenca lahko postala enaka človeški (Centa Strahovnik 2023, 853), kaže kot koristna, saj daje dobre utemeljitve v smeri trajnega razlikovanja med človekom in UI. Pomen naključne danosti, avtonomne moralne odgovornosti, zmožnost samorefleksije, ukoreninjenost v življenjski svet itd. – vse to predstavlja močne temelje za široko zavest o jasni razliki med njima. Habermas nam torej lahko pomaga, da glede na sedanji in prihodnji tehnološki napredek ozavestimo distanco med človekom in UI. Habermas sicer tehnološkemu napredku samemu po sebi ne nasprotuje (Feenberg 1996, 49–50). A s pomočjo njegovih teoretičnih izhodišč se kot družba lahko zavežemo, da bomo distanco med človekom in UI na vseh ravneh še naprej strogo vzdrževali. Nedotakljivost naravnih danosti, sposobnost moralne odgovornosti in zmožnost refleksije o sebi moramo vedno obravnavati kot posebne vrednote, ki jih imamo samo mi kot osebe. Na teh področjih meje nikoli ne smemo prekoračiti ne v eno ne v drugo smer: sistemom UI tako ne smemo nikdar pripisati statusa odgovornega moralnega subjekta, obenem pa našega moralnega delovanja ne smemo prepustiti odločitvam UI. Ne glede na to, kakšne sposobnosti bo UI v prihodnosti razvila – za moralo in refleksijo bo moral vedno poskrbeti človek.

Kratika

UI – Umetna inteligenca.

Reference

- Anderson, Joel.** 2005. Review of Jürgen Habermas, 'The Future of Human Nature'. *Ethics* 115, št. 4:816–821.
- Bognar, Melinda.** 2021. Prospects of AI in Architecture: Symbolicism, Connectionism, Actionism. *Journal of Architectural Informatics Society*, a1–a20.
- Centa Strahovnik, Mateja.** 2023. Identiteta in pogovorni sistemi umetne inteligence. *Bogoslovni vestnik* 84, št. 4:853–864. <https://doi.org/10.34291/bv2023/04/centa>
- Feenberg, Andrew.** 1996. Marcuse or Habermas: Two Critiques of Technology. *Inquiry: An Interdisciplinary Journal of Philosophy* 39, št. 1:45–70. <https://doi.org/10.1080/00201749608602407>

- Fisher, Eran.** 2021. Epistemic media and critical knowledge about the self: Thinking about algorithms with Habermas. *Critical Sociology* 48, št. 7–8:1309–1324. <https://doi.org/10.1177/08969205211044193>
- Geiger, R. Stuart.** 2009. Does Habermas Understand the Internet? The Algorithmic Construction of the Blogosphere/Public Sphere. *Gnovis: A Journal of Communication, Culture, and Technology* 10, št. 1. <https://escholarship.org/uc/item/60s6s0p8> (pridobljeno 7. 8. 2024).
- Gregorčič, Rok.** 2023. Tehnološki razvoj v luči Habermasove etike diskurza. *Bogoslovni vestnik* 83, št. 4:911–922. <https://doi.org/10.34291/bv2023/04/gregorcic>
- Habermas, Jürgen.** 1984. *The Theory of Communicative Action*. Zv. 1. *Reason and the Rationalization of Society*. Boston: Beacon Press.
- . 2002. *Religion and Rationality: Essays on Reason, God, and Modernity*. Cambridge: Polity Press.
- . 2003. *Truth and Justification*. Cambridge: Polity Press.
- . 2005. *Prihodnost človeške narave. Verjeti in vedeti*. Ljubljana: Studia humanitatis.
- . 2008. *Between Naturalism and Religion: Philosophical Essays*. Cambridge: Polity Press.
- Higgitt, Ryan.** 2011. Habermas' Communicative Rationality and Connectionist AI. *Culture, Theory and Critique* 52, št. 1:83–100. <https://doi.org/10.1080/14735784.2011.621664>
- Hovdelien, Olav.** 2011. Post-Secular Consensus? On the Munich-dialogue between Joseph Ratzinger and Jürgen Habermas. *Australian Journal of Theology* 18, št. 2:107–116.
- Khan, Muhammad Zubair, Ijaz Shafi Gilani in Allah Nawaz.** 2012. From Habermas Model to New Public Sphere: A Paradigm Shift. *Global Journal of Human Social Science* 12, št. 5:43–51.
- Monti, Paolo.** 2024. AI enters public discourse: A habermasian assessment of the moral status of large language models. *Etica & Politica* 26, št. 1:61–80.
- Morar, Nikolae.** 2014. An Empirically Informed Critique of Habermas' Argument from Human Nature. *Science and Engineering Ethics* 21:95–113. <https://doi.org/10.1007/s11948-013-9509-5>
- Nikitenko, Agris.** 2005. Intelligent Agent Control Using Inductive, Deductive and Case Based Reasoning. *Information Theories & Applications* 12:386–395.
- Ritzi, Claudio.** 2023. The hidden structures of the digital public sphere. *Constellations* 30:55–60.
- Salikov, Alexey.** 2018. Hannah Arendt, Jürgen Habermas, and Rethinking the Public Sphere in the Age of Social Media. *Russian Sociological Review* 17, št. 4:88–102. <https://doi.org/10.17323/1728-192x-2018-4-88-102>
- Steging, Cor.** 2018. Explainable AI: On the Reasoning of Symbolic and Connectionist Machine Learning Techniques. Magistrsko delo. University of Groningen.
- Strahovnik, Vojko.** 2023. Etični in teološki izzivi velikih jezikovnih modelov. *Bogoslovni vestnik* 83, št. 4:839–852. <https://doi.org/10.34291/bv2023/04/strahovnik>
- Takefuji, Yoshiyashu, in Ren Yamada.** 2019. A new deductive method with intelligence using pseudorandom number for solving coin-weighing puzzles: review of inductive and deductive methods. Preprints. 8. 5. <https://www.preprints.org/manuscript/201905.0090/v1> (pridobljeno 9. 8. 2024).
- Westerstrand, Salla, Rauli Westerstrand in Jani Koskinen.** 2024. Talking existential risk into being: a Habermasian critical discourse perspective to AI hype. *AI and Ethics* 4:713–726. <https://doi.org/10.1007/s43681-024-00464-z> (pridobljeno 13. 8. 2024).
- Žalec, Bojan.** 2023. Ali je umetna inteligenca inteligenca v pravem pomenu besede? Vprašanje psihičnih značilnosti in Splošnost. *Bogoslovni vestnik* 83, št. 4:813–823. <https://doi.org/10.34291/bv2023/04/zalec>