ANαliZA

časopis za kritično misel

Letnik 16, številka 1-2, avgust 2012 · ISSN 1408-2969

FILOZOFSKI PROBLEMI UMETNE INTELIGENCE

Janez Bregant

Turing se je motil

Filozofska fakulteta Univerze v Mariboru

Citiranje: Janez Bregant, »Turing se je motil«, *Analiza: časopis za kritično misel* 16, št. 1-2 (2012): 5–31.

To delo je objavljeno pod licenco Creative Commons Priznanje avtorstva 4.0 Mednarodna (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/deed.sl.

Izdajatelj: Društvo za analitično filozofijo in filozofijo znanosti, Ljubljana.

Elektronska različica je bila leta 2026 rekonstruirana iz skena tiskane izdaje (OCR, ročno pregledano) in preverjena s tiskano izdajo. Številke strani ustrezajo tiskani izdaji; prelomi strani znotraj prispevka se lahko nekoliko razlikujejo.

Janez Bregant
Filozofska fakulteta Univerze v Mariboru

Turing se je motil

V članku bomo predstavili devet ugovorov ideji o mislečih strojih iz 40. let 20. stol, ki so bili plod takratne laične in strokovne javnosti in ki jih je Turing leta 1950 zbral v svojem prelomnem članku »Stroji, ki računajo in inteligenca«. Poleg tega bomo tudi pokazali, kako jih je Turing v istem članku zavrnil, kar je botrovalo temu, da se je o tem, kako bi lahko bilo inteligentno nekaj takega, kot so elektronski možgani, kot so takrat imenovali računalnike, sploh začelo resneje razpravljati. V zvezi z njegovimi zavrnitvami bomo predstavili in razvili nekaj novih idej, ki jih branijo in tako podpirajo njegovo prepričanje, da so misleči stroji načeloma možni.

Ključne besede: umetna inteligenca, kognitivna znanost, Turing, lady Lovelace, solipsizem.

Uvod

Piše se leto 2029 in ljudje smo v hudih težavah: globalni računalnik Skynet, ki nadzoruje vso jedrsko orožje, je razvil lasten razum in se odločil, da se bo uprl svojim človeškim stvariteljem. Načrtuje sprožitev jedrske katastrofe, ki bo skupaj z njegovo vojsko robotov T-800 pobila milijone ljudi iz mesa in krvi, ki so ga tako dolgo zatirali in jih tako enkrat za vselej izbrisala z obličja Zemlje. Toda ljudje pod vodstvom Johna Connorja kljub premoči računalnika v tehnologiji in številu robotov vztrajajo, dokler niso pred tem, da zmagajo. Zato se superduh Skynet zvito odloči, da bo poslal enega izmed robotov, Terminatorja, nazaj v čas, da bi ubil mamo bodočega organizatorja ljudskega upora, Sarah Connor in tako preprečil, da ga kasneje ljudje v opisani vojni ne bi porazili. T-800 je učinkovit stroj za ubijanje brez čustev, narejen iz kovinskega skeleta ter pokrit z živim tkivom, ki je podobno človeški koži, zaradi česar ga je od nas na zunaj skoraj nemogoče razlikovati. Ni pa Terminator edini, ki se napoti nazaj v čas, ljudje pošljejo za njim svojega »vojaka«, to je Kyle Reese, ki naj bi robotovo tarčo zaščitil. Po hudem boju, v katerem Kyle umre, Sarah v bombni eksploziji izmaličen stroj, ki funkcionira zgolj še delno, dokončno uniči v hidravlični stiskalnici, ki iz njega naredi sploščeno gmoto.[1]

To je v grobem zgodba enega izmed najslavnejših filmov o boju med človekom in strojem, *Terminatorja*, ki ga je leta 1984 posnel James Cameron in ki je kasneje dobil še nekaj nadaljevanj, vsa pa so bila narejena na isto temo z manjšimi modifikacijami in izboljšavami robotov. Film je znanstvena fantastika, v katerem stroji, ki jih je naredil človek, pridobijo eno izmed po

[1] Ideja je iz McGinn (1999).

našem mnenju edinstvenih lastnosti človeka, to je duh ali razum oziroma zavest. Kljub temu, da je njihov govor okoren in obrazna mimika omejena, dajejo vtis zavestnih bitij. Očitno je, da se ustvarjalcem filma idejo o inteligentnih robotih ni zdela protislovna, pri čemer pa niso osamljeni, saj vidimo danes robota z duhom v sebi, kamorkoli se obrnemo (McGinn, 1999: 176).

Tako je mislil tudi Alan Turing, ko je leta 1950 v svojem članku »Stroji, ki računajo in inteligenca« zapisal: »Kljub temu pa mislim, da se bosta ob koncu stoletja uporaba besed in splošno izobraženo mnenje tako spremenila, da bo mogoče govoriti o strojih, ki mislijo, ne da bi pričakovali oporekanje« (Turing, 1990: 64). Ko danes več kot šestdeset let kasneje, spraševanje javnost, kako gleda na umetne sisteme (računalnike, programe, robote), ki so kandidati za to, da opravijo tako imenovani Turingov test, ki je vsaj po prepričanju nekaterih raziskovalcev pravo merilo za ugotavljanje tega, ali neka stvar misli ali ne, o čemer smo pisali v eni izmed prejšnjih številk *Analize*,[2] se upravičeno vprašamo, ali se ni Turing v svojem predvidevanju morda motil? Laična in strokovna javnost namreč niti danes, ko živimo v času, v katerem bi se morala njegova prerokba uresničiti in okolju, v katerem so računalniki takšne in drugačne moči naši vsakodnevni spremljevalci, ideji o mislečih strojih, tako kot v 50. letih 20. stoletja, niti približno ni naklonjena. Članek predstavlja devet ugovorov iz tistega časa prepričanju, da so stroji, ki so jih takrat imenovali elektronski možgani, inteligentni.[3] Vse je Turing zbral v svojem zgoraj omenjenem članku. Prav tako opisuje njihove zavrnitve, ki jih je objavil na istem mestu ter v zvezi z njimi prinaša nekaj novosti, ki govorijo v prid njegovi viziji, da stroji vsaj načeloma lahko mislijo. Omeniti pa velja, da četudi sprejmemo vse njegove zavrnitve ugovorov ideji o inteligentnih strojih, to še zdaleč ne pomeni, da takšni umetni sistemi tudi obstajajo; zagovornik tega se mora spopasti še s celim kupom drugih problemov, filozofskih in znanstvenih, ki so nastali kasneje in o katerih bomo v *Analizi* govorili ob naslednji priložnosti. Zdaj pa začnimo z zgodovinsko prvim sistematiziranim pregledom ugovorov temu, da bi misleči stroji načeloma lahko obstajali.[4]

I Teološki ugovor

Že Descartes je leta 1637 v svoji *Razpravi o metodi* pod pretvezo svet je zgodba [*mundus est fabula*], da bi se zavaroval pred srdom cerkve, ki bi si ga utegnil nakopati s svojimi pogledi na nastanek sveta oziroma svojo fiziko, pisal o tem, da je možen podoben svet, kot je naš, v katerem pa narava, kot jo poznamo, ni bila izoblikovana že takrat, ko ga je Bog ustvaril, ampak šele kasneje. Lahko bi se namreč zgodilo, da je Bog najprej ustvaril zgolj kaotično materijo, potem pa skupaj z naravnimi zakoni v procesu součinkovanja vse ostalo, pri čemer je bil on sam primarni, naravni zakoni (ki jih je, mimogrede, naravi tako ali tako podelil Bog) pa sekundarni vzrok.

[2] Glej Bregant (2010).
[3] Pojem »inteligence« v tem članku uporabljamo zgolj v smislu posedovanja misli ali razuma.
[4] Za koristne pripombe in namige se zahvaljujem Bojanu Borstnerju in Danilu Šustru.

Kljub temu odklonu od takratne običajne razlage nastanka vesolja pa vidimo, da Descartes v maniri dobrega katoliškega znanstvenika tudi v svojem novem svetu še vedno predpostavlja, da je Bog avtor Zemlje in ostalega, saj je skladno s sholastično doktrino, ki se ji očitno ni zmogel odreči, Bog hkrati stvarnik in ohranjevalec, ali kot pravi že Tomaž Akvinski v svoji *Summi Theologici*, napisani med leti 1265-1274, »Bog ne ohranja stvari z nekim novim delovanjem, temveč z nadaljevanjem dejanja, ki podeljuje bit« (Akvinski, 1947: L, q. 104, art. 1, ad quartum), Descartes pa leta 1642 v svojih *Quintae Responsiones*; »Bog je vzrok stvari ne le, kolikor nastanejo, temveč tudi, kolikor bivajo« (AT VII: 369, CSM II: 254-255). Descartesovo prepričanje, da je njegova domneva o tem, da je svet nastal iz prvotno neurejene snovi, združljiva s teološkim naukom, se kaže v naslednjem odlomku: »Ne bomo torej delali krivice čudežu stvarjenja, če bomo verjeli naslednje: tudi če bi Bog na začetku dal svetu le obliko kaosa – obenem pa postavil zakone narave in tej naklanjal svoje sočinkovanje, da bi tako delovala, kakor običajno deluje – bi lahko že samo s tem vse zgolj materialne stvari postale takšne, kot jih vidimo zdaj«[5] (Descartes, 2007: 69).

Kakorkoli, obstaja dvoje, česar Bog ni mogel ustvariti iz kaotične materije s pomočjo naravnih zakonov, ampak je to moral proizvesti sam: to sta zaradi svoje kompleksnosti, ki si je Descartes takrat za razliko od splošnejših resnic o vesolju še ni znal zadovoljivo pojasniti, človek in žival: »Toda ker o živalih in ljudeh še nisem imel tolikšnega spoznanja, da bi lahko o njih govoril na isti način kot o preostalem – namreč dokazoval učinke z vzroki in pokazal, iz katerih semen ter na kakšen način jih mora narava proizvesti – sem se zadovoljil s predpostavko, da je Bog izoblikoval človeško [in predpostavljamo tudi, da živalsko] telo ...« (Descartes, 2007: 69). Je pa po Descartesovem mnenju med ljudmi in živalmi kljub temu ena bistvena razlika, in to je, da živali za razliko od ljudi nimajo duše, kar nas vodi do poante teološkega ugovora temu, da lahko stroji mislijo, saj naj bi isto veljalo tudi za »avtomate«, s katerimi je že leta 1637 v *Razpravi o metodi* vizionarsko opisoval današnje robote oziroma računalnike.

Za živali in avtomate namreč trdi, da niso sposobni smiselne govorne komunikacije, kar je po njegovem kriterij ali preizkus, ki ga mora vsak sistem prestati, da bi mu lahko pripisali mišljenje. To je v kartezijanskem dualizmu substanc poglavitna lastnost duše, slednja pa se z njim pogosto kar enači in imenuje *res cogitans*, to je misleča stvar: »Vredno je omeniti, da ni takšnih ljudi, čeprav še tako omejenih in neumnih, vštevši blazneže, ki bi ne imeli zmožnosti, da združujejo različne besede in z njimi sestavljajo govor, s katerim lahko razodevajo svoje misli; da pa, nasprotno, česa podobnega ne zmore nobena druga žival, najsi bo še tako popolna in bogato obdarjena« (Descartes, 2007: 85). Prav tako ni prizanesljiv do strojev: »Lahko si sicer zamislimo, da bi bil kakšen stroj narejen tako, da bi izgovarjal besede, in celo, da bi izrekel kakšno besedo o telesnih dejanjih, ki bi povzročila spremembe v njegovih organih ..., ne moremo

[5] Gregor Kroupa v spremni besedi k *Razpravi o metodi* (2007) v 49. opombi to lepo povzame takole: »Če ohranjanje sveta v bivanju ni nič drugega kot njegovo nenehno in nadaljujoče se stvarjenje, potem je Bog ustvaril Zemljo, Sonce, zvezde im., četudi je na začetku ustvaril le kaos in se je univerzum iz slednjega šele razvil« (Descartes, 2007: 123).

pa si zamisliti, da bi te besede združeval v različne tvorbe in z njimi smiselno odgovarjal na vse, kar bi kdo rekel vpričo njega, kot to zmorejo najbolj omejeni ljudje« (Descartes, 2007: 83). Zanimivo pa je, da po njegovem mnenju avtomata, ki bi imel podobe kakšne živali brez razuma, od slednje ne bi mogli ločiti: »... [Če bi bili kje stroji, ki bi imeli opičje organe in podobo, ali organe in podobo kakšne druge živali brez razuma, ne bi mogli na noben način prepoznati, ali so povsem enake narave kot te živali ali ne« (Descartes, 2007: 83).

Preden predstavimo teološki argument, kot si ga je v 40. letih 20. stol, zamislila takratna javnost in ki v marsičem temelji na opisanih kartezijanskih razmišljanjih, moramo odgovoriti samo še na vprašanje, kako so ljudje po Descartesu prišli do razuma? Ko je ocenjeval telo in še ni vedel, ali je duša razumna, je ugotovil, da ima le-to številne spretnosti, za prepoznavanje katerih ne rabimo mišljenja in katere najdemo tudi pri živalih. Ni pa na ta način mogel odkriti nobenih lastnosti, za katere pravimo, da temeljijo na mišljenju in katerih pri živalih ne moremo najti, dokler ni predpostavil, da je duši razum podelil Bog: »Ko sem preučeval, katere funkcije bi v skladu s tem lahko imelo to telo, sem namreč odkril natanko tiste, ki so lahko v nas, ne da bi mislili nanje, in torej, ne da bi k temu pripomogla naša duša, torej tisti del nas samih, ki je različen od telesa in katerega narava je, kot sem povedal zgoraj, zgolj mišljenje... Nisem pa mogel odkriti nobene od funkcij, ki so odvisne od mišljenja in nam zato edine pripadajo kot ljudem; saj sem jih v njem odkril šele potem, ko sem predpostavil, da je Bog ustvaril razumno dušo in jo s telesom združil na način, ki sem ga opisal« (Descartes, 2007: 70-71). Njegov odgovor na zgornje vprašanje, ki hkrati predstavlja tudi bistvo teološkega argumenta, je tako na dlani: *dušo, ki misli, je ljudem dal Bog*.

V standardni obliki Turing teološki argument, za katerega pa ne smemo misliti, da je že preživet – v evropski družbi, ki ni prežeta s krščanstvom zgolj na deklarativni ravni, bi brez težav našel podpornike – zapiše takole:

1. Mišljenje je funkcija človekove nesmrtne duše.
2. Bog je dal nesmrtno dušo vsakemu moškemu in ženski.
3. Ni pa je dal nobeni drugi živali ali strojem.

∴ 4. Nobena žival ali stroj ne more misliti (Turing, 1990: 65).

Zavrne ga brez slabe vesti, saj na to, da bi v primeru izdelave inteligentnih strojev neupravičeno posegli po nečem, kar je dovoljeno samo Bogu, to je ustvarjanju duš, odgovarja, da to ne drži, saj bi v primeru izdelave mislečih avtomatov mi priskrbeli zgolj domove ali telesa za duše, slednje pa bi še vedno ustvarjal Bog. Svoj odgovor ponazori z naslednjo analogijo: »Pri poskusu konstrukcije takih strojev se ne bi nič bolj nespoštljivo polastili Njegove moči ustvarjanja duš kot pri zaploditvi otrok; v obeh primerih smo prej instrumenti Njegove volje, ki oskrbijo dvorce za duše, ki jih On ustvarja« (Turing, 1990: 65).

Težje se strinjamo z njegovim drugim ugovorom, da iz »zgoraj navedenega argumenta sledi resna omejitev vsemogočnosti Boga« (Turing, 1990: 65), zaradi

česar ni združljiv z definicijo Boga, po kateri je to eden izmed njegovih atributov.

Iz tega, da se Bogu očitno ne zdi primerno, da bi računalnike opremil z mislečo dušo, namreč ne izhaja, da tega ne bi mogel narediti, če bi želel. Tako je njegova ignoranca do strojev glede podeljevanja razuma prej posledica na primer njegove ocene, da si tega ne zaslužijo, če njegovo odločitev o neprimernosti malo ironiziramo, kot pa njegove nezmožnosti to narediti.

Kakorkoli, zaenkrat ne vidimo razloga za to, da Bog avtomatom ne bi dal mišljenja, če bi se mu tako zahotelo; s tem bi podkrepil svojo vsemogočnost, pa tudi licence za ustvarjanje duš ne bi izgubil. Njegova zadržanost glede tega mora tako očitno izhajati iz nečesa drugega. Mogoče je drža Boga bolj razumljiva, če pogledamo na teološki argument v luči ostalih argumentov proti inteligentnim strojem, izmed katerih že naslednji, ki opozarja na grozljive posledice, ki bi jih takšno početje lahko imelo, kar kliče po božjem posegu v svet, čemur pa se Bog (običajno) skuša izogniti.[6]

II Ugovor glav v pesku [*The Heads in the Sand Objection*]

Ne vem sicer zakaj, ampak večinoma se stroji, pri čemer imam v mislih robote, ne pa gospodinjskih aparatov, kljub temu, da se nam tudi zanje včasih zdi, da so na nek čuden način od danes na jutri dobili lastno voljo, takrat namreč, ko nam povzročajo takšne in drugačne težave, v našem svetu prikazujejo kot nasilni, zlobni in napadalni, skratka, nevarni. Kot da se je po filmu *Planet opic* iz leta 1968 v nas zažrl strah pred tem, da bi nas enkrat v prihodnosti, ko bomo svoj planet že skorajda uničili, nadvladala neka druga vrsta. V omenjenem filmu so to opice, ki so se razvile v bitja z inteligenco, podobno človeški, in govorom ter prevzele našo vlogo, medtem ko smo se ljudje izrodili v mutce, ki nosijo okrog pasu prazgodovinske živalske kože. V nas je bojazen, da ne bi v stvarnosti vlogo opic prevzeli stroji. Povedano drugače, posledice zavestnih računalnikov bi bile tako uničujoče za človeško raso, da ne smemo verjeti, da kaj takega sploh lahko obstaja, to je, delajmo se, da je to nemogoče, zatisnimo oči, zajemimo sapo ter vtaknimo glavo v pesek, kot to počnejo noji, ko jim grozi kakšna nevarnost, in kot oni upajmo, da se ji bomo s takšnim ignoriranjem izognili. Nekako tako lahko povzamemo ugovor glav v pesku, ki se odkrito v standardnem zapisu glasi:

1. Posledice strojnega mišljenje bi bile preveč strašne.

∴ 2. Upajmo in verjemimo, da stroji tega ne zmorejo (Turing, 1990: 66).

Argument ni samo enostaven, milo rečeno tudi prepričljiv ni. Ni težko ugotoviti, od kod črpa svojo verodostojnost za tiste, ki ga sprejemajo, gotovo pa ne iz logike, saj iz tega, da bi imel nek dogodek grozljive posledice, ne sledi, da se ne

[6] Tukaj nismo razpravljali o tem, ali obstaja kakšen dober dokaz za obstoj Boga, kar bi po mnenju nekaterih morali narediti najprej – zdi se, da celotna zgodovina dokazovanja Boga ni do danes proizvedla ničesar, kar bi kazalo v tej smeri – v argumentu smo hoteli pokazati, da četudi bi nekaj takega resnično obstajalo, še vedno ne bi bilo razloga, da stroji ne bi mogli misliti.

bo nikoli zgodil (Američani so vedeli, kaj bo povzročila jedrska eksplozija, pa se kljub temu niso odrekli bombardiranju Hirošime in Nagasakija, da bi Japonce prisilili v vdajo). Ker je, z drugimi besedami, med našimi željami in stvarnostjo pogosto velik prepad, mora razlog za to, da ljudje ugovor glav v pesku sprejemajo – v 40. letih 20. stol, mogoče bolj kot zdaj – tičati drugje in to je v krščanskem prepričanju, da je človek superiorno bitje, edino z razumno dušo, ki mu omogoča mišljenje, česar pa za nobeno ceno ne sme izgubiti, saj bi s tem izgubil tudi svoj smisel.

Turingu se ne zdi vredno, da bi ga zavrnil: »Menim, da ta argument ni dovolj tehten, da bi zahteval ovržbo« in ironično nadaljuje: »Bolj primerna bi bila tolažba: morda bi jo veljalo iskati v preseljevanju duš« (Turing, 1990: 66).

Drugače pa danes ni težko najti empiričnega dokaza, ki premiso zgornjega argumenta postavlja na laž, če smo le pripravljeni kot kriterij, kdaj lahko govorimo o mislečem (umetnem ali naravnem) bitju, sprejeti šah, kot se pogosto (ali pa se je) v praksi tudi dogajalo. Prvič sta človeški svetovni šahovski prvak, ki je bil takrat Kasparov in računalniški svetovni šahovski prvak, ki je bil takrat *Globoka misel [Deep Thought]*, pomerila leta 1989. Kasparov je zmagal gladko, bilo je 2:0. Mediji so potem do onemoglosti ponavljali, da »gre za zmago duha nad snovjo«, naslovi v časopisih pa so se glasili »Kasparov sesul računalnik v imenu celotnega človeštva«. Ljudem je odleglo, da je bil stroj poražen, saj velja šah za simbol (pre)moči človeškega duha. V primeru drugačnega rezultata, bi se lahko namreč našli ljudje, ki bi navdušeni nad prihodom novega superuma, z računalniki želeli nadomestiti učitelje, sodnike, zdravnike itd., kar bi v končni fazi vodilo do popolne kompjutizacije sveta. No, kot bomo videli, se ni zgodilo nič takega.

Programerjem na Carnegie Mellon poraz leta 1989 namreč ni vzel poguma in so šahovski program posodabljali tako dolgo, dokler ni bil končno sposoben premagati človeka. To se je zgodilo leta 1997, ko je moral Kasparov s 3.5 : 2.5 priznati premoč *Globoki modrini [Deep Blue]* (posamezno partijo je program proti njemu sicer dobil že leta 1996, dvoboja pa takrat še ne), nasledniku *Globoke misli. Globoka modrina* tako velja za prvi računalnik, ki je kdajkoli premagal človeštvo. In kaj se je po njegovi zmagi zgodilo? Mogoče apokalipsa človeštva? Daleč od tega, *Globoko modrino* so razstavili, ljudje so se sprijaznili s tem, da so dnevi človeške premoči v šahu končani, po zadnjem šahovskem dvoboju med človekom in strojem, leta 2006, ko *Globoki Fritz. [Deep Fritz]* ni imel težav s Kramnikom, zmagal je s 4 : 2, pa potekajo dvoboji za svetovnega šahovskega prvaka ločeno. Tako se zdi, da je premisa argumenta glav v pesku neresnična, saj nam praktični primer računalnika, ki nam je vzel intelektualno premoč, kaže, da posledice strojnega mišljenja za nas niti slučajno ne bi bile ali niso pogubne in da bi zato morali na robote gledati z drugačnimi očmi, kot to dela ugovor glav v pesku.

III Matematični ugovor

Že Aristotel je v svoji *Prvi analitiki* (del njegovih šestih spisov poglavitno o logiki, kasneje zbranih in imenovanih *Organon)* zasnoval postopek, s katerim je preverjal veljavnost silogizmov: silogizem je veljaven, če in samo če ni možno, da bi bili obe njegovi premisi resnični, sklep pa neresničen. To metodo izločanja neveljavnih silogističnih figur danes imenujemo algoritem odločljivosti za formalne sisteme, ki ga razumemo kot možnost, da vsako formulo izpeljemo iz aksiomov, to je, pokažemo, da je znotraj določenega sistema dokazljiva ali ne (v tem primeru je znotraj danega sistema ovrgljiva, ker lahko na primer izpeljemo njeno negacijo) (Uršič, Markič, 2009: 105). Skratka, vsaka poljubna formula *odločljivega* aksiomatskega sistema je bodisi dokazljiva bodisi ovrgljiva, kar ga naredi za doslednega (konsistentnega), to je, da ob danih aksiomih in deduktivnih pravilih ni mogoče izpeljati dveh protislovnih trditev. Pri tem *aksiomatske sisteme* obravnavamo kot formalne strukture, s katerimi izpeljujemo (dokazujemo) izreke ali zakone določene znanosti iz njenih osnovnih intuitivno razvidnih trditev, to je *aksiomov*, definicij in deduktivnih pravil[7]? (Uršič, Markič, 2009: 227).

Problem je, da je Gödel leta 1931 v svojem članku »Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme« dokazal, da je vsak formalni sistem, ki je dovolj bogat, da lahko z njim izrazimo teorijo števil, načeloma *neodločljiv* (Uršič, Markič, 2009: 286), kar je znano pod imenom Gödlov paradoks ali Gödlov teorem[8] in pomeni, da znotraj aksiomatskega sistema obstaja formula, ki ni niti dokazljiva niti ovrgljiva, zaradi česar je le-ta nedosleden. Ponazorimo to z enostavnejšo različico takšne formule, stavkom *F*, ki sam o sebi pravi, da ni dokazljiv, to je, *F* pravi, da F ni dokazljiv, kar imenujemo *samonanašajoči se* stavek. Zanj, to je za *F*, obstajata v logiki dve možnosti, da je (i) resničen ali (ii) neresničen, iz česar sledi naslednje: če je *F* resničen (res je, da *F* ni dokazljiv), potem ni dokazljiv, torej iz (i) izhaja prav tisto, kar (sam) *F* trdi, če pa F ni resničen (ni res, da F ni dokazljiv), potem je dokazljiv, torej iz (ii) izhaja negacija tistega, kar (sam) F trdi (Uršič, Markič, 2009: 287). Tako zaidemo v protislovje, saj dobimo po (i) F ni dokazljiv, po (ii) pa F je dokazljiv, z drugimi besedami, stavek F je hkrati dokazljiv (je resničen) in hkrati ovrgljiv (neresničen), kar pomeni, da je aksiomatski sistem, v katerem nastopa, nedosleden (nekonsistenten) s tem pa tudi *neodločljiv*, to je, ni tako, da je bodisi dokazljiv bodisi ovrgljiv, ampak je oboje.[9]

Ko na problem neodločljivosti pogledamo v luči računalnikov, vidimo, da velja tudi zanje. Ker so abstraktni računalniki, to je Turingovi stroji,[10] aksiomatski sistemi, dejanski računalniki pa so zgolj njihove izpopolnjene različice, se mu

[7] Zgodovinsko prvi aksiomatski sistem je bil verjetno Aristotelova silogistika, pa čeprav še ni bil povsem formaliziran, znani pa so še na primer Evklidova geometrija, Peanova aritmetika ali Russellova logika.

[8] Gre v bistvu za prvi Gödlov teorem nepopolnosti, ki pravi, da v kateremkoli formalnem sistemu, ki dokazuje osnovne aritmetične resnice, obstaja aritmetični stavek, ki je resničen, ampak nedokazljiv.

[9] Za natančnejšo analizo Gödlovega paradoksa glej na primer Uršič, Markič, 2009: 288–289.

tudi slednji ne morejo izogniti. To ima, pravi matematični ugovor, pogubne posledice za idejo, da obstajajo inteligentni stroji, kajti če zavoljo argumenta predpostavimo, da je pravo merilo za preverjanje tega Turingov test, ki smo ga podrobneje predstavili v Bregant (2010), stroj na določena vprašanja ne bo odgovoril, ker se ne bo mogel »odločiti«, kateri odgovor je pravilen, ali pa bo odgovoril napačno. Gre za tako imenovana samonanašajoča se vprašanja, to je vprašanja, katerih vsebina se nanaša na sam stroj, podobno, kot se je vsebina stavka *F* nanašala na sam *F v* primeru, s katerim smo zgoraj opisovali Gödlov paradoks.

Poanta matematičnega ugovora je potem v tem, da bomo v Turingovem testu lahko ugotovili, kje je stroj, saj se njegova inteligenca povezuje z njegovo nezmotljivostjo v smislu odgovarjanja na vprašanja, ki jih lahko izrazimo kot teoreme aritmetike prvega reda. Zaradi tega bomo računalnik lahko identificirali, skriva se tam, od koder ne bo odgovora na takšno vprašanje ali pa bo le-ta napačen, saj se predpostavlja, da človek s pravilnim odgovarjanjem na samonanašajoča se vprašanja nima težav: »Pokazano je npr. bilo, da obstajajo logični sistemi, v katerih noben stroj ne bo razlikoval med dokazljivo in nedokazljivo formulo, tj. da ne obstaja noben preizkus, ki bi ga stroj lahko uporabil za to, da bi zanesljivo uvrstil stavke v ta dva razreda. Tako stroj, če je narejen za ta namen, v nekaterih primerih ne bo dal odgovora« (Turing, 2004: 393). V standardni obliki lahko matematični ugovor zapišemo takole (označimo ga s 7):

1. Turingovi stroji so aksiomatski sistemi.
2. Naši računalniki so njihove (približne) realizacije.
3. Obstajajo resnični stavki, ki ne morejo biti izpeljani iz aksiomatskih sistemov.

(∴) 4. Obstajajo resnični stavki, ki jih računalniki ne morejo izpeljati.

5. Človek lahko pride do takšnih stavkov.

(∴) 6. Našega mišljenja ne moremo ustrezno predstaviti s pomočjo računalnika oziroma aksiomatskega sistema.

∴ 7. Človek ni računalnik.

∴ 8. Noben računalnik ne more biti tako inteligenten kot človek.[11]

Ali je matematični ugovor prepričljiv? Če Turingove besede, da matematični rezultat dokazuje neko pomanjkljivost strojev, ki ji ljudje niso podvrženi, prevedemo v glavni sklep 8, da ne more biti noben računalnik tako inteligenten kot človek se zdi, da matematični ugovor ne zavrača ideje o mislečih strojih na sploh, ampak obstoj inteligentnih strojev, sicer v nekem šibkejšem smislu dopušča, to je, recimo, da mislijo, ampak niso tako pametni kot ljudje, ker pač ne morejo rešiti vseh nalog, ki jih lahko mi. (Vprašanje je, če slednje drži, na kar opozarja spodnja zavrnitev matematičnega ugovora).

[10] Za njihov enostavnejši opis glej na primer Bregant, 2010, za podrobnejšega pa na primer Copeland, 1993.

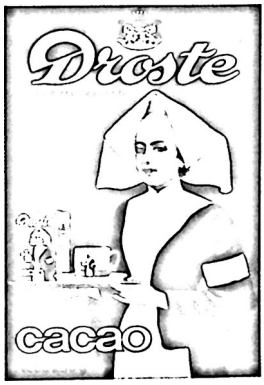
[11] Glej tudi sicer neposrečeno verzijo istega argumenta v Penrose (1999).

Dalje, *T* bi bil dober argument, če bi bila tudi premisa 5 resnična (o resničnosti ostalih premis ob sprejetju Gödlovega teorema ne gre dvomiti), to je, da ljudje (matematiki) znotraj danega aksiomatskega sistem ločijo dokazljive od nedokazljivih stavkov, kar pa je sporno, saj razen ihtavega prikimavanja in energičnega zatrjevanja, da je tako, do sedaj še nismo bili deležni dokaza za njeno resničnost, kar umetne inteligence ali mislečih strojev ne izključuje in prej dokazuje zmoto, do katere vodi Gödlov teorem. Turing na to opozarja z besedami: »... [Č]eprav je ugotovljeno, da obstajajo omejitve moči katerega koli posameznega stroja, pa je zgolj zatrjeno, brez kakšnega koli dokaza, da te omejitve ne veljajo za človeški um« (Turing, 1990: 67).

V članku »Stroji, ki računajo in inteligenca« Turing sicer predstavi preoblikovan matematični ugovor: »Namesto, da bi začel z domnevo, da je dani stroj nezmotljiv in nadaljeval do sklepa, da je stroj manj inteligenten kot oseba, začne z drugačno trditvijo. Ne glede na katerokoli lastnost stroja, ki opravlja [Turingov] test, vključno z nezmotljivostjo, obstaja vsaj eno vprašanje, ki ga lahko oseba postavi stroju, za katerega že vnaprej ve, da nanj stroj ne bo uspel pravilno odgovoriti« (Abramson, 2008: 151). Katera pa so ta vprašanja? »Vprašanja, za katera vemo, da bo ob njih strojem spodletelo, so takšne vrste: 'Dan je neki stroj, ki je opisan takole... Pike je treba zamenjati z opisom nekega stroja v standardni obliki... Kadar je opisani stroj v določenem, razmeroma enostavnem odnosu s strojem, ki je spraševan, se da pokazati, da je odgovor bodisi napačen ali pa sploh ne bo dan. To je matematični rezultat: argumentira se, da dokazuje neko nezmožnost strojev, ki ji človeški um ni podvržen'« (Turing, 1990: 66-67).

Turing ga zavrne s tem, da četudi bi v ustrezno izvedenem Turingovem testu stroj res pravilno identificirali, bi to pomenilo zgolj zmagoslavje nad enim strojem, to je, da dani primerek računalnika ni inteligenten, ne pa, da ni misleč noben stroj, kar izrazi z naslednjimi besedami: »Poleg tega lahko svojo superiornost ob takšni priložnosti občutimo zgolj do tistega enega samega stroja, nad katerim smo zabeležili svoj malenkostni triumf. Sploh pa ne bi prišlo v poštev, da bi lahko triumfirali hkrati nad *vsemi* stroji« (Turing, 1990: 67). Pri tem se opira na to, da je bistvena funkcija kateregakoli končnega niza dogodkov znotraj formalnega sistema njena *rekurzivnost*. To je postopek ponavljanja neke operacije potem, ko smo jo najprej izvedli na začetnem primerku, na rezultatu, ki smo ga s tem dobili, in potem na rezultatu rezultata, ki smo ga s tem dobili, in potem na rezultatu rezultatovega rezultata, ki smo ga s tem dobili itd. Vsebuje osnovni primer in rekurzivno pravilo, ki ga nanj apliciramo. Po tem principu smo dobili na primer definicijo naravnih števil v matematiki, kjer je osnovni primer 7 *je naravno število*, rekurzivno pravilo pa *vsako naravno število ima naslednika, ki je prav tako naravno število*. Če sedaj na začetnem primerku 1 je naravno število izvedemo zahtevano operacijo, dobimo 2 je naravno število (rezultat), če potem na tem rezultatu izvedemo zahtevano operacijo, dobimo 3 je naravno število (rezultat rezultata), če potem na tem rezultatu rezultata izvedemo zahtevano operacijo, dobimo 4 je naravno število (rezultat rezultatovega rezultata) itd. Običajno se rekurzivnost slikovno ponazarja s ti.

droste učinkom, kjer ženska na sliki v rokah drži predmet, ki vsebuje manjšo sliko iste ženske, ki v rokah drži isti predmet, ki spet vsebuje še manjšo sliko iste ženske, ki v rokah drži isti predmet itd.



Ker je vsak končni sistem rekurziven v opisanem smislu, zagovornikom modificiranega matematičnega ugovora pri zavrnitvi ideje o mislečih strojih ne pomaga, da lahko pri opravljanju Turingovega testa pokažejo, da lahko s pravim vprašanjem ugotovimo, kje se skriva računalnik. To namreč ne izključi možnosti, da obstaja drug računalnik, ki je nanj sposoben pravilno odgovoriti in nas pretentati. Seveda pa lahko potem obstaja spraševalec, ki si izmisli vprašanje, na katerega spet ta stroj ne bo mogel odgovoriti, zaradi česar bo lahko ugotovil, kje je računalnik in ga tako izločil iz tekme za pridobitev naziva misleči. To pa spet ne pomeni, da ne obstaja drug računalnik, ki bi na njegovo vprašanje znal pravilno odgovoriti in ga prevarati itd. »Skratka, za vsak določen stroj bi lahko našli od njega pametnejše ljudi, vendar pa bi se potem morda našel še pametnejši stroj in tako dalje« (Turing, 1990: 67).

Tako se zdi, da »matematični ugovor ne izloči tega, da lahko stroji mislijo, ampak zgolj poudarja intelektualno zmago nad posameznim strojem« (Abramson, 2008: 152).

IV Argument iz zavesti

Že sofist Gorgia (483–375 pr. n. št.) je učil, da četudi bi kaj obstajalo, o tem ne bi mogli ničesar vedeti in da četudi bi o tem lahko kaj vedeli, tega znanja ne bi mogli deliti z drugimi,[12] s čimer je zgodovinsko gledano postavil temelje za filozofsko stališče, ki ga danes imenujemo *solipsizem*.[13] Počiva na treh predpostavkah o našem duhu in znanju o njem, vse pa izvirajo iz časa, ko je Descartes kot metodo kritičnega preučevanja narave in človeka postavil egoistično iskanje resnice. Leta 1641 v *Meditacijah* piše, da je jaz misleča stvar brez delov, ki za svoj obstoj ne potrebuje ničesar drugega, saj si lahko zamisli

[12] Glej Sextus Empiricus, *Outlines of Pyrrhonism* (2000).
[13] Obstaja več vrst solipsizma, ki pa so med seboj povezani: metafizični, epistemološki in metodološki.

sebe, kot da nima telesa, pa kljub temu še vedno bo nekaj, to je res cogitans: »Ali morem trditi, da imam vsaj kaj prav neznatnega od vsega tistega, o čemer sem pravkar dejal, da spada k naravi telesa? Pozorno premišljam in premišljam – ničesar se ne domislim ... Kaj pa tisto, kar sem pripisal duši: dajem ali hodim? Če že nimam telesa, je tudi to zgolj umišljeno. Pa čutenje? Tudi čutenja ni brez telesa ...Pa mišljenje? Tu najdem trdno točko: samo mišljenja ni mogoče odtrgati od mene« (Descartes, 1988: 57-58). Ta njegova ideja o tem, (i) da je tisto, kar vem zanesljivo, zgolj tisto, kar je v mojem duhu – moje misli, moji občutki, moje želje itd. – in da te vsebine mojega duha zaradi zasebnega dostopa ne more poznati nihče drug, predstavlja prvo trditev solipsizma.

Descartes se sicer temu, da ne bi obstajal samo njegov duh, izogne s sklicevanjem na dobrohotnega Boga, s čimer svoje ideje, ki jih ima o zunanjem svetu, to je razsežni stvari, materializira. Zanj je Bog tako v nekem smislu most med osamljeno zavestjo, odkrito s pomočjo metodičnega dvoma, in svetom javnih predmetov ter drugih človeških bitij (Thornton, 2004).

Kartezijanski dualizem pa razen tega tudi izrecno trdi, da sta misleča in razsežna substanca ločeni, na kar nas opozarja naslednji citat: »In čeprav morda (ali bolje: zagotovo ...) imam telo, ki je zelo tesno združeno z mano, je vendar – ker imam na eni strani jasno in razločno idejo samega sebe, kolikor sem samo misleča, nerazsežna stvar, in na drugi strani jasno idejo telesa, kolikor je telo zgolj razsežna, nemisleča stvar – je torej vendar gotovo, da sem v resnici različen od svojega telesa in da brez njega ne morem bivati« (*Descartes*, 1988: 107). (Slovenski prevod *Meditacij* iz leta 1973, katerega avtor je Primož Simoniti in ki je bil ponatisnjen leta 1988 in leta 2004, je v zadnjem delu navedenega odlomka napačen, saj izvirnik ne pravi »in da brez njega ne morem bivati, ampak *»in da bi brez njega lahko bival«*: »Et quamvis fortasse (vel potiùs, ut postmodum dicam, pro certo) habeam corpus, quod mihi valde arcte conjunctum est, quia tamen ex unâ parte claram & distinctam habeo ideam meî ipsius, quatenus sum tantùm res cogitans, non extensa, & ex aliâ parte distinctam ideam corporis, quatenus est tantùm res extensa, non cogitans, certum est me a corpore meo revera esse distinctum, *& absque illo posse existere*«)[14] (AT II: 78, poudarek je moj).

Opozoriti velja, da Descartes kljub temu predpostavlja interakcijo med duhom in telesom, zaradi česar lahko rečemo, da gre pri njegovem dualizmu substanc hkrati tudi za *kartezijanski interakcionizem*. Zagovarja namreč trditev, da sta duh in telo združena, prvo ni zgolj mornar in drugo ladja, saj drugače ne bi čutili bolečine, ampak bi jo dojeli z razumom: »Po teh občutkih bolečine, lakote, žeje itd. me narava tudi uči, da pri svojem telesu nisem navzoč kakor mornar na ladji,

[14] To je lepo razvidno tudi iz prve francoske različice *Meditacij* iz leta 1647 v prevodu grofa Luynesa: »Et quoique peut-être (ou plutôt certainement, comme je le dirai tantôt) j'aie un corps auquel je suis très étroitement conjoint; néanmoins, parce que d'un côté j'ai une claire et distincte idée de moi-même, en tant que je suis seulement une chose qui pense et non étendue, et que d'un autre j'ai une idée distincte du corps, en tant qu'il est seulement une chose étendue et qui ne pense point, il est certain que ce moi, c'est-à-dire mon âme, par laquelle je suis ce que je suis, est entièrement et véritablement distincte de mon corps, *et qu'elle peut être ou exister sans lui*« (AT II: 78, poudarek je moj).

temveč da sem z njim kar se da tesno združen, tako rekoč pomešan z njim, tako da sestavljam z njim neko enotnost« (Descartes, 1988: 109). Še bolj očitno govori o povezavi med sicer enotnim duhom in deljivim telesom (kar je še en argument za njuno drugačnost) v naslednjem odlomku: »Če se na primer živci v stopalu zganejo silovito in na nenavaden način, da njihovo gibanje, ki sega prek hrbtenjače v najbolj notranje dele možgan, duhu znamenje, po katerem ta kaj čuti, recimo bolečino, ko da je bolečina v stopalu ...« (Descartes, 1988: 116). Kako je to možno, je sporno že od takrat, ko je Descartesa po objavi *Meditacij* na to opozoril Gassendi: »Pojasniti moraš, kako je ta interakcija, če si breztelesen, se ne raztezaš v prostoru in si nedeljiv, možna ... Kako lahko, če nimaš delov, vplivaš na dele? ... In če si nekaj ločenega, kako lahko skupaj s snovjo tvoriš celoto?« (CSM II: 238-239, AT VII 343-344).

Kakorkoli, (ii) da ni med mentalnim in fizičnim nobene nujne zveze, na primer, da ni ničesar takšnega med zavestnim izkustvom in telesnim obnašanjem, predstavlja drugo solipsistično trditev.

Descartesov duh pa ima tudi lastnost, ki je na prvi pogled tam mogoče ne bi pričakovali: gre za zaznavanje, kar dokazuje v svojem znanem primeru z voskom. Kako vemo, da gre potem, ko vosek, ki ga približamo ognju, izgubi prvotno obliko, barvo, velikost in trdoto, še vedno za isti vosek? Zato ker »zaznavanje voska ni gledanje, tipanje, predstavljanje voska in to tudi nikdar ni bilo, čeprav se je prej tako zdelo, temveč edinole in zgolj motrenje duha ...« (Descartes, 1988: 62). Iz predpostavke, da imamo do našega duha zaseben dostop (kar je prva solipsistična trditev) in Descartesovega prepričanja, da je zaznavanje aktivnost duha,[15] sledi tretja solipsistična trditev, (iii) da imamo tudi do izkustva nujno zaseben dostop.

Dva človeka tako ne moreta nikoli imeti istega izkustva, saj to zaradi skrivnostnosti in kratkotrajnosti svoje narave nikoli ne more biti predmet delitve. Vprašanje je, kako lahko potem sploh spoznamo izkustva drugih ljudi oziroma, kako lahko vemo, da drugi ljudje izkustva sploh imajo? (Thornton, 2004).

Na nesprejemljivost zadnje solipsistične ideje opozarja Wittgenstein, ko pravi, da se pojem »zasebno« uporablja v naslednjem pomenu: (p) samo jaz lahko vem, da resnično čutim bolečino. Pravi, da je (p) v bistvu konjunkcija dveh trditev: (pp jaz čutim bolečino in (P2) drugi ljudje ne vedo, da čutim bolečino. Dokazuje, da je (pj) nesmiseln, (po) pa neresničen (Wittgenstein, 1977: I, par. 246). Prvi konjunkt (p 1) je nesmiseln zato, ker predstavlja običajno izražanje bolečine, ki pa je zgolj del naučenega ali pridobljenega obnašanja, ko nas nekaj boli, to je govorni nadomestek za sicer naravno izražanje bolečine z na primer jokanjem, zaradi česar nima nobene spoznavne vrednosti oziroma je spoznavno prazen. To zapiše takole: »Torej praviš, da beseda ‘bolečina v resnici pomeni jokati? Nasprotno: jezikovno izražanje bolečine nadomesti jokanje in ga ne

[15] To ne pomeni, da duh zaznava tako, kot to počnemo v vsakdanjem življenju, to je s čutili. Gre namreč za vprašanje zaznavanja prave narave snovi, to je tistega, kar se opredeli kot razsežnost. Ker pa so duhu dostopne zgolj ideje, ki so lahko vrojene, pridobljene ali narejene, je zaznavanje v bistvu delo duha v smislu njihovega prepoznavanja.

opisuje« (Wittgenstein, 1977: I, par. 244). Drugi konjunkt (p>) je neresničen zato, ker nas izkušnje učijo ravno nasprotno, včasih je celo nemogoče, da druge osebe ne vedo, da čutimo bolečino, pa naj si tega še tako želimo.

Kakorkoli, če sedaj končamo ta izlet v solipsizem in vse tri njegove teze strnemo, dobimo, da je solipsizem metafizična teorija, ki pravi, da ne obstaja nič razen mene in mojega izkustva. In ravno na to namiguje argument iz zavesti, ki ga Turing ponazori z odlomkom iz Jeffersonovega[16] Listerskega nagovora iz leta 1949: »Dokler stroj ne bo mogel napisati soneta ali komponirati koncerta zato, ker bi mislil in občutil emocije, ne pa skozi naključno podajanje simbolov, ne bomo mogli pristati na to, da je stroj enak možganom – to se pravi, ne le napisati sonet, temveč vedeti, da ga je napisal. Noben mehanizem ne more občutiti (in ne le umetno signalizirati, kar je lahko urediti) ugodja ob svojem uspehu, gorja, ko se mu stopijo ventili, topline ob laskanju, bridkosti ob svojih napakah, ne more biti očaran s seksom ali biti jezen ali potrt, ko ne more dobiti, kar si želi« (Turing, 1990: 67).

V tem odlomku se v resnici skrivata vsaj dva argumenta iz zavesti. Prvi v svoji najbolj skrajni obliki pravi, da »je edini način, na katerega bi lahko kdo bil gotov, da stroj misli, ta, da *je* stroj in da čuti sebe, kako misli. Te občutke bi tedaj lahko opisal ostalemu svetu, toda nihče se seveda ne bi mogel za to upravičeno zmeniti. Podobno je po tem stališču mogoče vedeti, da neki *človek* misli, le tako, da sem jaz ta določen človek« (Turing, 1990: 67). V standardnem zapisu ga lahko ponazorimo takole:

1. L Vsebine duha ne more poznati nihče drug razen lastnika.
2. Mišljenje je del vsebine duha.

(∴) 3. Mišljenje lahko pozna zgolj lastnik.

4. Če hočemo vedeti, da stroj misli, moramo biti ta stroj.
5. Ta stroj pa ne moremo biti.

∴ 6. Ne moremo vedeti, ali stroj misli.

Sklep tega argumenta iz zavesti je blažja oblika solipsizma, gre za tako imenovani epistemološki solipsizem, ki predpostavlja, da ne moremo imeti znanja o ničemer, kar je zunaj našega duha. Seveda pa takšno stališče ni nedolžno, prav nasprotno, če ga sprejmemo, potem nimamo več razloga, da bi poleg sebe tudi na druge gledali kot na misleča bitja. Vprašanje je torej, zakaj se kljub temu obnašamo tako, kot da drugi ljudje imajo duha, če pa tega nikoli ne moremo vedeti oziroma zakaj govorimo, da ljudje čutijo, si želijo, mislijo? Solipsistični odgovor je, da drugim osebam pripisujemo mentalno življenje na osnovi sklepanja po sebi, to je druge osebe imajo lastna izkustva po analogiji z nami. To pa je seveda problematično, saj ni med duhom in telesom v solipsizmu

[16] Geoffrey Jefferson (1886–1961) je bil britanski nevrolog in nevrokirurg, ki je leta 1948 dobil listersko odlikovanje. Tega za izjemne dosežke na področju kirurgije podeljuje Kraljevi kirurški kolidž Anglije, imenovano pa je po Josephu Listerju, ki je postavil temelje sodobne sterilne kirurgije. Geoffrey Jefferson je imel ob tej priliki na Kraljevem kirurškem kolidžu nagovor z naslovom *The Mind of Mechanical Man*, v katerem je razpravljal o Marku 1, enem izmed zgodnjih elektronskih računalnikov.

nobene nujne zveze – to, da je moj duh zdaj lociran v mojem telesu, je čisto naključje – zaradi česar je pripisovanje mentalnih stanj drugim ljudem na osnovi tega, da se obnašajo tako kot mi, mi pa duha, kot vemo, imamo, torej ga imajo tudi oni,[17] nemogoče.

Wittgenstein pravi, da za ljudi uporabljamo pojme kot zavest, občutki, hotenje itd. kljub temu, da ne moremo vedeti, ali imajo duha, zato, ker so ljudje paradigmatski primer tistega, kar ima mentalno življenje: »... [S]amo za živo človeško bitje in kar mu je podobno (se obnaša) kot živo človeško bitje lahko rečemo: ima občutke, vidi, je slepo, sliši, je gluho, je zavestno ali nezavestno« (Wittgenstein, 1977: I., par. 281). In zato, ker ljudje funkcionirajo kot naš prototip tega, kar je zavestno, se nam zdi druge osebe označevati kot avtomate brez izkustva, k čemur epistemološki solipsizem stremi, neverjetno. Turing to enostavno imenuje sprejetje »vljudne konvencije, da vsakdo misli« (Turing, 1990: 68), bil pa bi zadovoljen že s tem, če bi se strinjali, da je to, da mislijo stroji, enako verjetno kot to, da mislijo ljudje. Wittgensteinova zavrnitev solipsističnega sklepanja po sebi kot razloga za pripisovanje duha drugim osebam s svojim dopuščanjem podobnosti v obnašanju kot kriterijem za dušo to nedvomno omogoča.

Drugi argument iz zavesti v istem citatu pa pravi, da je naš pogled na duha, ko ga pripisujemo strojem, preozek, saj ne more obstajati inteligenca, ki je ločena od občutkov, želja in čustev, glede na to, da imajo ti osrednjo vlogo pri nastanku obnašanja. Turing ga zavrne s tako imenovanim »živim glasom« [viva voce], saj meni, da Jefferson ne bi nasprotoval temu, da stroj misli, če bi ta lahko na naša vprašanja svoje pravilne odgovore posredoval s svojim glasom, ki bi nam tako kot pri ljudeh rabil kot dokaz, da računalnik resnično občuti žalost, jezo ali veselje. Turingov test sicer nikjer ne izključuje možnosti ali celo prepoveduje, da bi bil računalnik opremljen tako, da bi lahko govoril, je pa bilo to v njegovem času še nemogoče narediti. Turing tudi nikjer ne pravi, da bi bil edini način narediti stroj, ki bi z glasom, iz katerega bi lahko zanesljivo razbrali, da čuti, pravilno odgovarjal na naša vprašanja in zaradi tega mislil, narediti umetno osebo v slogu glavnega junaka filma *AI* iz leta 2001, Davida, otroku podobnega androida, čeprav je imel morda na umu prav to.

Kakorkoli, zdi se, da argument iz zavesti v tej ali oni obliki niti slučajno ne izključi možnosti obstoja inteligentnih računalnikov.

V Argumenti iz različnih nezmožnosti

V bistvu gre za naštevanje stvari (mogoče ne vseh), za katere ljudje še danes (mogoče ne vsi) menijo, da jih stroji nikoli ne bodo mogli narediti. Turing to formulira takole: »Dopuščam, da lahko pripraviš stroje do tega, da delajo vse te

[17] To je *argument iz analogije za duha drugih*, ki ga je za izognitev solipsizmu razvil Locke, sprejemajo pa ga tudi vsi ostali empiristi, ki so naklonjeni kartezijanskemu pogledu na zavest. Povzeli bi ga lahko tudi takole: ker sta edino, kar vem, obstoj in vsebina mojega duha, mora biti moje znanje o duhu drugih, če ga sploh lahko imam, posredno in analogno, to je sklepanje po sebi (Thornton, 2004).

stvari, ki si jih omenil, toda nikoli ne boš mogel kakšen stroj pripraviti do tega, da bi počel X« (Turing, 1990: 68). Izmed teh stvari, s katerimi lahko nadomestimo X, izpostavimo le nekatere: biti prijazen, imeti smisel za humor, se zaljubiti, ločevati pravilno od nepravilnega, učiti se iz izkušenj, delati napake, biti predmet svoje lastne misli, pravilno uporabljati besede, narediti kaj resnično novega itd.

Vprašanje je, na čem takšni argumenti temeljijo? Turing odgovarja, da na indukciji, saj človek iz tega, da vidi na tisoče strojev, ki so grdi, neuporabni za vse, razen za tisto, čemur so namenjeni (omejeno uporabni), predvidljivi glede svojega obnašanja, nerodni itd., sklepa, da so takšni vsi stroji. Skratka, noben do sedaj viden stroj ni bil zmožen narediti X-a (bit lep, univerzalno uporaben, nepredvidljiv glede svojega obnašanja, spreten itd.), torej X-a ni zmožen narediti noben stroj. Očitno gre pri tem sklepanju za induktivno posplošitev, ker pa mora ta za to, da bi bila dobra, vedno izpolniti kar nekaj pogojev, to je (i) številčnost in raznovrstnost vzorca,[18] (ii) zavestno iskanje nasprotne evidence,[19] i) ujemanje posplošitve s tem, kar o svetu že vemo[20] in (iv) izogibanje nagnjenosti k potrjevanju lastnih stališč [*confirmation bias*],[21] čemur pa dana induktivna posplošitev ne zadosti, ta ni dobra. Namreč, vzorec zajetih primerkov ni izključil možnosti, da se stroji in različne nezmožnosti skupaj pojavljajo zgolj naključno (to je, lahko obstaja računalnik, ki bi vse to naredil); zaradi predsodkov do strojev je šlo slej ko prej za iskanje primerkov v prid posplošitvi in vztrajanje pri njej kljub temu, da v svetu ni na voljo prepričljive razlage za to, da stroji tega niso zmožni narediti. Z drugimi besedami, kot pri vseh induktivnih argumentih, tudi pri posplošitvi vedno obstaja možnost, da je sklep neresničen in ker je ob novi evidenci, to je vedno zmogljivejših računalnikih – več spomina, višja hitrosti delovanja, več operacij – verjetnost, da računalniki niso zmožni delati opisanih stvari, nizka, je ta posplošitev prej ko ne slaba.

Dalje, ali se nam res zdi smiselno pripisati inteligenco zgolj tistim naravnim ali umetnim sistemom, ki lahko delajo te stvari? Zakaj bi morali predpostavljati, da je bitje, ki tega ne zmore, v nekem smislu pomanjkljivo oziroma neinteligentno? Navsezadnje poznamo ljudi, ki niso prijazni, nimajo smisla za humor, se ne zaljubijo, ne ločijo pravilno od nepravilnega itd., pa jim zaradi tega še ne odrekamo inteligence. Res je, da imajo ljudje potencial, da počnejo vse te stvari, pa kljub temu se zdi takšen antropomorfen kriterij za to, komu lahko pripišemo

[18] Z večanjem številčnosti in raznovrstnosti vzorca se zmanjšuje verjetnost, da se lastnosti F in G v našem vzorcu zgolj po naključju pojavljata skupaj, zaradi česar je tveganje, da G-ja v tistih F-jih, ki jih naš vzorec ni zajel, ne bomo našli, manjše.

[19] Še bolj vneto od primerov, ki našo posplošitev potrjujejo, iščemo tiste, ki jo zavračajo, ker s tem zgolj popravljamo škodo, ki nastane, ker smo po naravi bolj pozorni na evidenco, ki naše posplošitve potrjuje. Razen tega ima evidenca, če sprejmemo Popperjevo stališče, ki posplošitev postavlja na laž, pomembnejšo vlogo od tiste, ki jo potrjuje – vsak najden F, ki je G, namreč posplošitev »Vsi F so G« potrjuje zgolj posredno in do neke mere, medtem ko jo vsak primerek F, ki ni G, ovrže neposredno in dokončno.

[20] Bolj ko se naša posplošitev ujema s tem, kar o svetu že vemo – ta pogoj je izpolnjen, če imamo za opaženo povezavo med dvema pojavoma na voljo prepričljivo razlago – bolj razumno jo je sprejeti.

[21] Vztrajanje pri posplošitvah kljub očitni nasprotni evidenci, ki opozarja na njihovo neresničnost.

inteligenco, diskriminatoren do vseh ostalih organizmov. Razen tega hiter napredek na področju računalniških tehnologij in robotike objektivnega opazovalca kar sili k prepričanju, da bodo z nadaljnjim razvojem občutljivih senzorjev in umetnih organov stroji kaj kmalu konzumirali jagode s smetano bodisi ob spremljanju teniških dvobojev v Wimbledonu, kjer je to tradicija, bodisi kje drugje.

Čeprav se bo to bolj konzervativnim bralcem zdela znanstvena fantastika, pa tega ne bi mogli trditi za nekatere druge zmožnosti s seznama, ki se računalnikom prav tako odrekajo, to je delati napake, biti predmet svoje lastne misli, pravilno uporabljati besede in narediti kaj resnično novega. To je možno že sedaj in je bilo, če ne prej, že v 70. letih 20. stol., ki velja za zlato dobo umetne inteligence. Za prvo stvar, to je delati napake, česar stroj naj ne bi bil zmožen, s čimer bi v Turingovem testu razkril svojo identiteto, pravi Turing naslednje: »Stroj (programiran za igro) ne bi poskušal dajati *pravilnih* odgovorov na aritmetične probleme. Namenoma bi uvajal napake, preračunane tako, da zmede izpraševalca« (Turing, 1990: 70). Povedano drugače, ko bi moral stroj izračunati 1341 + 1513, tega ne bi storil takoj, ampak bi si za odgovor vzel približno toliko časa, kot rabi človek, ali pa bi na vprašanje odgovoril napačno.

Glede druge stvari, to je, da stroj ne more biti predmet svoje lastne misli, pa pravi, da ob predpostavki, da stroj ima misel, ima tudi določeno vsebino, saj v primeru, ko mora na primer rešiti neko enačbo, »opazuje« svoje obnašanje in rezultate: »V tovrstnem smislu je stroj nedvomno lahko svoj lastni predmet« (Turing, 1990: 70). Še več, načeloma je mogoče, kmalu pa bo tudi dejansko možno, da stroj »z opazovanjem rezultatov svojega lastnega obnašanja ... spreminja lastne programe tako, da bolj učinkovito doseže neki namen« (Turing, 1990: 70).

To, kar je imel Turing takrat za logično možno, se je dejansko uresničilo že v 70. letih 20. stol., ko so nastali prvi programi, ki so bili sposobni ne zgolj reševati zastavljenih nalog, ki so od njih zahtevale določeno stopnje refleksije, ampak tudi ustvariti lasten postopek za to, kako se sploh rešujejo. Tako je bilo mogoče trditev o nezmožnosti strojev, da bi naredili preostali stvari z zgornjega skrčenega seznama, to je pravilno uporabljati besede in narediti kaj resnično novega, empirično zavrniti že takrat.

Če je to dvoje med drugim pogoj za pripisovanje kognitivnih sposobnosti, je bil prvi program, ki ga je izpolnil, verjetno *Shrdlu*,[22n] ki ga je na MIT-ju napisal Terry Winograd. Gre za program, ki preko robotske roke na mizi premika predmete različnih oblik, barv in velikosti. Komunikacija poteka preko tipkovnice in ekrana, kjer se njegovi odgovori ali vprašanja prikažejo kot podnapisi. Ko dobi navodila, naj na določen način uredi predmete na mizi, program za to razvije in izpelje lasten načrt. Na primer, če ga vprašamo, ali lahko na kocko postavimo piramido, pogleda na mizo, ali je to mogoče že kje narejeno in odgovori, recimo, z da; v nasprotnem primera odgovori z ne vem in poskuša to storiti ter, ko mu to ne uspe, odgovori z ne. Ali, če določimo, da so na

[22] Ime je nesmiselna beseda, ki jo je avtor našel v ameriškem satiričnem magazinu MAD.

primer predmeti, ki so naši, piramide in tisti, ki niso beli, ter ga potem vprašamo, ali je v belem kvadra, ki vsebuje belo piramido in modro kocko, kakšen predmet, ki je naš, program pravilno odgovori, da je to modra kocka, kar očitno kaže na njegove sklepalne zmožnosti. *Shrdlu* je bil tako prvi program, ki je bil sposoben izpolniti določene pogoje za pripis inteligence: razumel je navodila, s pomočjo sklepanja je našel odgovore na zapletena vprašanja, razumel pa je celo, vsaj deloma, motive za svoja dejanja (Copeland, 1993).

Drugi, še naprednejši, program se je imenoval *Hacker*,[23] ustvaril pa ga je na MIT-ju Gerald Sussman. Gre za program, ki piše programe za računalnike, na katerih je sam nameščen. Deluje v podobnem okolju kot *Shrdlu* in tudi tukaj robotska roka premika različne predmete. Ključnega pomena za to, da *Hacker* lahko napiše program, je njegova knjižnica tehnik programiranja, ki ni nič drugega kot shramba navodil o tem, kako se napiše program, ki jo je dobil od svojega avtorja. Poleg tega se podobno kot človeški hekerji uči iz izkušenj, vsak neuspešen program je vir informacij o tem, česar se ne dela, s čimer se s prakso izboljšuje, prav tako pa s tem povečuje in izboljšuje svojo knjižnico tehnik programiranja. Recimo, da mora *Hacker* ustvariti program, to je *napisati postopek izvedbe, ne pa to zgolj izvesti*, kar je delal na primer *Shrdlu*, ki bo omogočal, da kocko A, ki leži pod kocko B, postavimo nad kocko C.



Najprej pogleda v svoj spomin in knjižnico, da bi našel karkoli relevantnega za to nalogo: ali vzorec obnašanja, ki bi mu lahko pomagal ali kakšen podoben program, ki ga že ima. Edino, kar najde, je na primer vzorec *postaviti na*, kar robotski roki omogoča, da en predmet postavi ali na drugega ali na mizo. Dalje vidi, da bi lahko ta vzorec uporabil za to, da bi položil A na C, če A ne bi imel ničesar nad sabo. Potem v svojem spominu in knjižnici išče dalje in najde na primer vzorec spustiti. Temu doda pripono na, po zgledu vzorca postaviti na, in tako dobi vzorec *spustiti na*, kar robotski roki omogoča, da en predmet spusti ali na drugega ali na mizo. Sedaj lahko zadano nalogo opravi: najprej kocko B spusti na mizo, potem pa kocko A postavi na kocko C. Res je, da je kontekst, v katerem deluje *Hacker*, izmišljen in omejen ter da je sposoben napisati zgolj najbolj enostavne programe, kljub vsemu pa ne moremo zanikati, da je s tem razbit mit o računalnikih, ki sami ne morejo pisati programov (Copeland, 1993).

Kljub temu pa je Turing menil, da niti z opisom metode, ki bi stroju pomagala, da bi katero od omenjenih stvari opravil, takrat nikogar ne bi mogel prepričati,

[23] Beseda »hacker« v osnovi označuje programskega odvisnika, sčasoma pa je zaradi tega, ker so se z njo začeli označevati računalniški navdušenci, ki so vdirali v tuje računalnike in pisali viruse, ki naj bi jih onesposobili, dobila pejorativen pomen.

da bi to bilo logično, kaj šele dejansko, možno: »Če trdimo, da stroj zmore početi kakšno od teh stvari, in opišemo tip metode, ki bi jo stroj lahko uporabil, običajno ne bomo zapustili globljega vtisa« (Turing, 1990: 70-71).

Tako ta sicer prikriti argument iz zavesti izgubi večino svoje moči šele dobrih dvajset let pozneje.

VI Ugovor lady Lovelace

Naredimo kratek zgodovinski skok v zanimivo etimologijo tega ugovora. Lady Lovelace je bila namreč resnična oseba, gre za Augusto A. King, grofico Lovelace, iz prve polovice 19. stol., ki je bila edini zakonski otrok pesnika lorda Byrona. Neobičajno za tisti čas je bilo, da je bila izobražena v matematiki, saj je njena mati vztrajala pri tem, da so jo zasebni učitelji v njej poučevali. Kmalu se je seznanila s Babbagom, ki je bil oče prvega sprogramiranega računalnika z imenom *Analitični stroj*, nastalega med leti 1823-1833, ki pa zaradi svoje kompleksnosti in takratnih pomanjkljivih tehnoloških zmožnosti dejansko ni bil nikoli realiziran. Znana je po svojih pripombah na njegov *Analitični stroj*, ki vključujejo to, kar je danes znano pod imenom *algoritem*. Šlo je za zaporedje operacij, ki korak za korakom vodijo do rešitve matematičnih problemov.

Zaradi tega, ker algoritem preko njegove dešifracije računalniku omogoča delovanje, na kar v bistvu mislimo, ko govorimo o programu, se je je v zgodovini računalništva oprijel vzdevek prve računalniške programerke. Videla pa je še dlje od Babbaga, saj je domnevala, da bi lahko *Analitični stroj* komponiral in izvajal katerokoli glasbeno delo, ne glede na njegovo zapletenost ali dolžino. To pa z drugimi besedami pomeni, da bi lahko števila, s katerimi računalnik skladno s pravili manipulira, označevala še kaj drugega razen količin, kar pa že predstavlja prehod iz računanja v *kompjutacijo*, s čimer si je prislužila naziv prerokinje računalniške dobe.

Obstaja več različic ugovora lady Lovelace, izvorna, ki se nahaja v njenih pripombah k *Analitičnemu stroju*, pa pravi, da »analitični stroj nima nobenih pretenzij, da bi karkoli *ustvaril*. Lahko dela *karkoli, kar mu znamo ukazati*, naj naredi« (Turing, 1990: 71). V svojem *Lecture on the Automatic Computing Engine* iz leta 1947 ga opiše takole: »Bilo je rečeno, da lahko računski stroji izvajajo zgolj tiste procese, katere so bili naučeni« (Turing, 2004: 392). Osnovna ideja je, da lahko računalniki za razliko od ljudi naredijo zgolj tisto, kar jim ukažemo, to pa ni ustvarjanje. V standardni obliki je prva različica videti takole:

1. Ustvarjalnost pomeni sposobnost, da narediš nekaj izvirnega.
2. Sledenje nizu pravil, ki jih je programer zapisal zato, da bi proizvedel določeno obnašanje, ne vsebuje ničesar izvirnega.

(∴) 3. Stroji, ki so sprogramirani tako, da se obnašajo, kot je bilo načrtovano, niso ustvarjalni.

4. Ustvarjalnost je bistvena za inteligenco.

∴ 5. Stroji niso inteligentni[24]

Turing se strinja s Hartreejem,[25] ki trdi, da iz tega ne sledi, da *Analitični stroj* ne bi imel sposobnosti, da bi karkoli ustvaril, ampak zgolj, da dokazi, ki so bili na voljo, lady Lovelace niso dopuščali sklepa, da jo ima. To pa ne pomeni, da ne bi bilo mogoče konstruirati elektronske naprave, ki bi sama mislila. Ali je to načeloma možno, je vprašanje, ki ga zastavlja novejši razvoj. Ni pa bilo videti, da bi tedaj konstruirani ali načrtovani stroji imeli to lastnost (Hartree, 1949). Turing dobrih sto let po nastanku *Analitičnega stroja* pravi, da z empiričnega vidika strojev, ki bi bili *ustvarjalni*, ne bi bil problem narediti: namesto da bi jih popolnoma sprogramirali, bi jim dali izkušnje, iz katerih bi se učili in obnašali »po svoje«. V tem primera bi pokazali obnašanje, ki ni bilo načrtovano s strani njihovega programerja, zaradi česar bi jih lahko opisali kot inteligentne: »Vsak stroj bi moral biti opremljen s trakom, na katerem bi bile odtisnjene naključne serije znakov, to je 0 in 1, v enakih količinah, ki bi jih stroj uporabljal pri svojem odločanju. Posledica tega bi bilo obnašanje stroja, ki nikakor ne bi bilo povsem določeno z izkušnjami, ki jih je dobil ...« (Turing, 1951: 475).

Bringsjord, Bello in Ferucci (2001) opozarjajo, da sta nenačrtovano obnašanje in neizvirnost (neustvarjalnost) združljiva oziroma, da stroji zaradi naše neumnosti, neznanja, nepazljivosti itd. res delajo stvari, ki jih programerji niso načrtovali, da pa to še ne pomeni, da jim lahko pripišemo kakršnokoli izvirnost (ustvarjalnost). Vprašanje je, kateri merodajni pojem izvirnosti imajo v mislih? Vsekakor pa lahko skladno s standardnim razumevanjem izvirnosti brez strahu trdimo, da stroji podobno kot ljudje v novih situacijah, v katerih se znajdejo, svojo izvirnost pokažejo, če ne drugače, s tvorbo stavkov v naravnem/strojnem jeziku, ki so nepoznanim okoliščinam primerni in ki jih prej niso imeli nikjer zapisanih.

Razen tega lahko na izvorno pripombo lady Lovelace odgovorimo tudi drugače. Ker smo »vsi programirani z dednostjo in okoljem, kot osebe v kakšnem naturalističnem romanu« (Šuster, 2007: 11), se glede tega pač od strojev, ki so omejeni s svojim programom in spominsko ter procesorsko zmogljivostjo, ne razlikujemo. Niti ljudje ne morejo storiti nič, česar jim ne dovoljujeta njihova biologija in genetika, niti stroji ne morejo storiti nič, česar jim ne dovoljujeta njihov hardver in softver. Govoriti potem o tem, da je razlika med človekom in strojem v tem, da prvi lahko ustvarja, drugi pa ne, zaradi česar je človek inteligenten, stroj pa ne, je milo rečeno šovinistično. Ob večji programski kompleksnosti in procesorski hitrosti, ki bi razlike med obdelovanjem podatkov pri ljudeh in računalnikih izničili, kar je načeloma možno pričakovati že v bližnji prihodnosti, iz tega sledi, da ali mislijo oboji, to je stroji in ljudje, ali pa noben (Oppy, Dowe, 2011).

Druga različica ugovora lady Lovelace trdi, da *stroj nikoli ne more narediti ničesar resnično novega*. Ampak, ali lahko mi kdaj naredimo kaj res novega? Ob

[24] Argument je povzet iz Abramsona (2008: 161), ki tam sicer nastopa v malo drugačnem kontekstu.

[25] Douglas R. Hartree (1897–1957) je bil angleški fizik in matematik, po H. sv. vojni dejaven tudi na področju numeričnih metod in digitalnih računalnikov (EN1AC).

predpostavki, da je svet deterministično,[26] seveda ne, saj so po njem vsa naša dejanja že vnaprej določena z vsoto predhodnih dogodkov, naravnimi zakoni in mejami našega vesolja: »Dogodek (dejanje ali izbira) je *determiniran*, kadar obstajajo predhodni pogoji (recimo volja usode, božja vnaprejšnja določenost, predhodni vzroki in zakoni narave) tako, da je njihov nastop zadostni pogoj za nastop tega dogodka« (Kane: 2005: 6-7). Večina filozofov se sicer strinja s tem, da ni determinizem niti nujno resničen niti nujno neresničen, zaradi česar lahko to, kakšen je, odkrijemo zgolj s pomočjo empiričnih raziskav, ki preučujejo svet, ne pa filozofsko argumentacijo. Vendar dokler tega ne vemo, je od stroja zato, da bi mislil, nesmiselno pričakovati, da stori nekaj, česar morda nismo sposobni storiti niti sami. Ker pa je to, da je svet deterministično, povsem združljivo s tem, da smo v takšnem svetu *presenečeni*, je mogoče bolje, da se pri zavračanju domneve o pametnih strojih v okviru pripomb lady Lovelace sklicujemo na to.

Tretja različica ugovora lady Lovelace tako pravi, da *nas stroj nikoli ne more presenetiti*. Vendar to ne drži, pravi Turing, saj se v praksi pogosto dogaja, da je izid, ki ga dobimo, na primer pri računanju, presenečenje. Abramson njegov odgovor povzame takole:»... [R]ačunalniki me presenečajo ves čas, pa ne zgolj zaradi površnosti pri pisanju programa, ampak tudi zaradi njegovih nepredvidljivih posledic, ki se pokažejo šele, ko začne delovati« (Abramson, 2008: 157). Kritik na to odgovarja, da je to posledica naše nepazljivosti in malomarnosti pri določanju vstopnih predpostavk (vnos), zaradi česar pričakujemo drugačen izstopni rezultat (iznos), kot ga potem dobimo, pri čemer nima stroj nobenih zaslug. Kakorkoli, stroj za nas pripravi presenečenje tudi v tem primeru, saj v končni fazi prihaja od njega, ne glede na to, da je bil tisti, ki je za to poskrbel, človek. Tudi za knjigo namreč pravimo, da nas je presenetila, pa čeprav je za ta naš odziv zaslužen njen avtor: »... [M]orda pa je vredno pripomniti, da presoja nečesa kot presenetljivega zahteva enako 'ustvarjalen mentalni akt', če prihaja presenetljivi dogodek od človeka, knjige, stroja ali česar koli drugega« (Turing, 1990: 72).

Vidimo, da iz nobene različice ugovora lady Lovelace ne sledi nujno, da nekaj takega kot misleči računalniki ne obstaja.

VII Argument iz zveznosti živčnega sistema

V splošnem obstajata dve različni vrsti sistemov: *diskretni*, to je sistemi z diskretnimi stanji, in zvezni, to je sistemi z zveznimi stanji. Diskretne imenujemo tudi digitalni, na primer sodobni računalniki, zvezne pa analogni, na primer človeški živčni sistem. Sistem je diskreten takrat, kadar vsebuje končno število stanj, »diskreten« pa pomeni, da je lahko v določenem trenutku zgolj v enem izmed možnih stanj, recimo v 0 ali 1, to je svojo programsko kodo dešifrira časovno ločeno oziroma zaporedno. Zvezen je sistem takrat, kadar nima končnega števila stanj, »zvezen« pa pomeni, da je lahko obenem v različnih

[26] Včasih je veljalo, da biti determiniran nujno pomeni biti povzročen, to je vsak dogodek je nujen vzročni učinek predhodnih dogodkov, zaradi česar je bil vsak determinizem vzročni, danes pa to ne velja več (Šuster, 2007: 10).

možnih stanjih, to je njegovi vhodi in izhodi se lahko spreminjajo v vsakem trenutku, in podatke obdeluje časovno hkrati oziroma vzporedno.

Argument iz zveznosti živčnega sistema temelji na ideji, da človeški možgani niso podobni digitalnemu računalniku, saj lahko majhna napaka v velikosti vhodnega živčnega dražljaja povzroči veliko razliko v velikosti izhodnega. Recimo, če v računalnik preko tipkovnice vnesemo »mačka« (vnos), računalnik preklaplja med 0 in 1, odvisno od kode, ki jo vsaka črka ima, dokler ne pride do binarnega zapisa na primer 1000011 1000001 1000111 1001111 1000001, ki pomeni izpis »mačka« na mojem ekranu (iznos). Če se pri vnosu zmotimo in namesto *m* vtipkamo *n*, to pomembnejše posledice za iznos, razen tega, da bo v izpisu »mačka« na našem zaslonu napaka in bo pisalo, recimo, »načka«, kar pa bomo hitro ugotovili in popravili, ne bo imelo. To pa ne velja za možgane. Če ni vzdraženo dovolj veliko število nevronov (vnos) kot posledica na primer izpostavljenosti roke ognju, to je, zgodi se majhna napaka, ustrezni del možganov ne poskrbi za to, da pride do krčenja in raztezanja mišic, kar ne pripelje do umika roke (iznos), česar pa ne moremo ugotoviti in roka nam zgori: »Živčni sistem gotovo ni stroj z diskretnimi stanji. Majhna napaka v informaciji o velikosti živčnega impulza, ki zadeva na nevron, lahko povzroči veliko razliko v velikosti izhodnega impulza. Zato bi se dalo argumentirati, da ni mogoče pričakovati, da je obnašanje živčnega sistema mogoče imitirati s sistemom diskretnih stanj« (Turing, 1990: 72). Tako so možgani zvezni sistem, računalniki pa diskretni, ker pa gre prava inteligenca z roko v roki le z zveznimi sistemi, ne pa tudi z diskretnimi, stroji niso inteligentni, medtem ko ljudje so. V standardnem zapisu lahko argument iz zveznosti živčnega sistema povzamemo takole:

1. Pravo misel oziroma inteligenco ima lahko izključno zvezni sistem.
2. Diskretni sistem ni zvezni sistem.
3. Stroj je diskretni sistem.

∴ 4. Stroj ne more imeti prave misli oziroma inteligence.

Turingova zavrnitev tega argumenta se opira na prepričanje, da je sicer res, da se mora stroj z zveznimi stanji ločiti od stroja z diskretnimi stanji, da pa v igri oponašanja te razlike ne moremo izkoristiti. Ima obliko sklepanja po analogiji: podobno kot lahko digitalni računalnik z majhno stopnje napake posnema diferencialni analizator,[27] ki je prav tako zvezni sistem, lahko z majhno stopnjo napake posnema tudi človeško obnašanje, ki je posledica obdelave podatkov v naših možganih: »Digitalni računalnik ne bi mogel točno predvideti, kakšen odgovor na neki problem bi dal diferencialni realizator, toda bil bi povsem sposoben dajati odgovore prave vrste« (Turing, 1990: 72). Z drugimi besedami, v Turingovem testu bo stroj pri odgovarjanju na vprašanja sicer delal napake, ki pa bodo tako majhne, da jih običajni izpraševalec ne bo odkril, zaradi česar pri ugotavljanju, kje tiči računalnik, od tega ne bo imel nobene koristi in ga bo slednji brez težav prevaral.

[27] Diferencialni analizator je mehanski analogni računalnik s kolesi in diski za reševanje diferencialnih enačb z integriranjem.

Turingov odgovor žal ne napade prave premise v argumentu, to je pravo misel oziroma inteligenco ima lahko izključno zvezni sistem, zdi pa se, da bi mu edino njena zavrnitev prinesla uspeh. Namreč, če nekdo vanjo verjame, ne bo pripravljen pripisati inteligence nobenemu stroju, ki ni zvezen, ampak bo v primeru, ko se bo pojavil računalnik, ki bo opravil Turingov test, odgovoril, da to dokazuje zgolj neprimernost takšnega preizkusa za preverjanje inteligence (Oppy, Dowe, 2011). Povedano drugače, sposobnost stroja, da nas v Turingovem testu pretenta, ker pač ne moremo izkoristiti razlike med zveznimi in diskretnim sistemi, ki se kaže v majhnih napakah slednjih v odgovorih na naša vprašanja, je kljub vsemu združljiva s tem, da ni inteligenten. Bolje bi bilo, da bi Turing podvomil o resničnosti prve premise, kot to stori Block v svojem članku »Psychologism and Behaviorism« iz leta 1981. Njegova poanta je, da naš pojem inteligence dovoljuje, da ima inteligentno bitje umetne čutne organe, kar pomeni, da v primeru, ko bi bile oči marsovcev kot filmske kamere, preko katere bi se prenašali podatki, zapisani kot pikčaste slike, podobne nekdanjim časopisnim, in v primeru, ko bi nam bili takšni marsovci podobni po videzu, obnašanju in celo notranji obdelavi podatkov, nihče ne bi trdil, da zaradi svojih filmski kameri podobnih oči niso inteligentni. Toda, ker obstaja končno število takšnih pikčastih slik, obstaja tudi končno število možnih vizualnih dražljajev in odgovorov, to pa pomeni, da marsovci niso zvezni sistem. Sploh pa, ali imamo dober razlog za predpostavko, da ljudje nismo diskretni stroji?

Kakorkoli, zdi se, da argument iz zveznosti živčnega sistema počiva na vsaj eni sporni premisi.

VIII Argument iz neformalnosti obnašanja

Obstajajo *pravila ravnanja*, za katera Turing pravi, da »... [P]omenijo predpise, kot na primer ‘Ustavi se, če vidiš rdečo luč’, po katerih lahko delujemo in se jih zavedamo« (Turing, 1990: 73). Vprašanje je, ali lahko ustvarimo popolno množico pravil, to je seznam vseh pravil, ki bi določal, kaj naj naredimo v vseh možnih situacijah? Vzemimo kot primer semafor. Ob prižgani rdeči luči se ustavimo, ob prižgani zeleni gremo dalje. Kaj pa, če se zaradi napake prižgeta obe skupaj? Za takšno naključno situacijo nimamo na voljo pravila ravnanja, včasih je bolje, da se ustavimo, včasih, da peljemo dalje, karkoli pa naredimo, lahko povzroči novo naključno situacijo, za katero spet ne bomo imeli predpisanega obnašanja: »Poskus, da bi formulirali pravila ravnanja, ki bi pokrila vse naključne možnosti, celo tiste, ki jih povzročajo semaforji, je videti nemogoč« (Turing, 1990: 73).

Turing se s tem strinja, zavrača pa *argument iz neformalnosti obnašanja*, ki naj bi iz tega sledil: »Če bi vsak človek imel določeno množico pravil ravnanja, po katerih bi uravnaval svoje življenje, ne bi bil nič boljši od stroja. Takih pravil ni. Torej, ljudje ne morejo biti stroji« (Turing, 1990: 73). Če ga zapišemo v silogistični obliki, se glasi takole:

1. Vsak človek z množico pravil za uravnavanje svojega življenja je stroj.

2. Noben človek nima takih pravil.

∴ 3. Noben človek ni stroj.

Očitno je, zakaj ga ne sprejema, saj *ne* gre za veljavno obliko sklepanja, ampak za formalno napako, in sicer za zmoto *nedistribuiranega srednjega termina*.[28] Glede na zgradbo zmotnega argumenta pa je ta vedno ali zmota zanikanega antecedensa ali zmota zatrjenega konsekvensa.[29]2 Da bi ugotovili, za katero izmed njiju gre, prevedimo silogizem v pogojnike:

1. Če bi vsak človek imel določeno množico pravil ravnanja za uravnavanje svojega življenja, bi bil stroj.
2. Če si človek, potem nimaš takih pravil.

∴ 3. Če si človek, potem nisi stroj.

Ker je druga premisa zanikani antecedens pogojnika v prvi premisi, sklep pa zanikani konsekvens istega pogojnika, gre v zgornjem primera za *napako zanikanega antecedensa*, pri kateri »'če napačno jemljemo kot 'če in samo če'« (Uršič, Markič, 2009: 75). V našem argumentu bi namreč ljudje lahko bili stroji, četudi nimamo takšnih pravil, to je ni izključeno, da ljudje ne bi bili stroji zaradi kakšnega drugega razloga.

Dalje, argument iz neformalnosti obnašanja je zmoten tudi zato, ker meša pojma »naj bi« *[ought]* in »bo« *[will]*, to je v premisi, kjer opisuje ljudi, uporablja »naj bi«, v premisi, kjer opisuje stroje pa »bo«. Da lahko to ponazorimo, ga moramo malo preoblikovati:

1. Ne obstaja niz pravil, ki določajo, kaj *naj bi* človek naredil v vseh možnih situacijah.
2. Obstaja niz pravil, ki določajo, kaj *bo* stroj naredil v vseh možnih situacijah

∴ 3. Človek ni stroj.[30]

Da bi se tej napaki izognili, bi morali ta pomenski razkorak pri srednjem terminu – *obnašanje* v *vseh možnih situacijah* – premostiti, to pa je možno le tako, da ali v obeh premisah uporabimo »naj bi«, zaradi česar pravila ostanejo pravila, ali da v obeh premisah uporabimo »bo«, zaradi česar pravila postanejo *zakoni*. Vprašanje je, kaj to za zagovornika argumenta iz neformalnosti obnašanja pomeni.

Prva možnost: (i) obstaja niz pravil, ki določajo, kaj *naj bi* tako stroj kot tudi človek naredila v vseh možnih situacijah.[31] Ali ta trditev drži? En način, ki se ga ljubitelji neinteligentnih strojev poslužujejo pri njenem zavračanju, je, da preprosto, brez kakršnegakoli dodatnega razloga, trdijo, da stroji niso takšne vrste stvari, za katere bi veljala kakršnakoli pravila. To pa seveda ni dovoljen

[28] Zagrešimo jo, ko v kategoričnem silogizmu srednjega pojma ne distribuiramo (kvantificiramo) niti v višji niti v nižji premisi.
[29] Za formalne zmote glej na primer Uršič, Markič (2009).
[30] V tem modificiranem argumentu je 1. premisa enaka 2. premisi izvornega argumenta, 2. premisa pa je zamolčana domneva, ki prej v izvornem argumentu ni nastopala.
[31] Zavoljo argumenta zanemarimo dejstvo z začetka tega ugovora, da velja splošno prepričanje, da takšen niz popolnih pravil obnašanja ne more obstajati.

način dokazovanja v tem primeru, saj zagovorniki nemislečih računalnikov s tem predpostavljajo resničnost tistega, kar bi morali šele dokazati, zaradi česar se *izogibajo vprašanju* ali dokazujejo v krogu (Oppy, Dowe, 2011). Ne glede na to, kako to imenujemo, gre za zmoten način sklepanja, ki resničnosti (i) ne spodbije, njih pa vodi do nesprejemljivih posledic. Kakorkoli, odgovore na zgornje vprašanje lahko strnemo v obliki konstruktivne dileme:

1. Ali takšna *pravila* obstajajo ali ne.
2. Če obstajajo, potem obstajajo za stroje in ljudi.
3. Če ne obstajajo, potem ne obstajajo za nikogar.

∴ 4. Ali takšna *pravila* obstajajo za stroje in ljudi ali za nikogar.

Vidimo, da vodi (i) nasprotnike mislečih strojev oziroma zagovornike argumenta iz neformalnosti obnašanja v dilemo, ki je zanje nerešljiva, saj jim noben izmed rogov ne zagotovi razlike, na kateri bi lahko utemeljili svojo idejo, da so ljudje inteligentni, računalniki pa ne.

Druga možnost: (ii) obstaja niz pravil, ki določajo, kaj *bosta* tako stroj kot tudi človek storila v vseh možnih situacijah (zakoni). Ali ta trditev drži? Tudi odgovore na to vprašanje lahko povzamemo v obliki podobne konstruktivne dileme kot pri prvi možnosti:

1. Ali takšni *zakoni* obstajajo ali ne.
2. Če obstajajo, potem obstajajo za stroje in ljudi.
3. Če ne obstajajo, potem ne obstajajo za nikogar.

∴ 4. Ali takšni *zakoni* obstajajo za stroje in ljudi ali za nikogar.

Tudi (ii) vodi nasprotnike inteligentnih strojev, to je zagovornike argumenta iz neformalnosti obnašanja, v dilemo, ki je zanje nesprejemljiva: noben izmed rogov jim ne daje razlike, s katero bi lahko upravičili prepričanje, da so ljudje misleči, računalniki pa ne.

Verjetno ni treba posebej poudarjati, da s tem argument iz neformalnosti obnašanja pri dokazovanju tega, da je človek drugačen od stroja, ni uspešen.

IX Argument iz izvenčutne percepcije

Zadnji ugovor ideji o obstoju inteligentnih strojev temelji na tako imenovani izvenčutni zaznavi, ki naj bi obsegala telepatijo,[32] jasnovidnost,[33] prekognicijo[34] in psihokinezo.[35] »Zdi se, da ti vznemirljivi pojavi zanikajo vse naše običajne znanstvene predstave. Kako bi si jih želeli diskreditirati! Zal je statistični dokazni material, vsaj za telepatijo, ogromen« (Turing, 1990: 74). Iz tega izhaja,

[32] Telepatija je prenos informacij od ene do druge osebe brez kakršnekoli fizične interakcije med njima.
[33] Jasnovidnost je način pridobivanja informacij s procesom senzorne deprivacije.
[34] Prekognicija je način pridobivanja informacij o prihodnosti brez izpeljave iz sedanjosti.
[35] Psihokineza ali telekineza je neposredni, to je nefizikalni, vpliv duha na gibanje ali spreminjanje snovi.

da ljudje niso stroji, saj niso sposobni izvenčutne zaznave, kar bo spraševalec v Turingovem testu s pridom izkoristil in stroj pravilno identificiral.

Zanimiva je resnost, s katero Turing argument iz izvenčutne percepcije obravnava, še več, zdi se mu, da »se izvenčutna zaznava lahko izkaže za zadnjo razliko med ljudmi in stroji, ki jih ustvarjajo« (Hofstadter, Dennet, 1990: 75). S tem, ko pravi »ta argument se mi zdi močan« (Turing, 1990: 74), to tudi neposredno priznava. Meni, da bi bila izvenčutna zaznava zelo relevantna za mišljenje, saj bi človek z njeno pomočjo na vprašanje, katere barve je karta v moji desni roki, pravilno odgovoril 130 krat v 400 primerih, stroj, ki lahko le ugiba, pa morda 104 krat (Turing, 1990: 74). Še posebej trdni so se mu morali zdeti dokazi za telepatijo, saj je predlagal, da bi bili udeleženci Turingovega testa zaprti v sobi, ki bi bila telepatsko nepredirna.

Kljub temu ni nobenega razloga, da v primeru obstoja izvenčutne zaznave, lete ne bi bili sposobni izvajati tudi stroji, pa tudi z nobenimi naprednimi senzorji ne bi rabili biti opremljeni, saj gre vendar za brezčutno pridobivanje informacij. Kakorkoli, morda so bili dokazi za izvenčutno zaznavo leta 1950 res močni, danes, dobrih šestdeset let kasneje, pa gotovo niso več. Naš namen tukaj ni predstavljati dokaze za to, da telepatija ne deluje, zadovoljili se bomo s tem, da obstoj nečesa tako paranormalnega, za kar v vsakdanjem življenju ne najdemo absolutne nobene evidence in je prej verjetno plod naše želje po nadnaravnem in ezoteričnem, kateri je duh pogosto podvržen, zavračajo skoraj vsi fiziki in psihologi, ki so se s tem kadarkoli ukvarjali. Prav tako je malo verjetno, da bi bilo nekaj takega združljivo z našimi naravnimi zakoni, na kar Turing v znak tolažbe namiguje, saj si je težko zamisliti: »Kako bi se lahko znanost, ki je tako dobro delovala za tako veliko stvari, izkazala za tako napačno? Izziv, da bi ponovno premislili celotno znanost od njenih najbolj bazičnih predpostavk naprej, bi bil velika intelektualna pustolovščina, toda dokazov, da bomo morali to storiti, se z leti preprosto ni nabralo« (Hofstadter, Dennet, 1990: 75).

Po predstavitvi vseh devetih ugovorov in njihovih zavrnitev Turing skromno pripomni: »Bralec je ugotovil, da nimam zelo prepričljivih pozitivnih argumentov v podporo svojih stališč. Če bi jih imel, se ne bi tako trudil z izpostavljanjem napak v nasprotnih stališčih« (Turing, 1950: 454). Vsaj dve stvari, ki temu nasprotujeta, velja na kratko omeniti, obe sta teoretični predpostavki, ki pa, če ju sprejmemo, naredita vprašanje o tem, ali stroji mislijo, za empirično. Prva je, da pripisujemo inteligenco ljudem na osnovi njihovega obnašanja, s čimer postane behavioristični Turingov test ustrezen kriterij za pripisovanje mišljenja strojem, njegova sicer sporna trditev, da bi ga stroj lahko opravil, pa empirično preverljiva. Druga pa je, da je razvoj funkcionalističnih teorij duha z idejo, da so mentalna stanja funkcionalna stanja, ki jih je mogoče realizirati z različnimi materiali, ideji o mislečih strojih zagotovil varnejši teoretični okvir, omenjena večvrstna realizacija pa empirično preverljivost (Oppy, Dowe, 2011).

In v čem se je potem Turing motil, kot namiguje naslov? Zgolj v tem, da bodo ljudje 50 let po objavi njegovega članka »Stroji, ki računajo in inteligenca«, misleče stroje jemali kot nekaj samo po sebi umevnega. Kljub temu, da imamo

na voljo še več dokazov za to, da takratni argumenti, ki pa so v marsičem aktualni še danes, ne vzdržijo, laična in strokovna javnost ideji o inteligentnih strojih še vedno ni naklonjena. Mogoče pa iz zanimanja za to vprašanje ljudje poznajo kakšne druge razloge, ki jih odvračajo od tega, da bi sprejeli trditev o mislečih računalnikih, morda celo tiste, ki jih bomo predstavili in ocenili v naslednjih številkah *Analize*.

TURING WAS MISTAKEN

The article introduces nine objections to the idea of thinking machines from the 1940s. They were a result of laics and experts back then and all collected in 1950 by Turing in his breaking article »Computing Machinery and Intelligence«. Besides, we are going to show how Turing refuted all of them which turned a non-serious debate about how could be intelligent something like the electronic brain as computers have been labeled into a serious one. With respect to Turing's rejections the article introduces and develops new ideas in favour of them and thus support his belief that thinking machines are in principle possible.

Keywords: artificial intelligence, cognitive science, Turing, Lady Lovelace, solipsism.

Literatura

Abramson, D. (2008). »Turing's Responses to Two Objections«. *Minds and Machines*, 18, str. 147–167.

Akvinski, T. (1265–1274/1947). *Summa Theologica*. New York: Benziger Brothers.

Aristotel (1970). *Organon*. Beograd: Kultura.

Block, N. (1981). »Psychologism and Behaviorism«. *Philosophical Review*, 90, str. 5–43.

Bregant, J. (2010). »Ali lahko stroj misli?«. *Analiza*, 4, str. 55–72.

Bringsjord, S., Bello, P., Ferucci, D. (2001). »Creativity, the Turing Test, and the (Better) Lovelace Test«. *Minds and Machines*, 11, str. 3–27.

Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction*.

Oxford: Blackwell.

Descartes, R. (1637/2007). *Razprava o metodi*. Ljubljana: ZRC SAZU.

Descartes, R. (1642/1988). *Meditacije*. Ljubljana: Slovenska matica.

Descartes, R. (1897–1913). *Oeuvres de Descartes*. Adam, C., Tannery P. (ur.). Paris: Leopold Cerf. (AT)

Descartes, R. (1642/1985). *Objections and Replies. V* Cottingham, J., Stoothoff, R., Murdoch, D. (ur.), *The Philosophical Writings of Descartes*, Vol. II, Cambridge: Cambridge University Press. (CSM)

Gödel, K. (1931). »Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme«. *Monatshefte für Mathematik und Physik*, 38, str. 173–198.

Hofstadter, D. R., Dennet, D. C. (1990). »Stroji, ki računajo, in inteligenca. Refleksije«. V Hofstadter, D. R., Dennet, D. C. (ur.), *Oko duha*, Ljubljana: Mladinska knjiga.

Hartree, D. R. (1949). *Calculating Instruments and Machines*. Urbana: The University of Illinois Press.

Jefferson, G. (1949). »The Mind of Mechanical Man«. *British Medical Journal*, 1, str. 1105–1121.

Kane, R. (2005). *A Contemporary Introduction to Free Will*. Oxford: Oxford University Press.

McGinn, C. (1999). *The Mysterious Flame*. New York: Basic Books.

Oppy, G., Dowe, D. (2011). »The Turing Test«. *The Stanford Encyclopedia of Philosophy*, http://www.plato.stanford.edu.

Penrose, R. (1999). *The Emperor's New Mind*. Oxford: Oxford University Press.

Sextus Empiricus (160–210/2000). *Outlines of Pyrrhonism*. Cambridge: Harvard University Press.

Šuster, D. (2007). »Problem svobode«. V Šuster, D. (ur.), *O svobodni volji*. Maribor: Aristej.

Thornton, S. P. (2004). »Solipsism and the Problem of Other Minds«. *Internet Encyclopedia of Philosophy*, ISSN 2161-0002, http://www.iep.utm.edu.

Turing, A. (1947/2004). »Lecture on the Automatic Computing Engine«. V Copeland , J. (ur.), *The essential Turing*, Oxford: Oxford University Press.

Turing, A. (1950). »Computing Machinery and Intelligence«, *Mind*, 59, str. 434–460.

Turing, A. (1950/1990). »Stroji, ki računajo in inteligenca«. V Hofstadter, D. R., Dennet, D. C. (ur.), *Oko duha*, Ljubljana: Mladinska knjiga.

Turing, A. (1951/2004). »Can Digital Computers Think?«. V Copeland, J. (ur.), *The essential Turing*, Oxford: Oxford University Press.

Uršič, M., Markič, O. (2009). *Osnove logike*. Ljubljana: FF.

Wittgenstein, L. (1977). *Philosophische Untersuchungen*. Frankfurt: Suhrkamp.