

Andrej Ule

Filozofska fakulteta Univerze v Ljubljani

Wittgensteinska kritika zamisli o mislečih strojih

V sestavku analiziram nekaj Wittgensteinovih kritik domnev, ki so ključne za velik del kognitivne znanosti. Menim, da teoretiki kognitivne znanosti še niso resno vzeli Wittgensteinovih kritik in začuda pogosto zamenjujejo vprašanje »kako kaj deluje« z vprašanjem »kaj to počne«. Predvsem pa je usodno zamenjavanje notranjih računskih (ali vzporednih) procesov v možganih, ki morda povzročajo mentalna stanja, z zunanjimi, socialno utemeljenimi vsakodnevnimi kriteriji prepoznavanja, klasifikacije in poznavanja vsebine mentalnih stanj.

Ključne besede: računski procesi v možganih, misleči stroj, Wittgenstein, Turing, sledenje pravilu

Znana je Wittgensteinova kritika ideje o strogo zasebnem jeziku, tj. jeziku, ki bi bil razumljiv in »dosegljiv« le posamezniku oz. le njegovi zavesti, možganom ipd. O mišljenju po Wittgensteinu namreč lahko govorimo le v kontekstu intersubjektivnih dejavnosti ljudi, ne zasebno, namreč v okviru sledenja pravil v različnih jezikovnih praksah in v medsebojnem prepletu raznih človeških dejavnosti. Teh dejavnosti ne moremo prevesti nazaj na procese v posameznikih ali na skupen njihovih posamičnih vedenj.

Wittgenstein je bil zato zelo skeptičen do zamisli o mišljenju (in podobno o zavesti) kot »notranjem procesu«, ki poteka v možganih posameznika. To zamisel je odklanjal prav tako močno kot metafizični psihofizični dualizem. Zgolj notranje-možganski proces mišljenja bi bil nekako sam sebi dovolj. To bi ga v končni fazi naredilo za prav tako »zasebnega«, kot so mentalna stanja in mentalni procesi domnevne nematerialne duševne substance. Dokler nimamo podane intersubjektivno preverljive (javno dostopne) in sistematične zveze med domnevnimi notranjimi stanji ali procesi v možganih in delovanjem posameznika v intersubjektivnih (socialnih) kontekstih, se izpostavljamu Wittgensteinovi kritiki zasebnega jezika, zasebnih stanj in procesov itd. Res lahko domnevamo, da takšna zveza ni načelno izključena in jo bomo morda kdaj lahko eksplicirali, vendar pa ostaja

mučen prelom med javno dostopno prakso delovanja ljudi v različnih situacijah (npr. jezikovnih igrach), ki določa pomen našim doživljanjem, željam, nameram, mislim, in notranjo, duševno ali »možgansko« naravo oz. strukturo teh doživljanj, želj itd.

Opozarjam, da ne gre le za konceptualni razkol med domnevno povsem računsko (ang. *computational*) določenim substratom mentalnih stanj in našim faktičnim doživljanjem teh stanj, temveč za razkol med tem, *kako delujejo* domnevni procesorji, ki tvorijo fiziološko-fizikalno osnovo mentalnih stanj, ter tem, *kaj* in *kako smo mi*, ko doživljamo ta stanja in jih izražamo v svoji družbeni praksi. K temu, kako smo, pa bistveno sodi, da ima naše doživljanje in delovanje svoj *pomen za nas*, da prav skozi svoje doživljanje in delovanje *občutimo svoj obstoj*. To ni le konceptualna razlika, temveč kategorialna razlika, za katero se nam niti ne sanja, kako bi jo lahko premostili.

Že v pojmu delovanja procesorjev tiči dvoumnost, namreč med opisom vzročnega procesa, tj. zaporedja stanj, procesorjev in pravilom delovanja procesorjev. Prvo je možno doseči s tem, da navedemo fizikalne zakone, ki regulirajo fizikalni potek dogajanja, in začetna stanja procesorjev v določenih časovnih trenutkih, drugo pa s tem, da navedemo npr. operacijsko tabelo prehodov stanj procesorja v odvisnosti od vhodnih podatkov in trenutnega stanja procesorja. Toda ta tabela je povsem idealizirana tvorba, ki kaže, kako *naj bi* procesor deloval v idealnih pogojih, ne pa, kako dejansko deluje. Je en od opisov *pravila* delovanja, ki naj bi ga implementiral procesor, ne opis procesa. Takšna pravila lahko včasih podamo s tabelo operacij, ki jih procesor lahko opravi ob ustreznih vhodnih podatkih in svojih notranjih stanjih. Kakorkoli že opredeljujemo pravila delovanja procesorja, moramo razlikovati med idealiziranim *operacijskim* delovanjem in dejanskim *potekom*, tj. dejanskim zaporedjem fizičnih stanj procesorja. Operacijsko delovanje procesorja je tisto, kar mu mi *pripišemo*, ko ga *tolmačimo* kot implementacijo določenega pravila ali tabelarično opredeljene operacije.

Vsaj v prvi, »računalniški« fazi kognitivne znanosti, so se avtorji pogosto sklicevali na model Turingovega stroja, ki naj bi predstavljal splošno formulacijo za računsko delovanje naravnih ali umetnih procesorjev. Vendar je ta model le idealizirana struktura delovanja sistemov, ki jim *pripisujemo* računsko delovanje, temu modelu pa ne ustreza noben realni sistem. Razlika ni le v tem, da ima Turingov stroj neskončen spominski trak, temveč tudi v tem, da se ta »stroj« lahko po nekaj korakih računanja vrne v natanko ista stanja, medtem ko se realni (končni) avtomati nikoli zares ne vrnejo v isto stanje, pač pa le v istovrstno stanje, ki sodi v enak ekvivalenčni razred stanj glede na določeno ekvivalenčno relacijo med stanji. Zato je vprašanje, ali pri njih sploh gre za računanje v smislu *operiranja s simboli*.

Cris Fields je v svojem prispevku »*Measurement and Computational Description*« (1996) v zborniku, posvečenem Turingu, izrecno poudaril, da moramo

razlikovati med vedenjem fizičnega sistema v času in (dinamičnim) opisom tega sistema s pomočjo algoritma, ki določa spreminjanje sistema v odvisnosti od (diskretnih) sprememb določenega parametra (ali parametrov), ki je opazljiv (merljiv) v času dogajanja. Tu gre za *naš* način opisa, ki omogoča preslikavo sprememb stanj sistema v korake algoritma. Pri tem moramo imeti na voljo določeno interpretacijo, ki preslika stanja sistema v zaporedje abstraktnih računskih stanj, pri čemer vsako stanje a_k sistema pridobimo z uporabo prvih k korakov algoritma ob določenem inputu (Fields, 1996: 170). Fields opozarja, da je takšna preslikava možna le, če vse operacije, ki so navedene v procesu preslikave, algebraično komutirajo (tj. kvantnomehansko gledano, ustrezne meritve ali opazovanja procesa so medsebojno neodvisna). Fields se nanaša na E. Dietricha, ki je opozoril, da lahko takšno interpretacijo sestavimo le tedaj, če sistem že vnaprej in neformalno jemljemo kot sistem, ki umešča (instancializira) funkcijo, ki bo izračunana z zadanim algoritmom, in če k korakov algoritma prepoznamo kot izračune poprej podanih podfunkcij (Dietrich, 1989).

Fields govori o takšni interpretaciji sistema kot o »virtualnem stroju« za dani algoritem. Takšna interpretacija pa je možna le tedaj, če lahko vedenje sistema od začetnega stanja (razumljenega kot vhodni podatek algoritma) vedno interpretiramo kot izvajanje danega algoritma pri danem vhodnem podatku. To je seveda le idealizirana predpostavka, oz. boljše predpostavka idealizacije, ne realni potek dogajanja v sistemu. Sistem torej »računa« (ang. *computes*) le v okviru naše interpretacije kot virtualni stroj, *ne po sebi*. Dinamičen opis sistema predstavlja le določeno relacijo, ki dovolj natančno podaja zaporedje (merljivih, opazljivih) fizičnih stanj sistema v zaporednih časovnih intervalih. To zaporedje pogosto lahko opišemo s kakšno lepo tekočo funkcijo, vendar to ni bistveno, to je že stvar interpretacije. A za računsko delovanje sistema bistveno potrebujemo takšno interpretacijo, ki zaporedna stanja sistema predstavi kot rezultate uporabe določenega algoritma (tj. sistema operacij) na ustreznih vstopnih stanjih sistema, ki predstavlja jo vhodne podatke virtualnemu stroju.

Fields dalje izvaja, da lahko vsak sistem, ki ga interpretiramo kot algoritemsko delujoči stroj, tudi *dejansko je tak* stroj, kajti že *simulirano računanje* je računanje. Torej takšna interpretacija dela tak sistem različnega od neinterpretiranega dinamičnega procesa, kot je npr. tok tekočine (Fields, 1996: 171). Drugi Fieldsov zaključek je »da lahko isti fizični sistem na podlagi različnih množic meritev interpretiramo kot različne virtualne stroje, namreč z interpretacijo vrednosti različnih množic spremenljivk kot pokazateljev stanja sistema« (prav tam).

Fieldsovo tezo, da je vsaka popolna simulacija računanja že tudi računanje, moramo brati tudi v obratni smeri, namreč da je vsako računanje svoja popolna simulacija, namreč takšna *interpretacija delovanja* sistema, ki ga predstavi kot popolno računanje (tj. vsako stanje sistema predstavi kot računski rezultat določenih vhodnih podatkov in notranjih stanj sistema). Potemtakem ni računanja po sebi,

tj. nobeno računanje ni potek dogodkov brez ustrezne (računske) interpretacije. To misel lahko povežemo z zanimivo Wittgensteinovo kritiko izenačevanja operacije in funkcije (predikacije), oz. operacij in relacij, ki jo je podal že v *Tractatusu*. Wittgenstein je namreč trdil, da so operacije sestavina našega simbolizma, oz. načina prikaza odnosov med formami stanj stvari, ne pa stanja stvari, dejstva ali dogodki v svetu. Operacije se lahko nanašajo na lastne rezultate kot bazo za operiranje, omogočajo matematična zaporedja predmetov, lahko izničujejo rezultate drugih operacij ali celo same sebe (Wittgenstein, 1976: 5.251, 5.252, 5.254). Predikacije tega načeloma ne zmorejo. Od tod med drugim izhaja tudi ugotovitev, da operacije niso nobena stanja stvari ali dejstva v svetu, temveč so *naš način prikaza* formalnih odnosov med različnimi opisi stanj stvari. V tem smislu tudi računanje, v kolikor so resnične operacije, podane s shemami Turingovega stroja, ne poteka »v naravi«, temveč le v *simbolnih sistemih* prikaza delovanja stroja. Za računanje je nujen operacijski (simbolni) opis dogajanja. Sledeč Wittgensteinu, mora računanje zajemati zaporedno (včasih tudi zaporedno in vzporedno) izvajanje operacij, ki jih lahko predstavimo le v določenem simbolnem sistemu, tj. ob uporabi ustreznih *formalnih izrazov*, spremenljivk, tabel delovanja ali algoritemskih postopkov. To pa ni isto kot opis »naravnega poteka« dogajanja v »stroju«, kajti ta opis ni nujno operacijski, lahko bi bil npr. le časovni diagram dogajanja.

V svoji kasnejši filozofiji je Wittgenstein opozarjal na podobno razliko med vzročnim delovanjem strojev in operacijskim delovanjem. Slednje po njegovem lahko poteka le v okviru predstave o stroju kot »svojem lastnem simbolu«. Te misli so zlasti pomembne za razpravo o možnosti mislečih strojev oz. za »strogi program« umetne inteligence. Po Wittgensteinu lahko stroj pojmujeemo idealizirano, tj. tako, da abstrahiramo od možnih okvar, nepopolnosti delovanja itd. V tem primeru se zdi, da »[s]troj ali sliko stroja uporabljamo kot simbol za določen način delovanja« (Wittgenstein, 2014: par. 193.).¹ To je povsem različno od napovedi dejanskega vedenja stroja. V slednjem primeru na splošno ne pozabimo na možnost napak, poškodb njegovih delov itd. Realni stroj se vedno lahko giblje povsem drugače kot napoveduje njegov simbolni dvojnik. Za simbolni dvojnik je možnost odpovedi *a priori*, tj. gramatično nemogoča. Wittgenstein zaključuje: »[P]remikanje stroja kot simbola je vnaprej določeno na drug način kakor premikanje danega dejanskega« (prav tam).

Toda tudi govor o idealnem ali idealiziranem stroju, ki naj bi se nahajal v stroju-simbolu, zavaja, npr. v govor, da »stroj ima (poseduje) svoje možnosti premikanja«. Wittgenstein ugotavlja, da ta idealna možnost gibanja ni gibanje, pa tudi ni podajanje fizičnih pogojev gibanja, saj se dejanski stroj lahko giblje povsem drugače, kot mu predpisuje njegova idealna podoba. Wittgenstein pravi, da se nam zdi, kot da je idealna možnost premikanja neke vrste »senca samega premikanja«.

¹ Podobno v Wittgenstein (1964: I., par. 125).

»Rečemo 'Izkustvo bo pokazalo, ali bo to dalo temu tečaju tako možnost premikanja', ne pa tudi 'Izkustvo bo pokazalo, ali je to možnost tega premikanja'« (prav tam: par. 194).²

V takšnem govorjenju napačno tolmačimo svoj način govora, ko zamenjujemo govor, ki opisuje operiranje v stroju-simbolu, z opisom realnega delovanja stroja. Wittgenstein pravi, da »[k]adar filozofiramo, smo kakor divji ljudje, ki slišijo način izražanja civiliziranih ljudi, ga napačno interpretirajo in iz te interpretacije nato izpeljejo najbolj nenavadne zaključke« (prav tam). Skratka v stroju-simbolu gre za poseben *izraz pravil* delovanja, oz. delovanja po pravilu, ne za dejansko delovanje. Podobno bi npr. »stroj za konjunkcijo«, podan z ustrezno tabelo prehodov stanj, predstavljal simbol za delovanje po pravilu, ne pa dejanski potek delovanja kakega »logičnega« stroja, npr. preprostega procesorja, ki implementira konjunkcijo.

Zopet smo torej pred načelno razliko med operacijami v nekem simbolnem sistemu (modelu) in med realnimi procesi v svetu. Še tako velika popolnost v realnem delovanju kakega stroja, npr. stroja, ki bi implementiral kako logično operacijo, ne pomeni, da stroj *izvaja* to operacijo, kajti še tako popolni stroj se lahko polomi, odpove, zastane ipd., medtem ko *izvajanje* te operacije načelno izključuje takšno možnost. Te možnosti so po Wittgensteinu kategorialno, gramatično izključene.

Vendar pa Wittgenstein ni dopuščal, da bi katerikoli izraz pravil oz. sledenja pravilom pomenil tudi dokončno ali izvirno *interpretacijo* ali *razlago* sledenja pravilu. V tem smislu tudi ne bi brezpogojno pristal na Fieldsovo formulacijo, da računska interpretacija delovanja sistema po pravilu računanja predstavlja že dejansko sledenje pravilu računanja in je torej simulacija računanja že računanje. Če namreč računska interpretacija pomeni le simbolni *izraz* preslikave fizičnih stanj sistema v algoritemska stanja računskega sistema, potem še ne gre za računanje. Šele *izvajanje* (ang. *executing*) te preslikave, torej dejansko izvajanje algoritma, je računanje. V tem smislu je računanje šele *dejanska* računske interpretacija, ne pa npr. le zamišljena interpretacija, ki zgolj formalno opiše algoritem računanja. Tabela preslikav ali tabela prehodov še ni operacija, pač pa je njen simbolni izraz. Če nekoliko pomešamo Wittgensteinovo terminologijo iz *Tractatusa* in v *Filozofskih raziskavah*, bi dejali, da se dejansko računanje le *kaže* v uporabi določenega algoritma, npr. v družbeni in javno dostopni *praksi* (Wittgenstein, 2014: par. 202), *tehniki* (par. 199) ali *običaju* (prav tam), v katerih moramo sami smiselno sodelovati, *opišemo* pa jo lahko s kakim formalnim simbolizmom.

Seveda se lahko vprašamo, kaj pa mi sami, ljudje, ali vsaj mi *dejansko operiramo* in podobno, vsaj mi *dejansko sledimo pravilu* in ne le zgolj interpretiramo svoje vedenje kot operiranje ali sledenje pravilom? Po Wittgensteinu oboje dejansko

² Enako v Wittgenstein (1964: I par. 125).

počnemo. To pa zato, ker je naše operiranje in sledenje pravilom *skupno uvežbana praksa*, ki jo pridobimo v delovanju, ki je v skladu z *življenjsko formo ljudi*. Kajti sledenje pravilom (in izvajanje operacij, ki je le posebna vrsta sledenja pravilom) je delovanje, ki poteka kot nekaj nedvornega in drugim ljudem razumljivega. To nedvornost moramo enostavno sprejeti kot neko osnovno dejstvo: »To, kar je treba sprejeti, kar je dano – bi lahko rekli –, so *življenjske forme*.« (prav tam II: par. 345). Človeška življenjska forma ni nekaj fiksnega, temveč je naš skupni način delovanja. Je preplet dejavnosti, ki opredeljuje naše pojmovanje in delovanje. O tem piše Wittgenstein v *Zvezkih (Zettel)* (1970):

Kako lahko opišemo človeški način delovanja? Pač, le tako, da prikažemo, kako vseprek mrgoli (ang. *swarms*) delovanje različnih ljudi. Ne, kar *sedaj nekdo misli*, posamezno dejanje (ang. *action*), temveč vse mrgolenje človeškega delovanja, je ozadje, na katerem opazujemo vsako dejanje, ki določa našo sodbo, naše pojme in reakcije. (Wittgenstein, 1970: par. 567)

Pravila oz. sledenja pravilom, prav tako operacije in izvajanja operacij so za Wittgensteina neločljiva sestavina človeške življenjske oblike oz. »mrgolenja« človeškega delovanja. Zato sledenja pravilom ne moremo abstrahirati in ga opazovati ločeno od tega ozadja, npr. tako, da bi se spraševali, ali lahko kak stroj ali organizem sam po sebi sledi pravilom.

Ker mišljenje predpostavlja sposobnost sledenja pravilom, se po Wittgensteinu tudi ni smiselno spraševati o tem, ali npr. stroj misli. Oz. točneje, takšno spraševanje bi predpostavljalo določeno podobnost človeku. O tem je tudi izrecno pisal v *Filozofskih raziskavah*:

Bi stroj lahko mislil? – Bi lahko imel bolečine? – No, ali bi o človeškem telesu lahko dejali, da je tak stroj? Temu, da bi bil tak stroj, je vendar kar sedita blizu. (Wittgenstein, 2014: par. 360).

Toda stroj vendar ne more misliti! – Je to izkustveni stavek? Ne. Le o človeku in tistem, kar mu je podobno, rečemo, da misli. To rečemo tudi o lutkah in celo o duhovih. Oglej si besedo 'misliti' kot orodje! (Prav tam: par. 361)

Wittgenstein torej vendarle ne zavrne povsem možnosti, da stroj misli, a ob predpostavki določene, četudi virtualne podobnosti s človekom. V čem bi bila takšna podobnost? Seveda, v tem, da si lahko zamišljamo delovanje stroja kot sestavino človeške življenjske oblike, tj. kot smiselni del mreže človeške dejavnosti, npr. mreže komuniciranja med ljudmi, sodelovanja ljudi v skupnih projektih, igranja itd. Lahko rečemo le to, če si lahko zamišljamo, da bi kak robot dobro posnemal delovanje ljudi v paradigmatičnih življenjskih situacijah ljudi in se enakopravno vključeval v komunikacijo med njimi, potem si ga lahko zamišljamo kot mislečega. Prav podobno si npr. zamišljamo vedenje »živih« lutk (npr. Ostržka) ali duhov. Toda od zamišljanja do uresničenja je dolga pot, morda nemogoča.

Wittgenstein pa z zadnjim stavkom zgoraj namigne, da je to podobno, kot da bi bilo mišljenje neko orodje. Ali je mišljenje lahko orodje? Morda nekako tako, kot so naše roke »orodje« za delo, ki se je razvilo v evoluciji človeka. Seveda lahko tudi tako gledamo na mišljenje, a to je le en vidik, ki pa ne zajame celote. Lahko gledamo na mišljenje *kot* orodje (evolucije, narave, boga itd.), toda prav to, da gledamo nanj kot na orodje, kaže, da mišljenje samo po sebi ni orodje. Na človeka npr. ne morem gledati, kot da je živ, ali kot da misli, temveč ga imam preprosto za živega, mislečega itd. Lahko pa gledam nanj, kot da je lutka, a to ni njegova narava. Podobno lahko na lutko ne gledam, kot da je simulacija človeka, temveč (za nas) *je* simulacija človeka. Lahko pa gledam nanjo kot živo, mislečo bitje, vendar pa to ni njena narava.

Wittgenstein se je med drugim kritično obračal tudi na Turingovo zamisel računalniških strojev, ki mislijo. V *Beležkah o filozofiji psihologije* je npr. l. 1947 zapisal:

Turingovi »stroji«. Ti stroji so vendar *ljudje*, ki računajo. To, kar pravi, bi lahko izrazili tudi v obliki *iger*. In sicer bi bile zanimive takšne igre, pri katerih bi po določenih pravilih prišli do nesmiselnih napotkov. Mislim na igre, podobne »tekmi«. Recimo, da dobite npr. ukaz »Nadaljuj enako dalje«, in to tedaj, ko je to povsem nesmiselno, npr. ko zaidete v krog; kajti ta ukaz ima smisel, vendar le na določenih mestih (Watson.) (Wittgenstein, 1984: par. 1096).

Wittgenstein sicer ni zapisal, kaj bi se zgodilo z ljudmi, ki bi se znašli v tovrstni igri in pred podobno nesmiselnimi napotki, verjetno pa je domneval, da bi ljudje prepoznali nesmisel takšnih napotkov in nekako »izstopili« iz igre.³ Wittgensteinova tiha domneva je, da v kolikor bi bili Turingovi stroji res zgolj računaljoči stroji, ne pa ljudje, ki računajo, potem ne bi prepoznali nesmisla napotkov in bi še dalje skušali »delovati« v skladu z njimi. Če pa bi »Turingovi stroji« ravnali podobno kot ljudje, tj. bi zaznali nesmisel napotka, potem bi bili dejansko »ljudje, ki računajo«, ne pa (kot je morda domneval Turing in za njim domnevajo zagovorniki stroge umetne inteligence), »stroji, ki se vedejo kot ljudje«. Zato pravi, da so »Turingovi stroji« pravzaprav ljudje, ki računajo.

Wittgenstein je nato dal še primer matematičnega postopka, ki po njegovem privede do podobnega brezpotja, namreč varianto Cantorjevega diagonalnega dokaza, da je množica naravnih (in podobno racionalnih) števil »šibkejša« kot množica realnih števil. Vsak poskus števne »urejanja« vseh števil, zapisanih kot neskončna decimalna števila (npr. vsa realna števila med 0 in 1) privede do paradoksa, ko pridemo do napotka, da zapišemo *n*-to decimalno *n*-tega števila, kajti tedaj bi mora biti *n*-ta decimalna obena enaka in različna od same sebe. To je po Wittgensteinu podobno kot napotek »Nadaljuj enako dalje«, torej do nesmiselnosti.

³ Wittgenstein ni navedel, na kateri Turingov tekst se je tu nanašal, domnevam, da gre za tedaj znan Turingov tekst o izračunljivih številih (Turing, 1937).

ga napotka in s tem do odkritja nesmiselnega pravila. Tudi tu ljudje lahko prepoznamo nesmiselnost, medtem ko stroj tega ne bi zmožel.

Posredno je s tem namignjeno, da Turingovi stroji ne bi prestali »Turingovega testa« (ta test je sicer Turing formuliral šele l. 1950 in ni znano, da bi Wittgenstein poznal njegov sestavek, ampak ideja za ta test je bila pri Turingu prisotna že precej pred tem), ker bi slej ko prej zašli v nesmiselne, neizvršljive korake pravil in se skušali še dalje »ravnati« po njih, medtem ko naj bi ljudje hitro prepoznali njihov nesmisel. Domnevam, da gre za tezo, da se ljudje sicer naučijo pravil s »slepim« (nedvomnim) sledenjem pravil, vendar pa to ne pomeni, da vseskozi slepo sledijo pravilu. Če se izkaže, da jih sledenje pravilu vodi v slepo pot, to opazijo in se ustavijo ali popravijo pravilo, medtem ko naj bi stroji tega ne zmoogli, ker pač zgolj »mehanično« sledijo zadanemu pravilu. To pa zato, ker poteka sledenje pravilom pri ljudeh v skupnostih, kjer se tudi naučijo smiselnega sledenja pravilom, smiselnega seveda glede na življenjsko formo ljudi.

Kot ugotavlja tudi Stuart Shanker v svojem sestavku o Wittgensteinu, Turingu in Churchu, je bila temeljna Turingova napaka (v njegovem spisu o izračunljivih številih iz 1930, na katerega se je Wittgenstein nanašal v svojih kritičnih opazkah o »Turingovem stroju«) v tem, da je nevede prešel od enostavnih podpravil, ki sestavljajo kompleksni aritmetični algoritem, k sklepu, da se takšna pravila lahko uresničujejo povsem mehanično. Vendar pa dejstvo, da ljudje lahko takšnim pravilom sledimo povsem slepo (brez razmišljanja, brez zavesti), nikakor ne dopušča sklepa, da so sama po sebi nekognitivna in da jim lahko *sledi* tudi kak stroj (Shanker, 1987: 641).

Wittgenstein bi seveda zavrnil sleherno sklicevanje na bojda povsem »notranja« (mentalna, možganska) stanja »razumevanja pravil«, ki naj bi omogočali in nadzorovali sledenje pravilom. Seveda bi Turingovi zagovorniki (ali zagovorniki sodobne UI) odgovorili, zakaj pa tega ne bi zmogli tudi kaki sodobni računalniki, še zlasti, ker je danes mogoče tudi računalnike učiti pravil v kaki »računalniški skupnosti« z »računalniško formo življenja«. Ne vem, kako bi Wittgenstein odgovoril na ta izziv, morda bi dejal: Če je res tako in bi res zgradili take računalniške sisteme, ja, potem bi pač imeli opravka z dejanskim sledenjem pravilom in bi morali takšne računalnike *obravnavati kot ljudi*, ne pa kot *stroje*. To bi verjetno vso razpravo postavilo na glavo, saj bi od tu dalje »normalni« zahodnjaki skušali dokazovati, kako stroji pač ne morejo biti ljudje, namreč zavestno misleča bitja. V to razpravo se na tem mestu ne bom spuščal, ker je podobno brezploдна kot razprava o tem, ali lahko stroji mislijo. Poudarjam pa, da Wittgenstein zavrača vsako argumentiranje v korist ljudi zato, ker imajo svoja »notranja mentalna stanja«, stroji pač ne, ker to vnaša v razpravo domnevo o »zasebnem jeziku«, ki jo je Wittgenstein na široko zavračal v *Filozofskih raziskavah* in drugje.

Ali Wittgensteinovo zavračanje mentalnih fenomenov kot notranjih stanj v posameznikovi duševni ali fiziološki strukturi pomeni, da se je odpovedal znanstveni psihologiji? Ali moramo psihološke pojave razlagati kot v celoti družbene ali vsaj medosebne, ne pa kot osebne? Res je, da je predstava o Wittgensteinu kot nasprotniku znanstvene psihologije dokaj običajna (gl. npr. Williams, 1999; Rey, 2003), vendar bi na podlagi Wittgensteinovih ostrih opazk o filozofiji enako ali še bolj upravičeno lahko govorili o njegovem nasprotovanju filozofiji itd. A takšni sklepi so po mojem mnenju preuranjeni. Prej bi lahko dejali, da se je Wittgenstein trudil otresti vseh psiholoških, filozofskih in tudi drugih teorij o človeški duševnosti, človekovem govoru, komunikaciji ipd., ker je skušal izpostaviti tisto, kar po njegovem ni nobena teorija več, temveč izraz »trdnih tal«, tj. nedvomnega obvladanja pravil in jezika. Ker so nekatere psihološke in filozofske teorije skušale zajeti tudi ta »tla« in jih razlagale z naravoslovnimi, psihološkimi in filozofskimi domnevami in teorijami, so s tem le dodatno oteževale »pregleden prikaz« (Wittgenstein, 2014: par. 122) in temu se je Wittgenstein silovito upiral. V *Beležkah o filozofiji psihologije* je Wittgenstein celo skiciral neke vrste načrt za analizo psiholoških pojmov, vendar kot vsakdanjih pojmov, ne pa pojmov, ki bi jih za svoj namen ustvarila znanost (1984: II, par. 63, 148). To pomeni, da pri gradnji psiholoških pojmov ne smemo pretrgati njihove zveze z vsakdanjim razumevanjem. To razlikuje psihologijo od nekaterih naravoslovnih ved, ki težijo k čim večji neodvisnosti od vsakdanjega razumevanja svojega predmetnega območja.

Ohranjanje sistematske zveze z vsakdanjim razumevanjem psiholoških pojmov in pojavov ne pomeni eliminacije znanstvene psihologije ali njenega zvajanja na behaviorizem. Kot meni Mary McGinn, Wittgensteinova kritika psiholoških pojmov ne pomeni, da psihološki pojmi terjajo vedenjske kriterije svoje uporabe, temveč »poskus pokazati, da zgolj z introspekcijo ne moremo izpeljati predstave o tem, kaj je dano psihično stanje« (McGinn, 1997: 130). Po McGinnovi argument zoper zasebni jezik ni v tem, da »morajo naši psihološki pojmi posedovati javne kriterije, temveč da le s spominjanjem na gramatiko naših običajnih psiholoških pojmov lahko dojamemo bistvo ali naravo psihološkega stanja dane vrste« (prav tam). Strinjam se s to interpretacijo.

Seveda je takšno pojmovanje psihologije in psiholoških pojmov v precejšnjem nasprotju z intencami dobršnega dela kognitivne znanosti, zlasti v nasprotju z radikalnimi zastopniki računskega pristopa.

Wittgensteinova kritika strojne metafore za razlago mišljenja jasno navaja načelno razliko med operacijami v simbolnem modelu stroja in njegovim možnim realnim delovanjem oz. nedelovanjem. Ne gre le za razliko med strojem, razumljnim na idealni način, tj. v abstrakciji od možnih odpovedi, in realnim delovanjem stroja, temveč za kategorialno razliko med sledenjem pravilom, ki jih vsebuje simbolni model, in časovnim zaporedjem njegovih realnih stanj. Če je torej računalniški algoritem res idealizirana razlaga dejanskega delovanja možganov, kot

menijo zagovorniki stroge računalniške koncepcije duha, potem dejansko operiranje ali dejansko sledenje pravilom ne poteka v možganih, temveč to počnemo *mi* kot ljudje, ki uporabljamo navedeni algoritem kot simbolni prikaz mišljenja. Zato je zelo nejasno, če govorimo o »mehanski realizaciji« algoritma v možganih, kajti takšna »realizacija« ni algoritem. Možgani »delujejo« algoritemsko le na podlagi naše (dejanske) interpretacije možganskega procesa kot računskega procesa.⁴ Zato sem kritičen do poskusov zblževanja Wittgensteinovega pojmovanja mišljenja in »računalniškim« pojmovanjem mišljenja (npr. pri Justinu Leiberju (1991) ali Georgu Rayu (2003)).⁵

Čeprav drži, da je Wittgenstein ostro zavračal dualistično pojmovanje odnosa med duševnostjo in telesom, domnevno predjezikovno, notranjo naravo duševnih stanj, introspekcijo kot psihološko metodo in da najdemo podobna stališča tudi pri zagovornikih računalniškega modela duha, pa to ne zadošča za trditev, da je bil Wittgenstein neke vrste »kognitivni naturalist« (kot meni Leiber) ali »funkcionalni mentalist« (kot meni Rey). Seveda se lahko računalnik »vede v skladu s pravilom«, pa vendar še ne *sledi pravilu*, kajti manjka mu ustrezna kompetenca. Ta se po Wittgensteinu razvije v skupnosti, kjer obstaja *utrjena praksa sledenja pravilu*, ki pozna določene običaje, vežbanje v uporabi pravila, razne tehnike sledenja pravilu ipd. Če bi hoteli npr. res dokazati obstoj sistematične podobnosti med (algoritemskim ali vzporednim) delovanjem računalnika in početjem ljudi, ko sledijo pravilom, bi morali razviti nekakšen nov »Turingov test«, ki bi pokazal, kako lahko takšno početje kar se da realistično predstavimo ali »realiziramo« kot računske procese, ki jih morda lahko simuliramo na računalniku. Podobno velja za »modernejš« poskuse modeliranja mišljenja ob pomoči nealgoritemskih, vzporednih informacijskih procesov v umetnih nevronskih mrežah. Tudi v tem primeru ne moremo trditi, da obstoj določenih funkcionalnih podobnosti med delovanjem možganov in delovanjem umetnih nevronskih mrež zadošča za utemeljitev teze, da je mišljenje (eden od) vzporedni(h) proces(ov) v biološki nevronske mreži, v človekovih možganih. Tudi tu bi morali poprej izvesti serijo testov, ki bi npr. pokazali, da obstaja sistematična podobnost med človeško prakso uvežbavanja pravil v skupnosti, kjer obstaja utrjena praksa sledenja pravilom, in procesom »učenja pravil« določenih nevronskih mrež. Vendar vsaj meni niso znani nobeni sistematični poskusi te vrste. Morda bi kdo dejal, da so dosedanji uspehi umetne inteligence, npr. v robotiki, kjer aparati umetne inteligence na vse več področjih uspešno zamenjujejo ljudi, vsaj posreden dokaz te vrste, namreč dokaz teze, da je človeška inteligenca »konec koncev« le nekoliko bolj kompleksna in bolj zmogljiva »računalniška« inteligenca.

⁴ Problematiko sledenja pravilom in Wittgensteinovo evolucijo pojma operacije obravnavam bolj obsežno v svoji knjigi *Operationen und Regeln bei Wittgenstein* (1997).

⁵ Obsežnejšo kritiko Leiberja in Reya sem podal v Ule (2005), kjer se opiram tudi na delo Diane Proudfoot in Jacka Copelandja »Turing, Wittgenstein and the Science of the Mind« (1994).

Teoretiki kognitivne znanosti in umetne inteligence po mojem mnenju še niso resno vzeli Wittgensteinovih kritik mišljenja kot »notranjega procesa« in začuda pogosto zamenjujejo vprašanja »kako kaj deluje« (ang. *how it works*) z vprašanjem »kaj počne« (ang. *what it does*), podobno zamenjujejo vzročno razlago mišljenja z operacijami v simbolnem modelu oz., splošneje, s sledenjem pravilom ter podobno zamenjujejo ravnanje, ki je v skladu s pravilom z dejanskim sledenjem pravilu. Predvsem pa je po moje usodno zamenjavanje notranjih računskih (ali vzporednih) procesov v možganih, ki morda povzročajo mentalna stanja, z zunanjimi, socialno utemeljenimi vsakodnevnimi kriteriji prepoznavanja, klasifikacije in poznavanja vsebine mentalnih stanj.

Wittgensteinian Criticism of Ideas on Thinking Machines

In this paper I analyze Wittgenstein's criticism of several assumptions that are crucial for a large part of cognitive science. In my opinion, theorists of cognitive science have not yet seriously considered Wittgenstein's criticisms and, quite surprisingly, frequently confuse the question "how does it work?" with "what does it do?" Their most "deleterious" mistake, however, is their confusion of the internal computational (or parallel) processes taking place in the brain (possibly causing mental states) with the external, socially-based, everyday criteria of recognition and classification and knowledge of the content of mental states.

Keywords: computational processes in the brain, thinking machine, Wittgenstein, Turing, rule-following

Literatura

Dietrich, E. (1989). »Semantics and the Computational Paradigm in Cognitive Psychology«. *Synthese*, 79, str. 119–141.

Fields, C. (1996). »Measurement and Computational Description«. V Millican, P. J. R. in Clark, A. (ur.), *Machines and Thought*, vol. 1, Oxford: Clarendon Press, str.165–177.

McGinn, M. (1997). *Wittgenstein and the Philosophical Investigations*. London in New York: Routledge.

Leiber, J. (1991). *An Invitation to Cognitive Science*. Oxford: Blackwell.

Proudfoot, D. in Copeland, B. J. (1994). »Turing, Wittgenstein and the Sciences of the Mind«. *Australasian Journal of Philosophy*, 72(4), str. 497–519.

Rey, G. (2003). »Why Wittgenstein Ought to Have Been a Computationalist?«. *Croatian Journal of Philosophy*, 3(9), str. 231–264.

- Shanker, S. G. (1987). »Wittgenstein versus Turing on the Nature of Church's Thesis«. *Notre Dame Journal of Formal Logic*, 28(4), str. 615–649.
- Turing, A. M. (1937). »On computable numbers, with an application to the Entscheidungsproblem«. *Proceedings of the London Mathematical Society*, 2(42), str. 230–265.
- Ule, A. (1977). *Operationen und Regeln bei Wittgenstein*. Frankfurt/M: Peter Lang.
- Ule, A. (2005). »How can one catch a thought-bird?: Some Wittgensteinian comments to computational modeling of mind«. *Synthesis Philosophica*, 20(2), str. 373–388.
- Williams, M. (1999). *Wittgenstein, Mind and Meaning: Towards a Social Conception on Mind*. London in New York: Routledge.
- Wittgenstein, L. (1964). *Remarks on the Foundations of Mathematics*. Cambridge, Mass.: The MIT Press.
- Wittgenstein, L. (1970). *Zettel*. V Wittgenstein, L., *Werke 5*, Frankfurt/M.: Suhrkamp.
- Wittgenstein, L. (1976). *Logično filozofski traktat*. Ljubljana: Mladinska knjiga.
- Wittgenstein, L. (1984). *Bemerkungen über die Philosophie der Psychologie, Bd. I. V Wittgenstein, L. Werkausgabe, Bd. 7*, Frankfurt/M: Suhrkamp.
- Wittgenstein, L. (2014). *Filozofske raziskave*. Ljubljana: Krtina.