■ *Pregledni znanstveni članek*

Eva Nike Cvikl, Dejan Dinevski

Teorija uma in njena uporaba na področju umetne inteligence

Povzetek. Razvoj umetne inteligence teži k vedno večjemu povezovanju robotov s človekom in človeško družbo. Ključne za sposobnost povezovanja ljudi v družbo so sposobnosti socialne kognicije in socialnih interakcij. Najpomembnejši nabor sposobnosti, potrebnih za tovrstno udejstvovanje, se imenuje teorija uma (angl. *Theory of Mind*). Gre za sposobnost razumevanja mentalnih stanj drugih posameznikov in razlikovanje le-teh od lastnih mentalnih stanj. Sposobnosti teorije uma so ključne za učinkovito gibanje v socialnem okolju ter tvorbo kakovostnih medosebnih odnosov. Obstaja že nekaj socialnih robotov, ki učinkovito simulirajo sposobnosti teorije uma ter vstopajo v smiselne socialne interakcije z ljudmi. Tovrstni roboti se uporabljajo v izobraževanju, storitvenem sektorju in v raziskovalne namene. Nekateri futuristi verjamejo, da gre pri ustvarjanju robotov s sposobnostmi teorije uma za korak, ki nas loči do zadnjega mejnika v robotiki – razvoja samozavedne umetne inteligence. Trenuten konsenz znanstvene skupnosti je, da je simulacija sposobnosti teorije uma še daleč od dejanske zaznave oziroma doživetja teh sposobnosti, ter da je mogoče ustvariti zgolj prepričljivo simulacijo samozavedanja, ne pa umetne zavesti kot take. V pričujočem prispevku bomo na podlagi spoznanj iz filozofije in medicinske (psihiatrične) stroke orisali koncept teorije uma ter probleme pri epistemoloških in praktičnih korakih implementacije tega koncepta na področju umetne inteligence.

Theory of Mind and Its Application in the Field of Artificial Intelligence

Abstract. The development of artificial intelligence tends to increasingly connect robots with humans and human society. Crucial to the human ability to connect in society are social cognition and social interaction skills. The most important set of abilities required for this kind of activity is called Theory of Mind. It is the ability to understand the mental states of other individuals and to differentiate them from our own mental states. The Theory of Mind abilities are crucial for effective navigation of the social environment and the formation of meaningful interpersonal relationships. There are already some social robots that effectively simulate the Theory of Mind abilities and are capable of entering meaningful social interactions with humans. Such robots are used in education, in the services sector and for research purposes. Some futurists believe that creating robots with Theory of Mind abilities is a final step that separates us to the last milestone in robotics – the development of self-aware artificial intelligence. It is important to emphasise that the simulation of the Theory of Mind abilities is far from the actual perception or experience of these abilities, and that it is possible to create only an extremely convincing simulation of self-awareness, not artificial consciousness per se. In this article, we will draw upon the findings of philosophy and the medical (psychiatric) profession to outline the concept of Theory of Mind and the problems in both the epistemological and practical steps of implementing that concept in the field of artificial intelligence.

■ **Infor Med Slov** 2020; 25(1-2): 25-32

Instituciji avtorjev / Authors' institutions: Univerzitetni klinični center Maribor (ENC), Medicinska fakulteta Univerze v Mariboru (DD)

Kontaktna oseba / Contact person: Eva Nike Cvikl, dr. med., Univerzitetni klinični center Maribor, Ljubljanska ulica 5, 2000 Maribor, Slovenija. E-pošta / E-mail: evanike.cvikl@student.um.si

Prispelo / Received: 12. 5. 2020. Sprejeto / Accepted: 30. 11. 2020.

Uvod

Ljudje smo socialna bitja. Socialni stiki so eden izmed temeljnih gradnikov naše družbe, ki nam omogočajo tvorbo odnosov in nadaljevanje vrste, izmenjavo informacij, ustvarjanje višjih družbenih struktur in institucij ter napredek civilizacije. Sposobnosti, potrebne za socialne interakcije, so v človeški vrsti prirojene in naravne, razvijemo pa jih preko procesov socialnega učenja, ki potekajo od rojstva dalje. . Ena izmed pomembnih sposobnosti, potrebnih za kvalitetne socialne interakcije, je teorija uma. Ta sposobnost nam omogoča prepoznavo mentalnih stanj sebe in drugih.

Področje umetne inteligence je že toliko razvito, da prihaja do rednih socialnih interakcij med inteligentnimi stroji in ljudmi. Naslednji korak je razvoj umetne inteligence s sposobnostjo teorije uma. V pričujočem prispevku predstavljamo področje socialne kognicije in sposobnosti teorije uma ter opisana dognanja umeščamo v kontekst sposobnosti različnih vrst umetne inteligence. V nadaljevanju prispevka si zastavljamo vprašanje, kako v razvoju umetne inteligence doseči naslednji korak po sposobnostih teorije uma – dejanski obstoj zavesti, ali je to sploh mogoče in kako bi obstoj zavesti v primeru umetne inteligence preverili.

Umetna inteligenca

Na področju računalništva se izraz »umetna inteligenca« uporablja za opis inteligence strojev, ki imajo sposobnost zaznave svojega okolja in odzivanja na okolje z nenaključnimi dejanji.[1] Ustreznejši slovenski prevod angleškega izraza »artificial intelligence« bi sicer bil »umetna inteligentnost«, vendar je dandanes »umetna inteligenca« že vsesplošno sprejeta in jo uporabljamo tudi v tem prispevku. Pogovorno se izraz »umetna inteligenca« uporablja, kadar stroj posnema kognitivne funkcije, ki jih ljudje povezujejo s sposobnostmi, kot sta učenje in reševanje problemov. V nadaljevanju predstavljamo vrste umetne inteligence:

Odzivni stroji

Gre za najstarejšo obliko sistemov umetne inteligence, ki imajo izjemno omejene, vnaprej določene kapacitete. Posnemajo sposobnost človeškega uma, da se odziva na različne vrste stimulusov. Te oblike umetne inteligence nimajo spominske funkcionalnosti, kar pomeni, da ne morejo uporabiti predhodno pridobljenih izkušenj, ki bi vplivale na njihove prihodnje poteze in odločitve. Z drugimi besedami, nimajo sposobnosti učenja. Uporabne so zgolj za direktno odzivanje na omejeno zbirko različnih stimulusov. Primer takega inteligentnega stroja je IBM-ova »Deep Blue«, ki je premagala Garryja Kasparova v šahu leta 1997.[2]

Stroji s sposobnostjo učenja

Druga oblika umetne inteligence so stroji, ki imajo ob sposobnostih reaktivnega stroja tudi sposobnost učenja iz preteklih izkušenj in odločanja na podlagi teh preteklih izkušenj. Skoraj vse v praksi uporabljene oblike umetne inteligence sodijo v to kategorijo. Sistemi, kot so tisti, ki sodelujejo pri globokem učenju, so priučeni z uporabo velikih količin podatkov za namen urjenja, ki jih nato shranijo v spominu kot referenco za reševanje prihodnjih problemov. Večina današnjih splošno uporabnih umetno inteligentnih sistemov, kot so virtualni asistenti, pogovorni roboti in tudi samovozeči avtomobili, uporablja to tehnologijo.[2]

Strojno učenje je znanstvena disciplina, ki se ukvarja s procesi, s katerimi se stroji »učijo« določenih dejanj iz podatkov. Pomembni komponenti strojnega učenja sta statistika, ki išče vzorce in zakonitosti v velikih zbirkah podatkov, ter programiranje, ki se ukvarja z izdelavo algoritmov. V grobem delimo strojno učenje na dve podvrsti: nadzorovano (zahteva označene podatke, iz katerih se algoritmi učijo reševanja problema) in nenadzorovano učenje (označeni podatki niso potrebni), v domeni nadzorovanega pa je pogosto omenjana kategorija tudi t. i. vzpodbujevano učenje.[3]

Globoko učenje je oblika strojnega učenja, ki največ navdiha črpa iz delovanja človeških možganov ter kognitivnih in nevrobioloških procesov. Globoko učenje poskuša modelirati abstraktne koncepte iz velikih podatkovnih zbirk z uporabo večslojnih umetnih nevronskih mrež, s čimer lahko kot vhodne informacije uporablja slike, zvoke in tudi tekst.

Osnova globokega učenja so umetne nevronske mreže. Razvoj umetnih nevronskih mrež je črpal navdih iz biološkega učenja že v 60. letih prejšnjega stoletja, ko so nevroznanstveniki prišli do prvih dognanj o strukturi možganske skorje. Ključno je bilo spoznanje, da ima možganska skorja notranjo strukturo, v kateri se informacije obdelujejo v posameznih slojih.[4] Nevronske mreže so bile sprva model, ki je pojasnjeval prenašanje informacij v možganih. Pri nevronskih mrežah vstopajo vhodni podatki v vhodni sloj mreže, se nato prenašajo v enega ali več skritih slojev, ti pa se nazadnje povezujejo z izhodnim slojem. Vsak sloj vsebuje procesne enote (analogne nevronom), ki so prek uteži (analogne sinapsam) povezane z enotami v predhodnih in

sledečih slojih. Enote v vhodnem sloju običajno enkodirajo spremenljivke, ki jih merimo v podatkovni zbirki. Vsak izmed globljih slojev običajno sodeluje z izgradnjo predhodnih enot. Učenje in optimizacija mreže za reševanje določenega problema prispeva k izboljšanju globljih slojev, kar posledično izboljša zgodnejše sloje. Algoritmi običajno vsebujejo enote, ki so uporabne za več različnih nalog, ter te nato prilagodijo za eno ali več specifičnih nalog.[5]

V sklopu globokega učenja se lahko uporabi tako nadzorovani kot nenadzorovani pristop k učenju, lahko pa se uporabi kombinacijo teh korakov. Kadar ima sistem za globoko učenje na voljo dovolj podatkov, lahko zgradi enote, ki so prilagojene za reševanje specifičnega problema ter nato te enote združi v sistem za predvidevanje.[5]

Teorija uma

Socialna interakcija

Socialna interakcija ali socialni stik je eden izmed temeljnih gradnikov človeške družbe. Omogoča nam tvorbo medosebnih odnosov, grajenje družbe in izmenjavo informacij o okolju. Socialna vedenja so pri otrocih prisotna že v najzgodnejšem razvojnem obdobju kmalu po rojstvu.[6]

Sposobnost vzpostavljanja primernih socialnih interakcij vsebuje več različnih procesov. Najprej mora osebek oziroma »socialni agens« prepoznati druge osebke kot živeče osebe, kar stori preko analize kompleksnih zaznavnih informacij, kot so obrazni izrazi, gestikulacija, drža, telesna govorica in glas. Ko se te informacije integrirajo, predstavljajo vhodne informacije za višje procese, ki omogočajo neposredno usklajevanje z doživetim občutjem ob procesiranih informacijah o stanju drugega (empatija), ter interpretacijo opazovanih vedenj drugega v smislu njihovih miselnih stanj (mentalizacija ali teorija uma). Z vplivom na sprejemanje odločitev bodo ti procesi vplivali na subjekt na način, da bo prilagodil svoje socialno vedenje. V opisanem procesu sodelujejo tri področja: socialna zaznava, socialno razumevanje in socialno odločanje, ki so ključne domene socialne kognicije[7].

Osnovna sposobnost, ki omogoča socialno zaznavo in udejstvovanje, je sposobnost razlikovati med objekti in subjekti (objekt je predmet, katerega vedenje je možno v celoti pojasniti preko fizikalnih pojavov, subjekt ali osebek pa ima notranja izkustva in doživetja, kot so motivacije, razlogi in nameni, zaradi česar njihovo vedenje ne more biti nikoli povsem predvidljivo).[7]

Socialna kognicija

Socialna kognicija zajema zbirko kognitivnih sposobnosti za dojemanje in obdelovanje informacij medosebnega in družbenega konteksta. Socialna kognicija osebi omogoča, da ugotovi, katera čustva čutijo posamezniki, s katerimi je v stiku, ter da na njih ustrezno odgovori. Socialna kognicija je pomemben korelat socialnih spretnosti. Propad socialnih kognitivnih funkcij lahko bolniku povzroči večjo funkcionalno škodo kot upad na drugih kognitivnih področjih, saj posamezniku onemogoča učinkovito komunikacijo ter tvorbo pristnih in kvalitetnih medosebnih odnosov. Raziskovanje socialne kognicije ter uporaba koncepta v klinični praksi je nujna za prepoznavo oškodovanosti bolnikov na tem področju ter za možnost uporabe ciljanih intervencij, ki bi bolniku vsaj deloma povrnile izgubljeno funkcionalnost oziroma preprečila izgubo le-te.[8] Ocena socialne kognicije je mogoča deloma skozi natančno klinično opazovanje, za natančnejšo opredelitev pa lahko uporabimo kognitivne naloge.[8]

Sposobnosti teorije uma

Teorija uma (angl. *Theory of Mind* – ToM) je sposobnost razumevanja lastnih mentalnih stanj in mentalnih stanj drugih. Sposobnost teorije uma se deli na afektivni del, ki vsebuje razumevanje čustvenih stanj sebe in drugih, in kognitivni del, ki pomeni razumevanje kognitivnih procesov, prepričanj, misli in namenov sebe in drugih.[8]

Izraz teorija uma in pristopi za oceno te zbirke sposobnosti so bili v znanstvenem svetu prvič uporabljeni leta 1978 v vplivnem članku Primacka in Woodroofa o sposobnostih teorije uma pri šimpanzih.[9,10] Sposobnosti teorije uma oziroma slabšanje ali propad le-teh so pomembno raziskovalno vprašanje na področju nevrodegenerativnih in nevrorazvojnih bolezni, nevroloških bolezni, možganskih poškodb in psihiatričnih bolezni.[11]

Razvoj sposobnosti teorije uma je ključen tekom razvojnega procesa. Otroci, ki sposobnosti še nimajo razvite, imajo pomembneje izraženo lastnost egocentričnosti – ne zmorejo še prevzemati perspektive drugih oseb. Z nevrološkim razvojem se ta sposobnost izboljšuje, največji premiki v razvoju teorije uma pa se zgodijo med 3. in 5. letom starosti.[12]

Sposobnost razumeti čustvena in miselna stanja drugih je ključna za naše socialno vključevanje in pravilno razumevanje socialnih interakcij.[12]

Stopnje teorije uma

Študije so pokazale, da otroci, ko pridobivajo zbirko sposobnosti teorije uma, le-te pridobivajo v stopnjah od najmanj do najbolj zahtevne:

1. Razumevanje, da imajo ljudje lahko različne želje.

2. Razumevanje, da lahko imajo ljudje različna prepričanja o isti stvari ali situaciji.

3. Razumevanje, da lahko ljudje ne vedo, da je neka stvar ali dejstvo resnično.

4. Razumevanje, da lahko imajo ljudje napačna prepričanja oziroma prepričanja, ki niso skladna z dejstvi.

5. Razumevanje, da lahko ljudje skrivajo svoje občutke in ravnajo drugače, kot občutijo.[13]

Študije so pokazale tudi, da so sposobnosti teorije uma lahko nestabilne in spremenljive. V razvojnem obdobju so pogosto vezane na posamezno situacijo, in se nato izboljšujejo tekom adolescence v odraslost. V odrasli dobi se s procesi socialnega učenja in nevroplastičnosti lahko izboljšujejo,[13] v primeru bolezni pa lahko tudi upadejo.[14]

Ocenjevanje sposobnosti teorije uma

Čustveno komponento teorije uma običajno ocenjujemo s pomočjo slikovnih dražljajev, ki predstavljajo kompleksna čustvena stanja, ali s pomočjo besednih pripovedi, ki opisujejo čustveno stanje lika. Pogosto uporabljana naloga za ocenjevanje čustvene komponente teorije uma je »Reading the Mind in the Eyes Test« (kratica MET, avtorji Baron-Cohen in sodelavci, 1997, revidirano 2001).[8] Naloga MET preiskovancem pokaže fotografijo oči vzorčne osebe, udeleženec pa mora opisati čustva, ki jih oseba na fotografiji čuti. Podobno kot pri zgornjih nalogah je tukaj pomembno kontrolirati za prisotnost druge oškodovanosti v kognitivnem procesiranju vidnih dražljajev.

Za ocenjevanje kognitivnega dela teorije uma uporabljamo naloge, ki vključujejo resnična in napačna prepričanja. Naloga z resničnimi prepričanji od udeleženca zahteva, da glede na informacije, ki jih prejme, razume, kaj oseba v opisani zgodbi ve o dejstvih situacije, pri čemer imata tako udeleženec kot oseba iz zgodbe enako točno informacijo o dejanskem stanju. Naloga z napačnimi prepričanji vsebuje neusklajenost med udeleženčevim vedenjem o situacijami ter prepričanjem o isti situaciji protagonista iz predstavljene zgodbe. Ta tip nalog je še posebej pomemben, saj preverja kompleksno sposobnost zanemarjanja lastnega vedenja o resničnosti in razumevanja možnosti, da imajo drugi posamezniki drugačna, tudi napačna, prepričanja o svetu[8].

Obstajajo kognitivne naloge, ki hkrati preverjajo tako kognitivno kot afektivno komponento sposobnosti teorije uma. Ena izmed takih nalog je »Strange Stories Test« (avtor Happé, 1994),[15] kjer udeleženci preberejo zgodbo o protagonistovem vedenju in morajo izkazati njeno razumevanje. Za uspešno razlago zgodbe je potrebno razumevanje mentalnih stanj protagonistov. Nekatera izmed mentalnih stanj v nalogi so kognitivne narave, nekatera pa afektivne, zato ta testna naloga pokriva obe področji sposobnosti teorije uma. Na podoben način deluje tudi naloga »Faux Pas« (avtorji Baron-Cohen in sodelavci, 1999),[15] kjer morajo udeleženci prepoznati protagonistove socialno neustrezne odzive. Tudi ta naloga preverja afektivno in kognitivno komponento sposobnosti teorije uma, saj zahteva prepoznavo občutkov kot tudi namenov protagonista.[8]

Uporaba teorije uma na področju robotike

Robotski sistem, ki bi imel razvite sposobnosti teorije uma, bi zmogel sodelovati v socialnih interakcijah med roboti in ljudmi na način, ki poprej niso bile mogoče. Tak stroj bi bil sposoben učenja od osebe z uporabo socialnih signalov na enak način, kot se uči človeški otrok; druge intervencije (kot je učenje iz zbirke podatkov, opisano zgoraj), ne bi bile potrebne. Prav tako bi bila taka tehnologija sposobna prepoznati cilje in želje oseb ali osebkov, s katerimi bi prišla v stik, ter se na ta način ustrezneje odzivala na njihova čustvena in miselna stanja. S tem bi imela možnost predvidevati odziv sogovornika in spremeniti svoje vedenje z ozirom na to.

Implementacija takega robotskega sistema je zahtevna, saj je za to potrebno pri robotih razviti veliko število koordiniranih procesov (zaznavnih, senzorno-motornih, pozornostnih in drugih kognitivnih). Primarna lastnost, ki jo mora imeti tak robot, je sposobnost razlikovati med živim in neživim, torej med objektom in subjektom, kar je tudi osnovna sposobnost človeške zbirke sposobnosti teorije uma.[16]

Predhodne raziskave so pokazale, da imajo roboti prejšnjih generacij pomanjkljive sposobnosti prilagajanja glede na človekova spremenljiva čustvena in motivacijska stanja, kar pogosto povzroči nezadovoljstvo človeka v socialnih interakcijah z roboti. Roboti s sposobnostjo teorije uma bi lahko omogočili kvalitetnejše tovrstne interakcije. Prvotne

implementacije teorije uma na področju umetne inteligence so bile osredotočene v glavnem na to, da so roboti prevzemali vidno perspektivo in upravljali s prepričanji, da bi razumeli, kakšno predstavo o svetu ima človek, s katerim so bili v interakciji. Uporaba teh metod je izboljšala socialne interakcije med človekom in robotom. Sodobnejši pristop se osredotoča na povratni inženiring človeških sposobnosti teorije uma, in sicer na sklepanje o človeških miselnih in čustvenih stanjih na podlagi vedenjskih pokazateljev. Naslednji veliki korak v sposobnostih umetno inteligentnih sistemov je zmanjšati razkorak med pravilnim sklepanjem o človeških čustvenih in miselnih pojavih ter sprejemanjem smiselne odločitve na podlagi prejetih informacij.[17]

Pomembno odprto vprašanje, na katerega naletijo eksperimentalni poskusi implementacije teorije uma na področju umetne inteligence, je iskanje koherentnega teoretičnega koncepta teorije uma. Če znamo na področju psihologije, medicine in družbenih znanosti opisati pojem sposobnosti teorije uma, ostajajo nejasnosti o procesih, ki sodelujejo pri uporabi sposobnosti teorije uma (npr. ali gre za izključno priučena znanja, ali pa sodelujejo notranji in prirojeni človeški mehanizmi, ki jih ni mogoče simulirati z uporabo umetne inteligence).[18] Prav tako je pomanjkljivo naše poznavanje nevroanatomskih korelatov procesov zavesti in socialne zavesti.[19] Ker je implementacija teh sposobnosti na področju umetne inteligence nujno odvisna od znanja in razvoja na področju človeške teorije uma,[18] pomanjkljivosti v našem razumevanju človeške teorije uma povzročajo težave pri tej implementaciji.

Trenutne implementacije teorije uma v robotiki

Trenutno obstaja nekaj delujočih robotov, ki sodijo v skupino »socialnih robotov«, torej robotov, ki so zmožni stopati v smiselne socialne interakcije z ljudmi[20]. Opisani roboti do določene mere uporabljajo sposobnosti teorije uma.

Najbolj poznan izmed socialnih roboto je Sophia, ki jo je razvilo podjetje Hanson robotics in jo prvič predstavilo leta 2016. Sophia je humanoidni robot, ki vsebuje napredno tehnologijo, s katero se lahko učinkovito vključuje v človeške socialne interakcije. Kamere v notranjosti oči v kombinaciji z algoritmom za predelavo zaznanih slik omogočajo, da »vidi« človeški obraz pred seboj, s čemer lahko sledi obrazni mimiki, vzdržuje očesni kontakt in prepozna posamezne obraze. Prav tako lahko procesira govorne podatke in se nanje v socialnem kontekstu smiselno odziva.[21] Potrebno pa je razumeti, da gre v tem primeru za zelo dovršeno simulacijo razumevanja in udeleževanja v komunikaciji, ne pa za dejansko procesiranje informacij, ki bi bilo primerljivo s človeškimi procesi socialne kognicije.[22]

Na podoben način delujeta tudi robota Pepper in NAO, ki ju je razvilo podjetje Software Robotics, in se uporabljata v izobraževanju ter na področju stikov s strankami (na primer v hotelirstvu),[23] ter robot ASIMO podjetja Honda and Kaspar, ki ga uporabljajo za pomoč pri delu z otroki s spektroavtistično motnjo.[24]

Samozavedna umetna inteligenca

Gre za končno stopnjo razvoja umetne inteligence, ki zaenkrat obstaja zgolj hipotetično.

Sposobnost samozavedanja

Da bi lahko govorili o robotih s sposobnostjo samozavedanja, potrebujemo najprej delovno definicijo zavesti. Barron in Klein sta leta 2015 napisala kontroverzni članek, ki je naletel na številne odzive v znanstveni skupnosti, v katerem utemeljujeta, da imajo žuželke primitivne živčne strukture, analoge strukturam višjih živali, ki le-tem omogočajo zavedanje lastnega telesa in gibanja tega v prostoru, kar je zadostni pogoj za samozavedanje.[25] Shelley Anne Adamo je v kritiki njunega prispevka izpostavila, da imajo glede na zastavljene kriterije sposobnosti samozavedanja že nekateri trenutno obstoječi roboti, ki zmorejo ustvarjati integrirano simulacijo sebe v prostoru z uporabo notranjih in zunanjih vhodnih podatkov.[26] Nekateri futuristi verjamejo v neizbežnost razvoja umetne inteligence s sposobnostjo samozavedanja in v to tudi vlagajo tako finančne kot intelektualne vire, a v znanstveni skupnosti vlada konsenz, da trenutno stroji še nimajo sposobnosti samozavedanja ter da še nimamo orodij, s katerimi bi lahko strojno zavest vzpostavili.[20] Problem oblikovanja umetne zavesti leži v številnih dilemah in nejasnostih, tako glede nevroanatomskih struktur, ki bi naj vzpostavljale funkcije zavesti, kot glede filozofskih dilem o tem, kaj zavest je in kako jo lahko opredelimo in preverimo.

Nevroanatomski korelati zavesti

Anatomski regiji, ki sta po dosedanjih dognanjih najtesneje povezane s pojavom zavesti, sta talamo-kortikalni sistem in ascendentni sistem vzburjenja. Ascendentni sistem vzburjenja in nekaj specifičnih talamičih jeder (vse strukture se nahajajo v možganskem deblu ali v njegovi neposredni bližini)

omogočajo difuzno aktivacijo obeh možganskih hemisfer. Ta aktivacija je ključna za vzpostavitev zavestnega zaznavanja. Aktivnost posameznih področij možganskih hemisfer korelira z vsebino zavestnega zaznavanja, vendar pa ni nobeno izmed višjih (kortikalnih) možganskih področij samo po sebi odgovorno za vzpostavitev zavedanja.[27]

Primerjalne analize so pokazale, da nekatera možganska stanja korelirajo z vzpostavitvijo zavesti. Študija iz leta 2003 ugotavlja, da se v stanju anestezije in komatoznem stanju zmanjša metabolna aktivnost v kortikalnih regijah možganov, specifično v področju prefrontalne in parietalne skorje.[28] Podobno so študije pokazale večjo metabolno aktivnost in povečano komunikacijo med kortikalnimi regijami ob ciljanem, zavestnem učenju novih nalog v primerjavi z izvajanjem naloge, ki je že priučena in avtomatizirana.[27] Elektroencefalografske študije so prav tako pokazale razlike v aktivnosti med budnim in zavedajočim stanjem ter komatoznim stanjem.[27]

Študije na področju nevroanatomskih korelatov zavesti so pomembne in koristne, vendar so dognanja pri pojasnjevanju funkcije zavesti lahko pomanjkljiva ali zavajajoča. V odmevnem članku iz leta 2015 Tononi in Koch opozarjata, da dognanja o možganskem dogajanju, ki naj bi sovpadala z zavestjo, niso enoznačna (npr. obstajajo stanja kortikalne metabolne ali električne aktivnosti brez zavesti). Poleg tega lahko gre pri opisanih nevroanatomskih korelatih zavesti dejansko za korelate aktivnosti, ki se zgodi pred samim zavestnim izkustvom ali tik za njim, ne pa za korelat samega izkustva. Definicija zavesti zgolj skozi prizmo nevroanatomskih korelatov je problematična tudi pri vprašanju o obstoju zavesti v primeru ljudi s pomanjkljivo razvitimi ali poškodovanimi strukturami, ki naj bi bile odgovorne za vzpostavitev zavesti (kot so pacienti po hudi poškodbi glave), ter seveda neprenosljiva na vprašanje obstoja zavesti živali in tudi možnosti obstoja zavesti strojev.[19]

Definicije zavesti v filozofiji

Filozofske šole, ki se ukvarjajo s problemom obstoja zavesti in njene definicije, se v grobem delijo v dve skupini – dualizem in monizem. Smer dualizma zagovarja, da človeško bit sestavljata dve različni in ločeni substanci – materialna in nematerialna. Telo je manifestacija materialne substance, medtem ko je um oziroma zavest manifestacija nematerialne substance. Dualizem v znanstveni skupnosti nima veliko privržencev zaradi nepreverljivosti hipotez in manjkajoče razlage, kako prideta nematerialno in materialno v medsebojno interakcijo, ki je kot rezultat očitna. Monizem zagovarja tezo, da je vse na svetu sestavljeno iz ene materije, pri čemer smer idealizma trdi, da je vse nematerialna substanca (kar v praksi pomeni, da izven naših misli materialni svet ne obstaja), smer materializma pa trdi, da je vse, tako fizično kot duhovno doživetje, produkt materialne substance. Materializem je najbolj obče sprejeta perspektiva na zavest v znanstveni skupnosti, še posebej pa njegova specifična struja, imenovana funkcionalizem, ki zagovarja trditev, da so mentalna ali duhovna stanja funkcionalna stanja možganov, kar pomeni, da njihovo kvaliteto duhovnosti oziroma mentalnosti definira njihova funkcija, ki je stanje substrata in mu je analogna. Z vidika funkcionalizma je razmeroma preprosto navesti določene funkcije zavesti – kognitivno kontrolo nad vedênjem, sposobnost integracije informacij, pozornosti, sposobnost opisati notranje dogajanje in o njem komunicirati. Prav tako si je mogoče predstavljati, da bi tovrstne funkcije lahko simulirali z uporabo robotske tehnologije. Težje pa si je zamisliti stroj, ki bi lahko izvajal funkcijo subjektivnega izkustva zavesti, se pravi pojasniti in simulirati občutek zavedanja oziroma »imeti zavest«.[27]

Problem druge zavesti

Vprašanje definicije zavesti odpira še eno pomembno vprašanje s področja epistemologije: kako lahko ugotovimo, da ima subjekt, s katerim smo prišli v stik, svojo notranjo zavest? Če lahko prisotnost svoje zavesti, bodisi kot korelat materialnega sveta bodisi kot ločeno duhovno substanco, potrdimo skozi dejanje samozavedanja,[29] pa enako ni mogoče trditi za druge zavesti. Na prisotnost njihovih notranjih stanj lahko sklepamo na podlagi njihovega vedênja in naše sposobnosti empatije in mentalizacije, ne moremo pa nikoli dostopiti do dejanskih notranjih stanj in s tem potrditi obstoja zavesti.[30] Aplikativne raziskave teorije uma pogosto zanemarjajo vprašanje druge zavesti in se osredotočajo na merljive korelate notranjih stanj, kar je tudi osnova raziskav na področju teorije uma v psihiatriji in nevroznanosti.[31] Vendar pa je brez odgovora na to temeljno vprašanje problematično vzpostavljati kriterije za ocenjevanje prisotnosti zavesti v primeru umetne inteligence.

Umetna zavest

Nejasnosti na področju razumevanja koncepta zavesti povzročajo tudi prepreke pri načrtovanju umetne inteligence s sposobnostjo učenja. Osnovno vprašanje, na katerega je potrebno poiskati odgovor, je: kako bomo vedeli, da smo ustvarili stroj s sposobnostjo samozavedanja? V preteklosti je bilo že več poskusov, kako definirati kriterije, po katerih bi

lahko ugotovili, da je robot samozaveden.[27] Najbolj poznan tovrsten poskus je Turingov test, ki preizkuša, ali je vedênje stroja nerazločljivo od človeškega vedênja.[32] Vendar pa, kot so kritiki testa opozarjali, pri tem ni mogoče reči, ali opazujemo dejansko delovanje zavesti ali zgolj simulacijo zavesti.[33] Kasnejši poskusi definicije kriterijev zavesti pri umetni inteligenci so se osredotočali na ugotavljanje dejanskih notranjih stanj in ne zgolj opazovanja vedênja. Aleksander in Dunmall sta leta 2003 predlagala pet preizkusov oziroma aksiomov, ki preverjajo prisotnost zaznavnih, emotivnih in domišljijskih notranjih stanj ter mehanizmov pozornosti in načrtovanja. Haikonen je leta 2007 predlagal test, ki preverja prisotnost notranjih predstav in notranjega govora, ter sposobnost opisati te vsebine ter jih prepoznati kot lastne. Leta 2010 je Arrables s soavtorji kot kriterij predlagal oceno prostorsko-časovnih vzorcev fizičnih stanj stroja v primerjavi s stanjem človeških možganov. Noben izmed naštetih poskusov vzpostavitve kriterijev zaenkrat ni univerzalno sprejet.[27]

Zaključek

Umetna inteligenca je področje hitrega razvoja, veliko aplikacij umetne inteligence pa že uporabljamo vsakodnevno v naših domovih ali na strokovnem in raziskovalnem področju. Ob napredku sposobnosti umetne inteligence postaja delo s tovrstnimi stroji manj podobno uporabi in bolj podobno interakciji oziroma sodelovanju. Za izboljšanje sodelovanja med človekom in strojem je pomemben naslednji korak v razvoju tehnologij umetne inteligence – razvoj strojev s sposobnostjo teorije uma oziroma izboljšanimi sposobnostmi socialnih interakcij, ki bi robotu omogočale razumevanje človeških mentalnih stanj in ustrezno odzivanje nanje. Poznavanje in razumevanje področja človeških socialnih kognitivnih sposobnosti je korak do ustvarjanja strojev s temi sposobnostmi. Ustvarjanje umetno inteligentnih strojev s sposobnostjo teorije uma je nujni premostitveni korak do razvoja umetne inteligence s sposobnostjo samozavedanja, ostaja pa odprto vprašanje, ali je slednje tehnično in metafizično sploh mogoče.

Poleg tega samozavedna umetna inteligenca odpira številne etične pomisleke, o katerih je potrebno poglobljeno razmišljati vzporedno z razvojem tehnologije. Ravnanje samozavedne umetne inteligence bi določala neke vrste morala ali etika, pri tem pa bi se soočali z vprašanjem, kako zagotoviti, da bi etična načela samozavednih strojev delovala za dobrobit človeka oziroma usklajeno s cilji človeštva. Odprla bi se tudi pravna vprašanja, in sicer, kako opredeliti pravice in odgovornosti samozavednih strojev ter kako zagotoviti njihovo spoštovanje in izpolnjevanje.[34] Psihologija, medicina, filozofija, nevroznanost in družbene znanosti lahko tukaj znanstveno-tehničnemu področju umetne inteligence nudijo pomembne odgovore in razumevanje problematike za nadaljnji razvoj.

Reference

1. Poole D, Mackworth A, Goebel R: *Computational intelligence: a logical approach*. New York 1998: Oxford University Press.
2. Joshi N: 7 types of artificial intelligence. *Forbes* 2019 https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/#18f25e2b233e (4. 4. 2020)
3. Deo RC: Machine learning in medicine. *Circulation 2015; 132*(20): 1920-1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593
4. Cao C, Liu F, Tan H et al.: Deep learning and its applications in biomedicine. *Genomics Proteomics Bioinformatics* 2018; *16*(1): 17-32. https://doi.org/0.1016/j.gpb.2017.07.003
5. Ching T, Himmelstein DS, Beaulieu-Jones BK et al.: Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018; *15*(141): 20170387. https://doi.org/10.1098/rsif.2017.0387
6. Gordon G: Social behaviour as an emergent property of embodied curiosity: a robotics perspective. *Philos Trans R Soc B Biol Sci* 2019; *374*(1771): 20180029. https://doi.org/10.1098/rstb.2018.0029
7. Arioli M, Crespi C, Canessa N: Social cognition through the lens of cognitive and clinical neuroscience. *BioMed Res Int* 2018: 4283427. https://doi.org/10.1155/2018/4283427
8. Henry JD, Cowan DG, Lee T, Sachdev PS: Recent trends in testing social cognition. *Curr Opin Psychiatry* 2015; *28*(2): 133-140. https://doi.org/10.1097/YCO.0000000000000139
9. Call J, Tomasello M: Does the chimpanzee have a theory of mind? 30 years later. *Trends Cogn Sci* 2008; *12*(5): 187-192. https://doi.org/10.1016/j.tics.2008.02.010
10. Premack D, Woodruff G: Does the chimpanzee have a theory of mind? *Behav Brain Sci* 1978; *1*(4): 515-526. https://doi.org/10.1017/S0140525X00076512
11. Schaafsma SM, Pfaff DW, Spunt RP, Adolphs R: Deconstructing and reconstructing theory of mind. *Trends Cogn Sci* 2015; *19*(2): 65-72. https://doi.org/10.1016/j.tics.2014.11.007
12. Cherry K: How the theory of mind helps us understand others. *VeryWell Mind* 2020. https://www.verywellmind.com/theory-of-mind-4176826 (13. 5. 2020)
13. Wellman HM, Fang F, Peterson CC: Sequential progressions in a theory-of-mind scale: longitudinal perspectives. *Child Dev* 2011; *82*(3): 780-792. https://doi.org/10.1111/j.1467-8624.2011.01583.x

14. Bora E, Berk M: Theory of mind in major depressive disorder: a meta-analysis. *J Affect Disord* 2016; *191*: 49-55. https://doi.org/10.1016/j.jad.2015.11.023
15. Beaudoin C, Leblanc É, Gagner C, Beauchamp MH: Systematic review and inventory of theory of mind measures for young children. *Front Psychol* 2020; *10*: 2905. https://doi.org/10.3389/fpsyg.2019.02905
16. Scassellati B: Theory of mind for a humanoid robot *Auton Robots* 2002; 12(1): 13-24. https://doi.org/10.1023/A:1013298507114
17. Görür OC, Rosman Benjamin S, Hoffman G, Albayrak S: Toward integrating theory of mind into adaptive decision-making of social robots to understand human intention. In: *Workshop on "The Role of Intentions in Human-Robot Interaction" in 12th ACM/IEEE International Conference on Human-Robot Interaction*. Viena 2017.
18. Bianco F, Ognibene D: Transferring adaptive theory of mind to social robots: insights from developmental psychology to robotics. In: Salichs M et al. (eds), *Social Robotics. Proceedings, 11th International Conference, ICSR 2019 November 26-29, 2019,*. Madrid 2019: Springer; 77-87. https://doi.org/10.1007/978-3-030-35888-4_8
19. Tononi G, Koch C: Consciousness: here, there and everywhere? *Philos Trans R Soc Lond B Biol Sci* 2015; *370*(1668): 20140167. https://doi.org/10.1098/rstb.2014.0167
20. Meissner G: Artificial intelligence: consciousness and conscience. *AI Soc* 2019; 35(2): 225-235. https://doi.org/10.1007/s00146-019-00880-4
21. Sophia the robot takes her first step. *The Telegraph* 2018. https://www.telegraph.co.uk/technology/2018/01/08/sophia-robot-takes-first-steps/ (19. 1. 2020)
22. Fitzsimmons C: Why Sophia the robot is not what it seems. *The Sydney Morning Herald* 2017. https://www.smh.com.au/opinion/why-sophia-the-robot-is-not-what-it-seems-20171031-gzbi3p.htm (19. 1. 2020)
23. SoftBank Robotics Europe. *Pepper and NAO in the service of Education and Research*. https://www.softbankrobotics.com/emea/en/industries/education-and-research (19. 1. 2020)
24. ASIMO by Honda: *The world's most advanced humanoid robot*. https://asimo.honda.com/ (19. 1. 2020)
25. Barron AB, Klein C: What insects can tell us about the origins of consciousness. *Proc Natl Acad Sci* 2016; *113*(18): 4900-4908. https://doi.org/10.1073/pnas.1520084113
26. Adamo SA: Consciousness explained or consciousness redefined? *Proc Natl Acad Sci U S A* 2016; *113*(27): E3812. https://doi.org/10.1073/pnas.1606942113
27. Reggia JA: The rise of machine consciousness: Studying consciousness with computational models. *Neural Netw* 2013; *44*: 112-131.
28. Baars BJ, Ramsøy TZ, Laureys S: Brain, conscious experience and the observing self. *Trends Neurosci* 2003; *26*(12): 671-675. https://doi.org/10.1016/j.tins.2003.09.015
29. Encyclopedia Britannica. *Cogito, ergo sum*. https://www.britannica.com/topic/cogito-ergo-sum (19. 1. 2020)
30. Avramides A: Other Minds. In: Zalta EN (ed.): *Stanford Encyclopedia of Philosophy*. Stanford 2020: The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University. https://plato.stanford.edu/archives/win2020/entries/other-minds/ (19. 1. 2020)
31. Leudar I, Costall A: On the persistence of the 'problem of other minds' in psychology: Chomsky, Grice and theory of mind. *Theory Psychol* 2004; *14*(5): 601-621. https://doi.org/10.1177/0959354304046175
32. Alan Turing Scrapbook. *The Turing test, 1950*. https://www.turing.org.uk/scrapbook/test.html (19. 1. 2020)
33. Searle JR: Minds, brains, and programs. *Behav Brain Sci* 1980; *3*(3): 417-424. https://doi.org/10.1017/S0140525X00005756
34. Müller VC: Ethics of artificial intelligence and robotics. In: Zalta EN (ed.): *Stanford Encyclopedia of Philosophy*. Stanford 2020: The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University. https://plato.stanford.edu/archives/win2020/entries/ethics-ai/ (2. 3. 2021)