Izvirni znanstveni članek/Article (1.01)
Bogoslovni vestnik/*Theological Quarterly* 83 (2023) 4, 853—864
Besedilo prejeto/Received:09/2023; sprejeto/Accepted:12/2023
UDK/UDC: 14:004.89:81
DOI: 10.34291/BV2023/04/Centa

Mateja Centa Strahovnik

Identiteta in pogovorni sistemi umetne inteligence
Identity and Conversational Artificial Intelligence

Povzetek: Prispevek obravnava spreminjajoče se dojemanje sistemov umetne inteligence (UI) v svetu, ki ga umetna inteligenca – s poudarkom na pogovornih sistemih UI ali klepetalnih robotih – vse bolj določa. Z razvojem človeku podobnih robotskih sistemov UI se povečuje potreba po razumevanju identitete, ki jo ti stroji s prevzemanjem vlog, prej rezerviranih za ljudi (npr. skrb za starejše ali izobraževanje in vzgoja otrok), prevzemajo. Osrednja teza prispevka je, da je za razumevanje identitete takšnih sistemov umetne inteligence treba upoštevati vidik naše lastne identitete, ki jo v interakcijah s temi sistemi oblikujemo in projiciramo. V dobi, ki jo opredeljuje umetna inteligenca, so raziskave identitete ter interakcij med umetno inteligenco in človekom izrednega pomena in so ključne tudi za nadaljnji razvoj umetne inteligence.

Ključne besede: umetna inteligenca, klepetalni roboti, veliki jezikovni modeli, etika, identiteta

Abstract: This paper explores the evolving perception of artificial intelligence (AI) systems in our increasingly AI-driven world, focusing on conversational AI or chatbots. With the rise of human-like robotic AI systems, there is a growing need to understand the identity these machines assume as they take on roles previously reserved for humans, such as caring for the elderly and educating children. The central thesis posits that comprehending the identity of such AI systems necessitates considering the aspect of our own identity that we shape and project onto them in our interactions with these systems. In an era defined by AI, these inquiries into identity and AI-human interactions are of paramount importance and are also vital for the very development of AI.

Keywords: artificial intelligence, chatbots, large language models, ethics, identity

1. Uvod

V prispevku nas zanima odgovor na vprašanje, kako dojemamo identiteto pogovornih sistemov umetne inteligence, s katerimi vse bolj prihajamo v stik.[1] Posebej pomembno se to vprašanje zdi ob možnosti sestavljenih sistemov, to je humanoidnih robotov in pogovornih sistemov umetne inteligence. Eden izmed ciljev razvoja teh sistemov je, da postanejo takšni, da bodo lahko prevzeli pomembna področja delovanja, ki so bila doslej v domeni ljudi. Tu velja omeniti zlasti področja skrbi za starejše, vzgoje otrok in druga področja, kjer je takšne asistenčne sisteme moč uporabiti (Spillane idr. 2019). V pomembnem smislu gre torej za delegiranje našega siceršnjega dela tem sistemom umetne inteligence, pri čemer to delo vse bolj postaja tudi osebno oz. čustveno delo (npr. skrb, ljubezen, sočutje; prim. Dorobantu idr. 2022). Gledano nekoliko ožje se v tem prispevku posvečamo pogovornim sistemom umetne inteligence oz. tako imenovanim klepetalnim robotom (angl. *chatbots*), za katere ni nujno, da jih spremlja določena vizualna upodobitev ali prisotnost. Teza, ki jo zagovarjamo, je, da moramo pri vprašanju dojemanja identitete takšnih in podobnih sistemov umetne inteligence nujno upoštevati tudi vidik naše lastne identitete, ki jo v interakciji s temi sistemi oblikujemo oz. prevzamemo.

2. Razsežnosti identitete in umetna inteligenca

Različne značilnosti sistemov umetne inteligence – vključno s pogovornimi sistemi – in pomen teh značilnosti v interakciji s človeškimi uporabniki so že kar nekaj časa predmet številnih študij. Eden izmed vidikov te interakcije je tudi dojemanje identitete teh sistemov in pomen te identitete, ki med drugim vključuje tudi to, kako ljudje te sisteme dojemamo npr. z vidika spola (West, Kraut in Ei 2019), rase (Yuting in He 2020), zaupanja, vrednosti (Kraus, Seldschopf in Minker 2021), prijaznosti, sočutnosti (McKee idr. 2021) itd. Ni pa osrednje vprašanje teh raziskav sama identiteta oz. se je ne lotevajo celostno. Zato v prispevku najprej izpostavljamo za to vprašanje osrednje vidike identitete. Posebej nas bodo zanimali vidiki moralne, spoznavne in psihološke identitete – v nadaljevanju podajamo kratko opredelitev teh vidikov.

Identiteto običajno razumemo kot zbir razlikovalnih značilnosti in lastnosti, ki posameznika ali skupino posameznikov opredeljujejo. Ta nabor vključuje različne vidike – vključno z osebnimi, družbenimi in kulturnimi gradniki, ki oblikujejo posameznikov občutek o sebi in njegovem mestu v svetu. V nadaljevanju izpostavljamo vidike identitete, ki jih z vidika interakcije s sistemi umetne inteligence razumemo kot posebej pomembne. Prvi vidik se dotika osebne identitete, ki zadeva

[1] Prispevek je nastal v okviru raziskovalnega projekta Z6-2666 „Kognitivna teorija čustev v kontekstu teologije čustev: telesni občutki, spoznanje in moralnost“, raziskovalnega programa P6-0269 „Religija, etika, edukacija in izzivi sodobne družbe“, ki ju sofinancira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije, ter v okviru raziskovalnega projekta „Epistemic Identity and Epistemic Virtue: Human Mind and Artificial Intelligence“ s podporo fundacije John Templeton (program New Horizons for Science and Religion in Central and Eastern Europe).

zlasti edinstvene lastnosti, izkušnje, prepričanja, vrednote in spomine, po katerih se oseba razlikuje od drugih – in po katerih se prepozna sama oz. jo prepoznavajo drugi. Vključuje številne gradnike in določila: ime, starost, spol, telesni videz, sposobnosti, značajske lastnosti ipd. Pri pogovornih sistemih umetne inteligence lahko hitro vidimo, da so določeni oz. jih določamo glede na omenjene gradnike: pogosto imajo ime, glas, spremlja pa jih določena upodobitev.

Drugi vidik tvori posameznikova družbena (in kulturna) identiteta, ki se nanaša na pripadnost širši družbeni skupnosti in na družbene kategorije, ki jim posameznik pripada – kot so npr. narodnost, etnična pripadnost, vera, socialno-ekonomski status, poklic in pripadnost različnim družbenim skupinam (družina, prijatelji, družabne skupine). Po drugi strani kulturna identiteta zajema skupne običaje, tradicije, prepričanja, vrednote, vedenje in tudi posebne predmete, ki določeno skupino ali družbo opredeljujejo. Hkrati jo določajo tudi dejavniki, kot so jezik, vera, narodnost, dediščina ter kulturne prakse in norme, ki jih posamezniki podedujejo in sprejemajo. Klepetalni roboti sicer družbene ali kulturne identitete nimajo, a lahko so zasnovani tako, da glede na kontekst uporabe družbeno in kulturno identiteto do neke mere simulirajo. Poleg tega so lahko programirani tako, da razumejo različne jezike in kulturne norme ter se nanje odzivajo – kar jim omogoča učinkovito komunikacijo z uporabniki. Podobno velja za vidike psihološke identitete, ki se nanašajo zlasti na posameznikovo subjektivno občutenje samega sebe – na njegovo samopodobo, samospoštovanje in pojmovanje sebe (Cote in Levine 2002). To med drugim vključuje to, kako se posamezniki dojemajo v odnosu do drugih, njihove notranje misli, čustva, prizadevanja in pripovedi, ki jih oblikujejo o svojem življenju. Razvijalci lahko klepetalnim robotom poleg osebnega imena in značilnosti, ki ustrezajo določeni družbeni ali kulturni identiteti, vgradijo še bolj osebnostne vidike, npr. prijazno in profesionalno osebnost za sistem, ki prihaja v stik z kupcem – medtem ko ima klepetalni robot za igralno platformo lahko bolj igrivo osebnost.

Naslednji posebej pomemben vidik identitete je tudi moralna identiteta. Moralna identiteta zajema človekov moralni značaj, vrednote, načela in stališča, povezana z ideali, in se oblikuje v smiselno, obstojno in razmeroma stabilno predstavo o posameznikovem moralnem jazu (Doris 2002) – vse to pa se odraža v dejanjih in projektih, ki jim v življenju sledi. Moralna identiteta, razumljena na tak način, očitno vpliva na človekove moralne presoje, dejanja in odločitve: bodisi z oblikovanjem prostora alternativ, ki jih posameznik vidi kot ustrezne, dopustne ali vredne, da si zanje prizadeva (omejujejo naše izbire), ali z dajanjem prednosti izbrani alternativi pred drugo (Strahovnik 2011). Avtorji, kot je Charles Taylor (1990), nas posebej svarijo pred poskusi, da bi etiko utemeljili na osiromašenem pojmovanju identitete, in zahtevajo, da je treba upoštevati posameznikov značaj, ideale, vrednote in odgovornost ter tudi vpetost te identitete v posameznikovo skupnost. En primer takšnega osiromašenega modela bi bilo Sartrovo razumevanje človekove (moralne) identitete. Zanj »[človek] ni nič drugega kot to, kar namerava, obstaja samo toliko, kolikor se uresničuje, ni torej nič drugega kot vsota svojih dejanj, nič drugega kot to, kar je njegovo življenje« (Sartre 1946). Pogovorni sistemi umetne inteligence kot taki nimajo zavesti in sposobnosti oblikovanja moralnih pre-

pričanj ali vrednot, vseeno pa so lahko zasnovani na etičen oz. odgovoren način in upoštevajo vidike morebitnega vpliva interakcij z uporabniki. Navsezadnje je to težnjo mogoče razbrati tudi iz usmeritev glede razvoja teh sistemov, kjer je v ospredju govor o zaupanja vredni, odgovorni in na človeka osrediščeni umetni inteligenci (Evropska komisija 2019).

Preden se pomaknemo k vprašanjem dojemanja in spreminjanja identitete na podlagi interakcije s sistemi umetne inteligence, na kratko izpostavimo še vidik identitete, za katerega lahko uporabimo izraz spoznavna identiteta. Spoznavna identiteta se med drugim nanaša na posameznikov odnos do vednosti in tudi do samega procesa spoznavanja. Kot takšna vključuje načine oz. vzorce, glede na katere posamezniki vednost dojemajo, pridobivajo, potrjujejo in uporabljajo – pa tudi njihova prepričanja in stališča do znanja in njegovih virov (Caltabiano, Adam in Denham 2019). Vidike spoznavne identitete oblikujejo različni dejavniki, vključno s kulturnim in izobraževalnim kontekstom. Ne gre samo za posamezne učne sloge, o katerih je bilo v zvezi s spoznavno identiteto precej govora v preteklosti. Pomembno je, da jo razumemo na način, ki vključuje posameznikova prepričanja o zanesljivosti različnih virov znanja, kot so osebne izkušnje, intuicija, avtoriteta, znanstvena metoda ali kulturna tradicija. Kot takšna lahko širše vpliva na pristop posameznika k reševanju problemov, odločanju ter sodelovanju s prepričanji in idejami drugih, na načine vstopanja in sodelovanja v razpravo, ki vključuje nasprotujoča si stališča, in na odprtost za nove perspektive. Podobno kot pri moralni identiteti tudi spoznavno identiteto v pomembni meri tvorijo spoznavne vrline, kot so npr. spoznavna odgovornost, pazljivost, premišljenost, spoznavni pogum, smisel za pomembnost in za bistveno, spoznavna integriteta, temeljitost, intelektualna treznost in poštenost, nepristranskost, spoznavna ponižnost idr. (Strahovnik 2022). Pogovorni sistemi umetne inteligence spoznavne identitete neposredno nimajo, zato besedilo razbirajo in ustvarjajo na podlagi vzorcev, prisotnih v besedilnih podatkih, na katerih so se učili. Lahko pa so dojeti oz. delujejo tako, da se zdijo njihovi odgovori kompetentni, resnični, preverjeni, poznavalski – hkrati pa so takšni sistemi zmožni tudi logičnega razmišljanja, spomina, ipd., kar je vse povezano z vidiki spoznavne identitete.

3. Pogovorni sistemi umetne inteligence

Pogovorni sistem umetne inteligence oz. klepetalni robot (angl. *chatbot*) je računalniška aplikacija, ki je zmožna sporazumevanja z ljudmi v naravnem jeziku (McTear, Callejas in Griol 2016) in je znana tudi kot pogovorni ali virtualni pomočnik (ang. *conversational agent, voice assistant*). Takšni sistemi za pogovore z uporabniki v naravnem jeziku uporabljajo umetno inteligenco (npr. strojno učenje razumevanja oz. procesiranja naravnega jezika) – v tem smislu je njihov namen simulirati človeku podobne interakcije in v tem okviru podajati ustrezne in smiselne odgovore na naša vprašanja oz. pozive (Mariani idr. 2023). Eden prvih razvitih klepetalnih robotov je bil program ELIZA, ki ga je leta 1966 ustvaril oz. spisal Joseph Weizenbaum. ELIZA je znala simulirati besedilni pogovor in je bila ustvarjena z namenom

pokazati, kako zlahka je uporabnike preslepiti, da verjamejo, da se pogovarjajo s dejansko osebo – čeprav ELIZA kot program ni prestala slovitega Turningovega testa (Prinz 2022, 24; McTear , Callejas in Griol 2016, 57).

Obstajata dve glavni vrsti klepetalnih sistemov. Prva temelji na sledenju vnaprej podanim pravilom, ki so lahko nadalje podprta z določeno obliko umetne inteligence. Običajno gre pri teh sistemih za vnaprej programirane odzive na vprašanja oz. pozive. Druga vrsta je zanimivejša, in sicer gre za sisteme, ki so odprti in nimajo vnaprej določenih odgovorov – torej v celoti temeljijo na velikih jezikovnim modelih umetne inteligence, ki so sposobni učenja oz. razvoja. Zanimali nas bodo predvsem ti sistemi. Običajno takšni klepetalni roboti temeljijo na tvorbi besedila ali govora, ki potem predstavlja medij interakcije z uporabnikom. Področja uporabe teh sistemov so številna – od zahtevnejših področij avtomobilske, vojaške in drugih tehnologij, varnosti, izobraževanja, področij jezika (prevajanje), zdravstva do nekoliko preprostejših področij, kot so nakupovanje, kuhanje, zabava, upravljanje doma itd. Klepetalni roboti so lahko del spletnih mest, aplikacij za sporočila, drugih mobilnih aplikacij, naprav z glasovno podporo, kot so mobilni telefoni, pametni zvočniki, pametne ure, tablice ipd. Najbolj znani so Google Assistant (Google), Siri (Apple), Alexa (Amazon) in Bixby (Samsung). Naprave s tovrstnimi sistemi postajajo vse bolj razširjene, saj so praktične in preproste za uporabo ter prilagodljive tudi za starejše ljudi in tiste, ki so gibalno ali kako drugače ovirani. Z njimi lahko uporabnik sam nadzoruje vse naprave, ki so s temi sistemi združljive: radijski sprejemnik, klimatsko napravo, pralni stroj, luč, osebni koledar ipd. – tako ves čas ostaja obveščen in ohranja občutek nadzora. Rast glasovnih asistentov v okviru potrošniške tehnologije naj bi bila primerljiva celo s porastom uporabe in razvoja pametnih telefonov. Ti sistemi tako v okviru omenjene rabe običajno zagotavljajo informacije in pomoč, odgovarjajo na vprašanja in izvajajo naloge – pomagajo reklamirati nedelujoč izdelek, naročiti prevoz, nastaviti opomnik, predvajati izbrano glasbo ipd. Gre torej za obdelavo in razlago uporabnikovih vhodnih podatkov, ustvarjanje kontekstualno ustreznih odzivov in pogosto tudi učenje iz interakcij, kar jim omogoča da svoje delovanje sčasoma izboljšajo. Ena izmed njihovih vlog je tudi, da vzbudijo pozornost potrošnikov in ponudijo personalizirane odzive, ki so tako bolj prilagojeni, kot bi bili ob zgolj statičnem naboru informacij. Raziskave kažejo, da torej ne gre le za to, da bi uporabniki s pomočjo teh sistemov dobili odgovore in opravljali naloge, ampak do teh sistemov izražajo tudi določena čustva, oblikujejo naravnanosti – in z njimi ustvarjajo odnose (Malodia idr. 2021).

V zadnjem času se pogosto pojavlja tudi uporaba, prvenstveno namenjena t. i. digitalnemu oz. virtualnemu družabništvu (npr. sistemi Replika, Kuki, Mitsuku, Cleverbot), duševnemu zdravju (npr. WYSA, Woebot, Elomia, Mindsum) – pa tudi klepetalni roboti, namenjeni pogovorom o veri oz. svetih besedilih (npr. Bible Buddy, Chat KJV). Takšni sistemi se razvijajo predvsem zaradi porasta zanimanja in potreb na področju duševnega zdravja ter duhovne rasti ljudi, ki bi si želeli dostopati do spletnih svetovalcev, terapevtov, duhovnikov ali družabnikov (Kraus, Seldschopf in Minker 2021). Takšna oblika svetovanja, pomoči ali družabništva je ljudem lahko bližja zaradi različnih razlogov, kot je npr. lokacija, dostopnost in prilagodljivost, kadar svetova-

lec zaradi različnih razlogov ni dostopen – spletni svetovalec pa je na voljo povsod in vedno, kjer in kadar ima posameznik dostop do spleta. Nekateri klepetalniki so dostopni tudi brezplačno, kar pomeni, da morebitne finančne stiske posameznikov, ki so osamljeni ali želijo pomoč, niso ovira. Še dodaten razlog pa je, da si uporabnik morda želi ostati anonimen, kar mu ti sistemi omogočajo. Najti je mogoče še druge razloge, kot so bile npr. omejitve gibanja ljudi ob nedavni pandemiji.

Da bi lahko razumeli, zakaj so ti sistemi v porastu in za ljudi privlačni, je treba razumeti, kako so oblikovani oz. ustvarjeni (Kraus, Seldschopf in Minker 2021). Eden od razlogov je, da lahko delujejo kot ‚pristen' sogovornik, če npr. izvzamemo vizualni faktor, ki sicer na občutja uporabnikov tudi pomembno vpliva (Youting in Jiangen 2020). Drugi razlog pa je, da so svetovalni klepetalni sistemi običajno oblikovani na podlagi dobro uveljavljenih svetovalnih pristopov, ki sicer potekajo v osebnem stiku s terapevtom, kot so npr. nevrolingvistično programiranje, kognitivna vedenjska terapija, klasični štiristopenjski svetovalni pristop (angl. *four motivational interviewing processes*) idr. Slednjega izpostavljamo kot primer, kako lahko ustvarjalci izbrani proces upoštevajo ali mu sledijo. Pristop tvorijo štiri stopnje, in sicer vzpostavitev odnosa (ang. *engaging*), osredotočanje na problem (ang. *focusing*), spodbujanje in iskanje motivacije (ang. *evoking*) ter načrtovanje (ang. *planning*) (He idr. 2022). Ključno je, da posameznika spremljamo in vodimo tako, da motivacijo za spremembo, ki se je mora lotiti, da bi dani problem razrešil, poišče sam (Miller in Rollnick 2013). Če si torej ogledamo, kako morajo biti svetovalni ali družabniški klepetalniki po tem pristopu ustvarjeni, je pomembno izpostaviti, da morajo ustvarjalci sistemov oblikovati program, ki skozi pogovor med uporabnikom in pogovornim sistemom umetne inteligence spodbuja vzpostavljanje navidezne sodelovalnosti in vzbuja sočutje – vse z namenom, da bi uporabnik premostil določen problem, ki ga predloži v reševanje.

Sicer pa takšni pogovorni roboti, ko je npr. družabniški sistem Replika, svoje odgovore ustvarjajo in pogovor izvajajo na podlagi velikega jezikovnega modela – deloma glede na vnaprej predvideno pogovorno vsebino. Pri tem je pomembno, da lahko razvijalci sistemov vanje vgradijo določena zasnovna načela oz. postopkovna pravila, npr. ton in dolžina odgovorov, ton pozivov, ki jih sami podajajo, pogostost pozivov ipd.

Morda na tem mestu omenimo še, da lahko v tudi Svetem pismu najdemo zanimivo mesto, kjer podobna ‚jezikovna tehnologija' nastopa kot mistična tehnologija oz. oblika sporazumevanja z Bogom (2 Mz 28,30; 3 Mz 8,8; 1 Sam 28,6), in sicer gre za omembo *urima* in *tumina*. Interpretacije, za točno kakšne vrste tehnologijo oz. pripomočka gre, so sicer zelo raznolike. Tukaj izpostavljamo zgolj tiste, ki so za našo temo najbolj zanimive. Gre za neke vrste pripomoček oz. način, ki je bil v starem Izraelu dovoljen za povpraševanje po prihodnosti – po tem razumevanju sta bila *urim* in *tumim* običajno dva kamenčka ali dve paličici, s katerima so duhovniki Boga spraševali za bodisi pritrdilen bodisi odklonilen odgovor (1 Sam 28,6, op., 6). Glede na eno izmed razumevanj – tj. v rabinski literaturi – lahko najdemo podrobnejše opise pripomočka in njegovega delovanja. To razumevanje pravi, da če je oseba želela dobiti odgovor na neko vprašanje, je morala najprej stopiti pred velikega duhovnika, ki je bil obredno oblečen v osem oblek in obrnjen k Bogu (lahko

tudi k Skrinji zaveze). Oseba je nato jasno in kratko izgovorila eno vprašanje in pričakovala odgovor. V tem trenutku je velikega duhovnika prevzel Sveti Duh. Duhovnik je nato vizijo posredoval v obliki črk preko naprsnika oz. tablic, ki so visele na njegovih prsih. *Urim* in *turim* naj bi bila svetloba, ki je prihajala skozi dragulje na tablicah na prsih (naprsniku) duhovnika. Dragulji so predstavljali črke – tisti, skozi katere je šla svetloba, so bili del sporočila oz. odgovor na vprašanje (Hirsch, Muss-Arnolt in Blau 1906; prim. tudi Maimonidesovo delo *Mishna Torah for Rambam, Book of Work*, *Laws of the Temple Vessels and the Worshipers in It 10,11*). Vidimo lahko, da gre po nekaterih izročilih oz. interpretacijah pri urminu in turimu za tehnologijo tvorjenega oz. tako ali drugače posredovanega besedila, ki so ga razbirali iz tablic – podobno kot lahko sedaj mi prebiramo odgovore pogovornega sistema.

Iz zgoraj povedanega vidimo, da se klepetalni roboti uporabljajo za vrsto opravil: v prodaji in podpori potrošnikom, zdravstvu, financah, izobraževanju, na področju zabave – vse do terapij in duhovne podpore. Postajajo nekakšni sopotniki in sooblikovalci človekovega vsakdana – in s tem posameznika samega. V nadaljevanju se bomo zato posvetili vprašanjem, pomislekom in izzivom identitete, ki jo ti sistemi lahko imajo ali pri posamezniku izzovejo. Vse to odpira tudi pomembna etična vprašanja. Bolj očitni etični izzivi so npr. v tem, da se s takšno uporabo oži polje zasebnosti posameznika, saj naprave s temi sistemi delujejo neprekinjeno, so v pripravljenosti, poslušajo naš govor ipd. Problem je tudi posredovanje ali dostop do osebnih podatkov, ki jih ti sistemi za delovanje potrebujejo ali pa jih lahko pri delovanju pridobijo – npr. uporabnikov e-poštni račun, telefonska številka, seznam stikov, fotografije ipd. Poleg tega se ti sistemi posredno seznanijo z navadami uporabnika, ki jih zberejo iz vseh naprav – npr. kdaj gre spat ali kdaj zjutraj običajno vstane, kdaj zapusti dom ali kakšno glasbo posluša. Sledi tudi problem kraje identitete, saj obstaja tveganje, da uporabniku vdrejo v sistem, dostopijo do njegovih osebnih podatkov, celo vstopijo v dom. Na drugi strani so nekoliko manj izpostavljeni etični izzivi, ki pa imajo lahko za posameznike in človeštvo nasploh še pomembnejše posledice. Eden večjih izzivov je odnos, ki se vzpostavi med uporabnikom in sistemom umetne inteligence. Kot vsak odnos tudi ta na življenje posameznika in tudi na celotno dinamiko drugih odnosov, v katere je vključen, sovpliva. Tu je potem še vprašanje uporabnikove psihološke identitete v odnosu s sistemom, saj – kot smo videli v zgornjih primerih – se uporabnik v odnos s sistemom tudi čustveno oz. bolje rečeno osebno vplete, tako da ta vpliva na njegova prepričanja, vero, pogled na odnose in življenje ipd.

Zaključimo lahko, da težava ni toliko v tem, da se umetna inteligenca in uporabnik spreminjata ali razvijata – izziv je razumeti in bolje nadzirati, kako in v kaj se v teh procesih preoblikujeta. Zanima nas, ali lahko predvidimo, kakšne so dolgoročne posledice tega odnosa, in kaj lahko naredimo, da bi znali omenjene sisteme oz. orodja koristno usmeriti, oblikovati in uporabiti za dvig kakovosti življenja posameznika in skupnosti. To pa ne pomeni zgolj učenja in prakse čuječnosti uporabnika in ustvarjalcev ob uporabi teh sistemov, temveč tudi sodelovanje pri ustvarjanju in tudi njihovem poznejšem preoblikovanju. Kot pravijo Mckee idr., »bo človekovo dojemanje sistemov umetne inteligence oblikovano ne le na podlagi delovanja teh sistemov, ampak tudi na podlagi konteksta, v okviru katerega

do odnosov in interakcije s temi sistemi prihaja« (2021). V nadaljevanju se zato posvečamo izzivom identitete; še prej si bomo ogledali nekaj relevantnih izsledkov raziskav, ki se našega raziskovalnega vprašanja dotikajo.

4. Izzivi identitete sistemov in naše lastne identitete

Kot smo lahko videli v prejšnjem razdelku, so sistemi, na katerih so pogovorni roboti zgrajeni, načrtovani premišljeno – in pogosto tudi usmerjeni na določeno ciljno skupino. To hkrati pomeni, da več podatkov kot uspejo zbrati, boljše nam bodo lahko svetovali oz. nas vodili. Del teh ključnih podatkov pa seveda razkriva tudi našo identiteto. Sistemi pa ne pridobivajo zgolj podatkov, ki jih sami razkrijemo – in za katere vemo, da smo jih razkrili –, ampak zbirajo podatke tudi iz drugih virov ali pa so jih sposobni sami uganiti ali določiti. Zato nas v prispevku zanima predvsem odnos uporabnikov do pogovornih sistemov umetne inteligence: natančneje, kako uporabniki klepetalnih sistemov dojemajo identiteto teh sistemov, kako dojemajo sebe, kaj o sebi spoznajo – in kako lahko to vpliva na nadaljnji razvoj obravnavanih tehnologij.

Eno izmed osrednjih področij, kjer je ta izziv v ospredju in ki smo ga že omenili, je gotovo področje duševnega zdravja, kjer se raziskave sicer trenutno ukvarjajo prvenstveno z dostopnostjo pomoči in uporabnostjo sistemov umetne inteligence (Kraus, Seldschopf in Minker 2021). Ena izmed raziskav, povezana z identiteto, je bila osredotočena na rasno in etnično podobnost oz. ujemanje med uporabniki in svetovalnimi sistemi umetne inteligence v vlogi terapevta oz. svetovalca (v nadaljevanju UI terapevti) (Yuting in He 2020). V raziskavi so ugotovili, da so udeleženci, ki so bili iste rase ali etnične pripadnosti, kot je bila predstavljena rasa – zunanja podoba, ki jo privzame terapevt – njihovih UI terapevtov, lahko z njimi vzpostavili tesnejšo vez oz. globlji odnos ter da bi z istim UI terapevtom želeli sodelovati tudi v prihodnosti – in bi ga priporočili svojim bližnjim. Zanimivo je, da so udeleženci, ki so bili v interakciji z nepersonificiranim UI terapevtom, poročali skoraj enako – do odstopanj glede zaupanja in tesnosti vezi je prihajalo le tam, kjer se rasa oz. etnična pripadnost ni ujemala. Raziskava je pokazala tudi, da so se udeleženci iste rase kot UI terapevt težje odprli (zaradi občutka, da jih bo nekdo sodil, dobrega vtisa, strahu pred družbeno stigmatizacijo, potrebe po pripadnosti). Da bi omenjeno težavo rešili, so raziskovalci predlagali, da se v pogovorni sistem vgradi več spodbudnih besed, morda tudi prek čustvenih simbolov; ko pa demografskih podatkov o uporabnikih ni dovolj, se svetuje nepersonificiran UI terapevt ali možnost, da si uporabnik videz UI terapevta po svojih potrebah priredi sam (438–439). Gre pa seveda pri rasi zgolj za en vidik identitete – treba je upoštevati tudi druge vidike: od spola, telesnih značilnosti, do vere in pogleda na duhovnost itd.

Pri UI asistentih je pomenljiv tudi premislek, ko gre za predsodke glede spola, ki so bili (v)kodirani že med samim procesom oblikovanja UI sistema. Če si ogledamo primere najbolj razširjenih in znanih pogovornih asistentov, so – Amazonova Alexa, Applova Siri, Microsoftova Cortana ali pa Googlova asistentka – večinoma predobli-

kovani kot mlade in razmeroma podrejene ženske (Chin in Robison 2020, 82–104). UNESCO (West, Kraut in Ei 2019; Cercas Curry, Robertson in Reiser 2020) to razume kot tveganje za krepitev spolnih stereotipov. V primerih zlorabe pri uporabniku, ko UI asistentka ne odgovori ‚primerno', pa se to tveganje še poveča – in škodljiv vzorec komunikacije lahko prenese v druge odnose (Cercas Curry Robertson in Reiser 2019). Raziskave so bile zato posvečene tudi iskanju alternativnih oblik pogovornih asistentov. Ugotavljajo, da bi večina udeležencev želela, da je glas UI asistenta robotski, nato spolno nevtralen, nato ženski in nazadnje moški. Trenutna analiza pa kaže, da ima velika večina pogovornih robotov ženski glas. Večina udeležencev je tudi želela, naj bo pogovorni robot star med 25 do 40 let, ali pa jim je bilo vseeno, koliko let ima. Skoraj nihče pa ni izrazil želje, da bi pogovorni robot imel glas, ki odraža starost 24 let in manj. Slednje je povsem v nasprotju s trenutnim stanjem – največ robotskih asistentov odraža glas do 20 leta starosti. Večina udeležencev bi želela robotske asistente, ki so upodobljeni kot ljudje, nato kot živali in nazadnje kot roboti; pa da ne bi imeli naglasa, da bi bili prijazni, ustrežljivi, dejansko v pomoč in da bi imeli tudi smisel za humor. Ugotavljajo tudi, da pri UI asistentih ni pomembna smo osebnost, izražena skozi glas, ampak tudi osebnost, izražena v drugih vidikih vedenja, vsebine pogovora in načina pogovora (Cercas Curry 2020, 75–76). Zato je verjetno, da se bodo ti sistemi razvijali tudi v smer specifičnih prilagoditev glede na želje uporabnikov. Trenutno namreč interakcija med ljudmi in vodilnimi UI sistemi poteka tako, da so ti z vidika uporabnika v vlogi tekmecev (pri igrah), pomočnikov in svetovalcev. V prihodnosti pa bodo ljudje z UI asistenti stopali tudi v drugačne odnose, kot so npr. učenec in učitelj, pacient in zdravnik, stanovalec in načrtovalec mesta ipd. – in jih bodo dojemali oz. jih želijo dojemati kot člane skupnosti (McKee idr. 2021, 6; 24).

Izpostavimo lahko še, da v pogovorih s pogovornimi roboti, ki imajo značilnosti človeka (kot so ime, spol, način govora), te posameznika samodejno pripeljejo do počlovečenja teh sistemov in čustvene navezanosti nanje (Malodia idr. 2021). To pa vodi v večjo raven zaupanja in posledično tudi do večjega obsega razkritja osebnih podatkov (Ischen idr. 2020, 43–44; Miklavčič 2021). In čeprav do sedaj prevladujejo raziskave, ki so usmerjene na zadovoljstvo uporabnikov omenjenih sistemov, je kljub temu mogoče najti tudi zanimive izsledke za vidike zaupanja in identitete. V okviru ene izmed takšnih raziskav so v razvoj UI terapevta vključili tudi elemente pomenkovanja (angl. *small talk*), empatične odzive in dejavni oz. samoiniciativni sistem, kar naj bi na razvoj in poglabljanje odnosa med človekom in robotom vse vplivalo pozitivno. Raziskava pa je pokazala nasprotno – da vsaj za prva dva elementa ni nujno, da pozitivno vplivata na razvoj zaupanja do UI terapevta pri uporabniku. Predvidevajo, da so se uporabniki zaprli oz. UI terapevtu niso zaupali, ker se je pogovor dotikal osebnih in občutljivih vsebin, kar je uporabnike spodbudilo, da so postali previdni. Pozitivni učinek na uporabnika in njegovo zaupanje je medtem imelo samoiniciativno odzivanje sistema, ki naj bi bilo bolj zanesljivo in za uporabnika razumljivo (Kraus, Seldschopf in Minker 2021).

Po zgoraj povedanem vidimo, da je ne glede na hiter (celo prehiter) razvoj pogovornih sistemov umetne inteligence pred njimi oz. njihovim ustvarjanjem še veliko izzivov. Morda je ravno tukaj del težave, saj pri tako hitrem ritmu ni dovolj

časa za zadostno refleksijo (Vallor 2016) in nato vpeljevanje novih uvidov v razvoj – kaj šele proučevanje dolgoročnih posledic za človeštvo. Kot izpostavljajo mnogi, je eden pomembnejših korakov zlasti vključevanje v razvoj teh tehnologij, skupno sodelovanje strokovnjakov, ki poleg strokovnjakov inženirstva vključuje tudi strokovnjake s področij etike, filozofije, psihologije, teologije, oblikovanja in tudi drugih področij. To bi namreč spodbudilo razprave o perečih in ključnih vprašanjih (Lee idr. 2022; Vallor 2016), ki se dotikajo tudi vprašanja odnosa med uporabnikom in umetno inteligenco oz. natančneje vprašanja o tem, kakšnega človeka ta odnos oblikuje, kako (lahko) en na drugega sovplivata ipd.

Vse to razkriva pomembnost tega, da je pri interakciji s pogovornimi sistemi umetne inteligence prisotno tudi vnovično določanje oz. oblikovanje naše identitete. Do tega lahko pride na različne načine. Prvi izmed njih se nanaša na privzemanje vlog in perspektiv ter posledično igranje različnih vlog (učenec, učitelj, zaupnik, stranka), ki lahko povratno vplivajo tudi na našo lastno identiteto. Nekatere od teh vlog so lahko takšne, da jih spodbudi bodisi pogovorni robot neposredno ali pa so posledica naše interakcije z njim – ne pa nujno nekaj, za kar bi se zavestno in avtonomno odločili sami. Drugič, do tega lahko pride na podlagi same vsebine pogovora oz. pri razkrivanju podatkov o samem sebi. Klepetalni roboti uporabnike pogosto spodbujajo, da delijo tudi osebne podatke ali pa poročajo o svojih mislih in občutkih. Takšno samorazkrivanje lahko posameznike spodbudi ne zgolj k razmisleku o svoji identiteti, ampak tudi k spremembi identitete same. To se lahko zgodi, če pride do deljenja vidikov, ki jih običajno z drugimi ne delijo. Ali pa vprašanje, ki ga zastavi pogovorni robot, pri posamezniku na raven zavestnega mišljenja spravi spomin, ki je bil sicer zakrit – to pa nato v pomembni meri vpliva na posameznikovo samopodobo in identiteto. Gre torej za vpliv procesa samorefleksije, ki pa ga pomembno določa interakcija s klepetalnim robotom. Zato je pomembno, da se pri načrtovanju in implementaciji klepetalnih robotov zavedamo tudi etičnih razsežnosti in premislekov, ki so povezani z dostojanstvom posameznika ter njegovim občutkom odprtosti in ranljivosti. Tretji način vključuje zunanjo potrditev: klepetalni sistemi lahko uporabniku takšno zunanjo potrditev zagotovijo, nato pa se na podlagi te potrditve, ki je npr. ne bi nujno prejel od drugih, s katerimi vstopa v odnose, oblikuje posameznikova identiteta. Četrti način vključuje oblikovanje identitete na podlagi (so)oblikovanja preferenc uporabnika in prilagojenih priporočil ali predlogov (priporočilni sistemi UI), ki lahko sčasoma oblikujejo posameznikove preference, interese in vedenje. To vpliva na to, kako dojemamo svojo identiteto, saj jo uskladimo z izbirami in priporočili, ki jih podaja klepetalni robot.

5. Zaključek

Zgornji premisleki kažejo, da oblikovanje in dojemanje identitete pogovornih sistemov umetne inteligence ni enosmeren proces, zato moramo pri tem upoštevati tudi, kako se pri interakciji s temi sistemi lahko spremeni naša lastna identiteta. Prav to odpira cel niz etičnih, pa tudi širše antropoloških vprašanj in izzivov, na katere dobrih odgovorov še ni. Sistemi umetne inteligence predstavljajo novo resničnost,

s katero se človek srečuje ali jo – bolje rečeno – vsakodnevno živi. Poleg tega, kako dojemamo njihovo zunanjo ‚podobo' (spol, glas, barva kaže, ipd.), bo vse večji pomen dobivalo tudi vprašanje, kako dojemamo druge vidike njihove identitete. Morali bomo tudi uskladiti načine govora o teh sistemih – npr. ko rečemo, da je neki sistem vreden zaupanja ali dober terapevt ali pa pravičen. Ali gre tu za metaforo ali dejansko za pripis identitete? Ali lahko ti sistemi posedujejo vrline (Vallor 2016)?

Etična in antropološka vprašanja so povezana tudi s tem, kako se ti sistemi predstavljajo in z uporabniki komunicirajo. Pri tem lahko pride do primerov lažnega ali napačnega predstavljanja. Uporabniki lahko namreč s pogovornimi sistemi umetne inteligence vzpostavijo čustvene vezi. Če pa klepetalni robot pravo identiteto skriva, lahko uporabniki odnos razvijejo na podlagi napačne predpostavke – kar lahko ob odkritju resnice privede do čustvene stiske. Oblikovanje klepetalnih robotov, ki posnemajo človeške osebnosti in čustva, tako sproža vprašanja o etičnih posledicah poosebljanja umetne inteligence – to namreč lahko zabriše meje med človekom in strojem, kar lahko privede do nesporazumov. Izpostaviti je mogoče tudi vidik predstavljanja kulture in kulturne identitete, saj lahko ta utrjuje stereotipe in je žaljiva. Interakcija z temi sistemi lahko privede tudi do odvisnosti in vpliva na to, kako takšni uporabniki komunicirajo z drugimi (Dorobantu idr. 2022). In če identiteta ni jasna, je določitev odgovornosti za dejanja (npr. odgovore, priporočila), ki jih izvajajo klepetalni roboti, lahko zahtevna. Etični pomisleki so v ospredju tudi, kadar sistemi umetne inteligence manipulirajo s čustveno ranljivostjo uporabnikov ali jo izkoriščajo. Za odpravo teh pomislekov je zato ključno, da razvijalci in organizacije sprejmejo etične smernice, ki dajejo prednost preglednosti in poštenosti identitete klepetalnega robota. Uporabniki morajo biti jasno in vnaprej obveščeni, da so v stiku z umetno inteligenco (Evropska komisija 2019). Glede na skokovit razvoj teh sistemov ter njihovo uporabo v povezavi z roboti gre pri vseh teh vprašanjih za nenehno iskanje ravnovesja med koristmi in škodljivimi posledicami, ki se mu ne moremo izogniti – zato terja temeljit premislek.

Reference

Caltabiano, Marie L., Raoul J. Adam in Rebecca Denham. 2019. Epistemic Identity and Undergraduate Students' Understandings of Psychology. *International Journal of Education, Psychology and Counseling* 4, št. 30:299–314.

Cercas Curry, Amanda, in Verena Rieser. 2019. A Crowd-based Evaluation of Abuse Response Strategies in Conversational Agents. V: *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. *Association for Computational Linguistics*, 361–366. ACL Anthology. https://aclanthology.org/W19-5942.pdf (pridobljeno 20. 4. 2023).

Cercas Curry, Amanda, Judy Robertson in Verena Rieser. 2020. Conversational Assistants and Gender Stereotypes: Public Perceptions and Desidetata for Voice Personas. V: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing Association for Computational Linguistics,* 72–78. ACL Anthology. https://aclanthology.org/2020.gebnlp-1.7.pdf (pridobljeno 20. 4. 2023).

Chin, Caitlin, in Mishaela Robison. 2020. *How AI Bots and Voice Assistants Reinforce Gender: AI in the Age of Cyber-Disorder Actors, Trends, and Prospects.* Milano: Ledizioni LediPublishing.

Cote, James E., in Charles G. Levine. 2002. *Identity, Formation, Agency, and Culture: A Social Psychological Synthesis.* New York: Psychology Press.

Doris, John M. 2002. *Lack of Character: Personality and Moral Behaviour.* Cambridge, MA: Cambridge University Press.

Dorobantu, Marius, Brian Patrik Green, Anselm Ramelow in Eric Salobir. 2022. Being Human in

the Age of AI. Research gate. https://www.researchgate.net/publication/365945288_Being_Human_in_the_Age_of_AI?channel=doi&linkId=6389b3b82c563722f22e84df&showFulltext=true (pridobljeno 8. 5. 2023).

Evropska komisija. 2019. Etične smernice za zaupanja vredno umetno inteligenco. Evropska komisija. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/JURI/DV/2019/11-06/Ethics-guidelines-AI_SL.pdf (pridobljeno 20. 4. 2023).

He, Linwei, Erkan Başar, Reinout Wiers, Marjolijn Antheunis in Emiel Krahmer. 2022. Can Chatbots help to motivate Smoking Cessation? A Study on the Effectiveness of Motivational Interviewing on Engagement and Therapeutic Alliance. *BMC Public Health* 22, 726.

Hirsch, Emil G., William Muss-Arnolt, Wilhelm Bacher in Ludwig Blau. 1906. Urim and Thummim. V: *Jewish Encyclopedia*. Zv. 12, 384–386. New York: Funk and Wagnalls.

Ischen, Carolin, Theo Araujo, Hilde Voorveld, Guda van Noort in Edith Smit. 2020. Privacy concerns in chatbot interactions. V: Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger in Petter Bae Brandtzaeg, ur. *Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019*, 34–48. Cham: Springer.

Kraus, Matthias, Philip Seldschopf in Wolfgang Minker. 2021. Towards of Trustworthy Chatbot for Mental Health Applications. V: Jakub Lokoč, Tomáš Skopal, Klaus Schoeffmann, Vasileios Mezaris, Xirong Li, Stefanos Vrochidis in Ioannis Patras, ur. *MultiMedia Modeling*, 354–366. Cham: Springer.

Lee, Minha, Jaisie Sin, Guy Laban, Matthias Kraus, Leigh Clark, Martin Porcheron, Benjamin R. Cowan, Asbjørn Følstad, Cosmin Munteanu in Heloisa Candello. 2022. Ethics of Conversational User Interfaces. *CHI EA '22: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, članek 80:1–7. ACM Digital Library. https://dl.acm.org/doi/10.1145/3491101.3503699 (pridobljeno 24. 4. 2023).

Mariani, Marcello M., Novin Hashemi in Jochen Wirtz. 2023. Artificial Intelligence Empowered Conversational Agents: A Systematic Literature Review and Research Agenda. *Journal of Business Research* 161:113838.

Malodia, Suresh, Nazrul Islam, Puneet Kaur in Amandeep Dhir. 2021. Why Do People Use Artificial Intelligence (AI)-Enabled Voice Assistants? *IEEE Transactions on Engineering Management* 71:491–505. https://doi.org/10.1109/tem.2021.3117884

McKee, Kevin R., Xuechunzi Bai in Susan Fiske. 2021. Humans Perceive Warmth and Competence in Artificial Intelligence. *PsyArXiv* 26.

McTear, Michael, Zoraida Callejas in David Griol. 2016. *The Conversational Interface: Talking to Smart Devices*. Cham: Springer International Publishing.

Miklavčič, Jonas. 2021. Zaupanje in uspešnost umetne inteligence v medicini. *Bogoslovni vestnik* 81, št. 4:935–946. https://doi.org/10.34291/bv2021/04/miklavcic

Miller, William R., in Stephen Rollnick. 2013. *Motivational Interviewing: Helping People Change*. 3. izd. New York: The Guilford Press.

Prinz, Konstantin. 2022. *The Smiling Chatbot: Investigating Emotional Contagion in Human-to-Chatbot Service Interactions.* Wiesbaden: Springer.

Sartre, Jean-Paul. 1946. Existentialism is a Humanism. V: Walter Kaufman, ur. *Existentialism from Dostoyevsky to Sartre*, 289–311. New York: Meridian Publishing.

Spillane, Brendan, Emer Gilmartin, Christian Saam in Vincent Waade. 2019. Issues Relating to Trust in Care Agents for the Elderly. V: *Proceedings of the 1st International Conference on Conversational User Interfaces, CUI '19*, članek 20. ACM Digital Library. https://dl.acm.org/doi/10.1145/3342775.3342808 (pridobljeno 24. 4. 2023).

Strahovnik, Vojko. 2011. Identity, Character and Ethics: Moral Identity and Reasons for Action. *Synthesis Philosophica* 26, št. 1:67–77.

– – –. 2022. Identiteta, etika prepričanja, razumnost in resonanca. *Bogoslovni vestnik* 82, št. 3:547–559. https://doi.org/10.34291/bv2022/03/strahovnik

Sveto pismo Stara in Nova zaveza: SSP študijska izdaja. 2020. Ljubljana: Društvo Svetopisemska družba Slovenije.

Taylor, Charles. 1990. *Sources of the Self: The Making of the Modern Identity.* Cambridge, MA: Cambridge University Press.

Vallor, Shannon. 2016. *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting.* New York: Oxford University Press.

Yuting, Liao, in Jiangen He. 2020. Racial Mirroring Effects on Human-Agent Interactions in Psychotherapeutic Conversations. V: *Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI'20*. ACM Digital Library. https://dl.acm.org/doi/10.1145/3377325.3377488 (pridobljeno 24. 4. 2023).

West, Mark, Rebecca Kraut in Chew Han Ei. 2019. I'd Blush if I could: Closing Gender Divides in Digital Skills through Education. Unesco. https://unesdoc.unesco.org/ark:/48223/pf0000367416.locale=en (pridobljeno 24. 4. 2023).