**ANαliZA**

časopis za kritično misel

Letnik 14, številka 4, december 2010 · ISSN 1408-2969

FILOZOFIJA IN UMETNA INTELIGENCA

Janez Bregant

# Ali lahko stroj misli?

Filozofska fakulteta Univerze v Mariboru

**Citiranje:** Janez Bregant, »Ali lahko stroj misli?«, *Analiza: časopis za kritično misel* 14, št. 4 (2010): 55–72.

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 izvirna tiskarska datoteka številke, ki ji je bil leta 2026 dodan ta naslovni list. Strani in prelomi ustrezajo tiskani izdaji.*

Janez Bregant
*Filozofska fakulteta Univerze v Mariboru*

# Ali lahko stroj misli?

V tem članku bomo govorili o umetni inteligenci kot delu kognitivne znanosti, ki izhaja iz tega, da naš duh deluje po kompjutacijskih načelih. V prvem delu si bomo ogledali vizionarje, ki so že pred več stoletji ugotavljali, da je mišljenje zgolj računanje. Drugi del članka predstavlja formalne pogoje, ki jih mora nek sistem izpolniti, da bi ga lahko imenovali računalnik. Tretji del opisuje Turingov test, to je preizkus sposobnosti stroja, da pokaže svojo inteligenco, ki ga področje umetne inteligence jemlje kot dokaz za to, da smo stopili v dobo mislečih računalnikov, s čimer pa se filozofi ne strinjajo. Štiri osnovne ugovore si bomo ogledali v četrtem delu, najbolj obetavnega, ki se imenuje *ugovor iz simulacije*, pa bomo skušali zavrniti z dopolnitvijo Turingovega testa in ga tako ohraniti pri življenju.

*Ključne besede*: umetna inteligenca, kognitivna znanost, Turingov stroj, Turingov test, klasična UI, konekcionizem.

## 0 Uvod

Ali lahko stroj misli"? To je vprašanje, ki se ni prvič pojavilo šele v 20. stol., ko so bili računalniki zreli za množično proizvodnjo, ampak že daljnjega leta 1637, ko je Descartes trdil, da noben stroj nikoli ne bo sposoben simbolne manipulacije, ki bi mu omogočala tvorjenje smiselnih stavkov, na osnovi česar bi mu lahko pripisali mišljenje. Šele po stoletjih čakanja in abstraktnih strojih, ki računajo, je ideja o mislečem stroju s prihodom prvih elektronskih računalnikov v 50. letih 20. stol. dobila nov zagon. Leta 1956 se je na slavni darthmouthski konferenci rodila nova disciplina, ki jo lahko preprosto imenujemo *umetna inteligenca* (UI).[1] Njen cilj ni zgolj posnemati mišljenje, ampak narediti stroj, ki misli, v polnem pomenu te besede: »… [V]časih je definirana kot ›znanost o izdelavi strojev, ki so sposobni narediti stvari, za katere je po naših merilih potreben um‹[2] (Copeland, 1993: 1).

[1] S pojmom »inteligenca« mislimo na izvajanje umskih/spoznavnih operacij, s pojmom »umetna« pa na to, da jih ne izvaja človek, ampak stroj/sistem (to je nekaj, kar ni rezultat narave, ampak človekovega ali strojevega dela).

[2] Ali »Umetna inteligenca je znanost o tem, kako narediti in/ali programirati računalnike, da bodo sposobni istih stvari kot um« (Boden, 1990: 1).

Nekateri teoretiki predlagajo drugačno definicijo umetne inteligence in jo razumejo kot »*znanost o inteligenci na sploh* …« (Boden, 1990: 1), katere cilj je pojasniti ali celo reproducirati različne zmožnosti duha in ugotoviti, ali lahko inteligenca nastane zgolj v sistemih, ki imajo podobno zgradbo kot možgani ali pa jo je možno realizirati tudi na kakšen drug način. Ker to ne vključuje poznavanja in uporabe računalniške tehnologije, ampak razvijanje abstraktnih principov, ki so organizirani na ustrezen način, razlike med umetno inteligenco, psihologijo in filozofijo duha izginejo. Predmet raziskav je vedno enak, ne glede na to znotraj katere izmed naštetih disciplin se gibljemo in katero orodje je tisto, ki ga skladno s tem uporabljamo. Iz tega se je rodila novo področje z imenom *kognitivna znanost*,[3] ki vključuje vse našteto in predstavlja skladno raziskovanje spoznave, inteligence in duha (*glej* Haugeland, 1986: 5).

Katerokoli definicijo umetne inteligence sprejmemo, ne moremo mimo tega, da so njen razcvet v 20. stol. zaznamovala tri odkritja: formalna teorija kompjutacije, prihod računalnikov, ki so sposobni realizirati predpisano »računanje«, in odkritje nevrona. Iz tega sta se danes na področju umetne inteligence razvili dve smeri, in sicer *klasična UI* in *konekcionizem*.[4] Prva stavi pri raziskovanju tega, kako mora biti zgrajen sistem, da bi bil inteligenten, na idejo, ki je znana pod imenom *simbolna hipoteza*, trdi pa, da lahko računalniki manipulirajo s poljubnimi »figurami« na katerikoli način, ki jim ga predpišemo, tako da moramo za to, da bi mislili, namesto figur samo še uporabiti simbole in manipulacije označiti kot racionalne. Druga pa se pri iskanju organizma, ki je zmožen proizvesti mišljenje, običajno zateka k izdelavi medsebojno povezanih mrež, sestavljenih iz enostavnih in uniformnih enot, izmed katerih so najbolj običajne tako imenovane *nevronske mreže*, v katerih enote predstavljajo nevrone, povezave med njimi pa sinapse.

Pri tem ne gre za znanstveno fantastiko, ampak za pravo znanost, ki je utemeljena na ideji, da smo v bistvu tudi sami računalniki. Umetna inteligenca tako nima ničesar opraviti z naprednimi tehnologijami, vsaj ne v smislu, da bi morali do potankosti poznati delovanje sodobnih strojev ali jih celo razvijati, ampak zgolj s teoretičnimi predpostavkami, ki jih je možno ali pa tudi ne, realizirati v današnjih računalnikih. Obe smeri črpata navdih iz prepričanja, da delajo računalniki res enake stvari kot naši možgani, razlikujeta se zgolj v načinu izvrševanja tega. »Na enak način, kot je v računalniku zakodirana vsebina računalniških reprezentacij – simbolov, je tudi v možganih zakodirano znanje. V obeh primerih gre za fizično uprimerjene simbolne strukture (Markič, 1997b: 73). Skratka, mišljenje je kompjutacija, ki vključuje simbolno manipulacijo, ta pa je lahko realizirana ali v računalnikih kot sistemih, ki to »prevajanje« izvajajo preko preklapljanja med 0 in 1 kot elementoma binarnega sistema (zaporedna obdelava podatkov), ali v, na primer, naprednih robotih, ki so sestavljeni iz umetnih možganov in to prevajanje izvajajo preko hkratnega »gorenja« s sinapsami povezanih umetnih nevronov, kot delov umetnih nevronskih mrež (vzporedna obdelava podatkov) .

[3] »Kognitivna znanost« ali »Nova znanost duha«, kot jo imenuje Howard Gardner, je interdisciplinarno področje, ki ga sestavljajo filozofija, psihologija, računalništvo, lingvistika, nevroznanost in antropologija« (Markič, 1995: 1).

[4] Za razliko med njima *glej* Markič, 1997b.

V tem članku bomo govorili o umetni inteligenci kot delu kognitivne znanosti, ki izhaja ravno iz te stare ideje, da naš duh deluje po kompjutacijskih načelih, kar lahko imenujemo »*računalniška metafora*« (Miščević, 2001: 9). V prvem delu si bomo ogledali mislece, ki so že pred več stoletji ugotavljali, da je mišljenje zgolj kompjutacija, pa tega zaradi različnih problemov niso bili sposobni zadovoljivo pojasniti in ki so celo že predvideli nastanek avtomatov. Drugi del članka predstavlja formalne pogoje, ki jih mora nek sistem izpolniti, da bi ga lahko imenovali računalnik, in iz zgodovinskega vidika osvetli njihov prihod. Tretji del opisuje Turingov test, to je preizkus sposobnosti stroja, da pokaže svojo inteligenco, ki ga (ali so ga) številni znanstveniki s področja umetne inteligence jemali kot dokaz za to, da smo stopili v dobo mislečih umetnih sistemov, čemur filozofi s svojimi vprašanji in ugovori nasprotujejo. Štiri osnovne ugovore si bomo ogledali v četrtem delu, ostale pa si bomo pustili za kdaj drugič. Najbolj obetavnega, ki se imenuje *ugovor iz simulacije*, bomo skušali zavrniti z dopolnitvijo Turingovega testa in slednjega tako zaenkrat ohraniti kot kriterij za preizkus inteligentnosti stroja.[5]

## I Vizionarji

»Mišljenje ni nič drugega kot računanje« pravi Hobbes leta 1651 v *Leviathanu* (Hobbes, 2006: 32). Ker gre za prvo formulacijo misli v pojmih računanja, ga lahko upravičeno imenujemo praoče umetne inteligence. Omenjene besede vključujejo dve osnovni ideji: (i) mišljenje je mentalni proces, ki obsega simbolno manipulacijo, ki se dogaja v možganih. Zaradi tega so primerki misli v bistvu primerki možganov, ali kot jih imenuje Hobbes, »fantazme«: »Ko človek razmišlja ne dela ničesar drugega kot, da si zamišlja končno vsoto, ki jo dobi s seštevanjem delov, ali ostanek, ki ga dobi z odštevanjem ene vsote od druge … Te operacije se ne dogajajo samo pri številih, ampak pri vseh načinih stvari, ki jih lahko združujemo ali odvzemamo« (Hobbes, 2006: 32). In (ii), mišljenje je najbolj čisto, ko zvesto sledi metodičnim pravilom. Rečemo lahko, da je mišljenje mehanski proces, je nekaj takega kot upravljanje mentalnega abaka: fantazme premetavamo sem in tja skladno s pravili razuma, podobno kot na pravem abaku skladno z računskimi pravili sem in tja premikamo kroglice (*glej* Haugeland, 1986: 23). Hobbesov materializem dovoljuje obstoj umetnih sistemov/živali, ali kot jih sam imenuje *avtomatov*, saj po njegovem življenje ni nič drugega kot gibanje telesnih organov, ki je je podobno gibanju zobnikov, vzmeti in koles pri stroju: »Kaj pa je srce drugega kot vzmet, živci strune in sklepi kolesa, kar telesu skupaj omogoča gibanje …?« (Hobbes, 2006: 9).

Če se zdi Hobbes umetni inteligenci naklonjen, pa tega z zgodovinskega vidika ne bi mogli reči za Descartesa. Pri odgovoru na osrednje vprašanje znotraj filozofije duha, to je, kakšno je razmerje med mentalnim in fizičnim, ki je znano pod imenom *problem duha in telesa*,[6] se zateče v *dualizem*, ki predpostavlja obstoj dveh različnih in ločenih substanc, misleče/mentalne – *res cogitans* in raz-

[5] Za koristne namige in pripombe se zahvaljujem Bojanu Borstnerju in Danilu Šustru ter Moniki Košarkoski, tudi za to, ker je moj neusahljiv vir navdiha.

[6] Zanj *glej* na primer Kim, J., 2001.

sežne/fizične – *res extensa*. V *Razpravi o metodi* to leta 1637 zapiše takole: »Potemtakem je ta Jaz, to se pravi um, po katerem edinem sem, kar sem, reč, ki je popolnoma različna od telesa in tudi mnogo lažje dostopna spoznavanju kakor telo. Povsem jasno je, da bi um sam ne prenehal biti vse, kar jè, čeprav telesa ne bi bilo« (Descartes, 1957: 60–61). Lastnosti, ki jih pripisuje duhu, so na primer misel, želja, prepričanje itn., lastnosti, ki jih pripisuje telesu pa na primer razsežnost, velikost, oblika itn. Skratka, duh je stvar, ki misli in čuti, telo pa stvar z maso in lokacijo. Obe substanci sta različni, med njima pa poteka vzajemno delovanje (interakcija). Tako je njegova rešitev problema duha in telesa, da so mentalni pojavi, na primer ljubezen, drugačni in ločeni od fizičnih pojavov, na primer »gorenja« nevronov. Osnovna težava dualizma substanc, na katero so Descartesa opozorili že njegovi sodobniki, je, kako natanko poteka vzročna veriga od na primer želje po vodi, kar pripada duhu, do dviga kozarca, napolnjenega z njo, in pitja, kar pripada telesu, če pa misleča stvar sploh nima lokacije v prostoru. Kraljica Elizabeta Češka je na primer Descartesa spraševala, kako lahko človeški duh vpliva na aktivnosti telesa (na primer krčenje in raztezanje mišic), če pa je zgolj misleča stvar. Drugi Descartesov sodobnik Pierre Gassendi pa je temeljni problem dualizma substanc izpostavil z besedami: »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?« (Descartes, 1985: 238). Kakorkoli, Descartes si je bil že v tistem času, govorimo o letu 1637, sposoben zamisliti govoreči stroj, ki je občutljiv na tip: »Sicer si je lahko predstavljati kak stroj, ki je tako zgrajen, da izgovarja določene besede, in celo to, da izreka marsikaj, kar povsem ustreza prisotnosti predmetov, ki delujejo na njegove zunanje organe. Če se na primer dotaknemo določenega mesta, bo stroj vprašal, kaj hočemo od njega, če drugega, bo zavpil, da mu prizadevamo bolečino, in podobno« (Descartes, 1957: 82). Ni pa dopuščal tako imenovanega *mehanskega mišljenja*, saj je bil prepričan, da noben stroj ni sposoben racionalno manipulirati s simboli, kar pa je vsaj po simbolni hipotezi pogoj za to, da lahko osebi ali sistemu pripišemo mišljenje: »Toda stroj ne more iz lastnega nagiba nikoli tako postavljati besed, da bi smiselno odgovarjal na vse, kar se izreče vpričo njega, kot zmorejo to katerikoli ljudje, pa naj bodo še taki topoglavci« (Descartes, 1957: 82). To pa je seveda v nasprotju z željami raziskovalcev s področja umetne inteligence, ki si, vsaj v njeni tradicionalni različici, prizadevajo ustvariti misleči stroj.

Alternativa opisanemu dualizmu je *monizem*, to je obstoj zgolj ene substance. Slednja je danes snovna/materialna, takšna teorija pa predpostavlja, da je vse, kar obstaja, zgrajeno iz fizičnih delcev, zato se tudi imenuje *fizikalizem*. Fizikalizem izgleda z vidika umetne inteligence precej bolj obetavno kot dualizem substanc, vendar pojasnitev mišljenja, ki je v Hobbesovih pojmih z racionalnimi pravili skladna manipulacija simbolov znotraj integriranega sistema, v prvenstveno fizičnem svetu vodi do paradoksa. Ta izhaja iz tega, da omenjena definicija mišljenja zahteva obstoj manipulatorja, ki predvidene operacije izvaja, ime, ki se je zanj udomačilo, pa je *homunkulus* – človeček. Paradoks lahko zapišemo v obliki konstruktivne dileme: ali manipulator upošteva, kaj simboli

in pravila pomenijo, ali pa ne. Če njihov pomen upošteva, potem manipulacija ne more biti povsem mehanska, saj je pomen vsaj deloma mentalni pojav. Če pa njihovega pomena ne upošteva, potem manipulacija ne more biti primerek mišljenja, saj je to, kaj nekaj pomeni, glavni kriterij za to, ali je nekaj miselno. Torej ali manipulacije niso mehanske ali pa niso mišljenje.[7] Zdi se, da v nobenem primeru ne dobimo tistega, kar želimo, to pa je umetna inteligenca. Ali lahko to dilemo zavrnemo?

Vzemimo njen prvi rog, imenujmo ga *miselni*, to je, da manipulator upošteva pomen simbolov. Četudi predpostavimo, da je pomen povsem fizični pojav, paradoksa mehanskega mišljenja ne rešimo. Zakaj? Ker predpostavka o inteligentnem človečku, ki bere simbole in ugotavlja, kaj pomenijo (*vnosi*) ter se glede na to odloča, katera pravila razuma bo uporabil, da bo ustvaril nove (mentalne) simbole (*iznosi*), to je na primer tisto, kar bomo mislili v naslednjem trenutku ali katero odločitev bomo glede nečesa sprejeli, (najverjetneje) predpostavlja, da ima tudi sam v glavi človečka, ki zanj počne natanko iste stvari, kot jih on za nas. In tako v nedogled. Tako mi nismo nikoli manipulatorji naših misli. Še več, lastnih misli sami tudi razumeti ne moremo. Naše misli so simboli, ki jih je zmanipuliral in razumel naš homunkulus.[8] Miselni rog dileme nam tako ne prinese rešitve. Kaj pa njen drugi rog, imenujmo ga *mehanski*?

Hume je bil tisti, ki je v sredini 18. stol. skušal na ta način izraziti računsko naravo mišljenja. Za razliko od Hobbesa, ki trdi, da *so* ideje fizični delci, kar zahteva obstoj inteligentnega homunkulusa, ki iz ozadja vleče prave poteze, pravi Hume zgolj, da so ideje *kot* fizični delci, katerih interakcije so razložljive v pojmih naravnih moči, ki so podrejene naravnim zakonom. Res je, da se s tem izogne manipulatorju simbolov – saj na vprašanje, kako ideje ubogajo naravne zakone, ne rabi odgovarjati, reče lahko preprosto, da pač jih. Kljub temu, da ne predpostavlja človečka, ki bi ga s tem, ko bi zanj opravljal manipulacije, oropal mišljenja, to rešitve paradoksa mehanskega mišljenja še ne prinese. Zakaj? Ker mora sedaj predstaviti teorijo o tem, kaj je tisto, kar upošteva pomen simbolov in kar sledi pravilom razuma, kar v končni fazi naredi ideje in interakcije med njimi za misleče, če to ni manipulacija s simboli. Zdi se, da Hume zgodbe, ki bi pomen in racionalnost potem, ko ju je izločil iz svoje ontološke slike, spet pripeljala nazaj, nima. Pri njem preprosto ne vemo, kaj je tisto, kar je v zvezi z duhom mentalno (*glej* Haugeland, 1986: 43–44). Tako nam tudi mehanski rog dileme ne prinese rešitve.

Paradoks mehanskega mišljenja, ki grozi s tem, da v primeru, ko je pomen pomemben za manipulacijo, procesi niso izkjučno mehanski (vključujejo homunkulusa, ki opravlja racionalne operacije), v primeru, ko ni pomemben, pa procesi v resnici niso racionalni (so le nesmiselne, strojem podobne interakcije) se zdi z zgodovinskega vidika neosvojljiva trdnjava.

[7] 1. Manipulator upošteva, kaj simboli in pravila pomenijo ali ne pomenijo.
2. Če njihov pomen upošteva, potem manipulacija ne more biti povsem mehanska.
3. Če njihovega pomena ne upošteva, potem manipulacija ne more biti primerek mišljenja.
4. Manipulacije niso mehanske ali pa niso mišljenje (*glej* Haugeland, 1986: 39).

[8] Tudi on ni manipulator svojih misli. Njegove misli so simboli, ki jih je zanj zmanipuliral in razumel njegov homunkulus itn.

## II Kaj je računalnik?

Poleg paradoksa mehanskega mišljenja in problema interakcije med nesnovnim duhom in snovnim telesom, pa je obstajala še ena dilema, ki je zavirala širitev ideje o umetni inteligenci: imenujemo jo lahko *subjektivnost misli*. Ta preprečuje, da bi lahko misli objektivno opazovali in opisovali. Tako se je vsa psihologija skrčila zgolj na opazovanje obnašanja, iz katerega je sklepala, katere misli so mu botrovale. Drugače kot izpeljani vmesniki misli niso mogle sodelovati v procesu odločanja in delovanja. Govorimo o rojstvu *behaviorizma*, ki je pravo življenje opisoval izključno v ne-mentalnih pojmih in ki se je še v 30. letih 20. stol. zdel precej pri močeh, kasneje pa je na srečo na veljavi hitro izgubil. Njegova težava je ta, da nam opisovanje sveta v pojmih obnašanja nikoli ne da zanesljivega odgovora na to, katere ideje ali motivi so pravi vzroki za naša dejanja, saj jih vedno obstaja cela vrsta, ki bi za to prišli v poštev, mi pa nimamo orodja, s katerim bi izmed njih izbrali pravega. Omenjenih treh dilem, ki so več stoletij preprečevale, da bi prepoznali pravo vrednost Hobbesove ideje o mišljenju kot računanju, je bilo, če ne že prej, konec, ko se je pojavil v sredini 50. let 20. stol. prvi šahovski program.[9] Ko danes verjetno najmočnejši šahovski program Fritz 12[10] igra, sledi splošnim pravilom igre in pravilom izračunavanja, katera poteza je v danem položaju najboljša. To pa ni nič drugega kot uresničitev Hobbesove definicije mišljenja v umetnem sistemu.[11]

Videli smo, da so že filozofi 17. in 18. stol. prepoznali temeljno vprašanje, ki se z vidika umetne inteligence glasi, *kako je možno, da v mehanskem svetu obstaja mišljenje* (vprašanje je bilo sicer bolj splošno in se je nanašalo na fizični svet in človeka), niso pa nanj znali odgovoriti. Stvari so se spremenile šele čez stoletja s prihodom računalnika. Kaj je torej računalnik? Računalnik je *formalni sistem*, to je funkcionalna celota med seboj načrtno povezanih odvisnih elementov, ki deluje na predpisan in ustaljen način. Igra je, recimo, primer formalnega sistema, kjer gre za to, da na primer figure postavljamo skladno s pravili tako, da na igralni plošči dobimo želen položaj. Takšni formalni sistemi so vedno sestavljeni iz treh delov: (i) *označevalcev*, s katerimi se igra, (ii) *manipulacije*, ki je prestavljanje, spreminjanje, dodajanje ali odvzemanje figur, in sicer skladno s pravili igre, kjer je določeno, kaj so figure, kakšen je njihov začetni položaj in kako se premikajo ter (iii) *digitalnosti*, ki pomeni, prvič, da je igra *samozadostna*, to je, da na njen potek vplivajo zgolj njene figure, poteze in položaji, ne pa tudi zunanji svet, na primer to, ali je šahovska deska moja ali ukradena, drugič, *določena*, to je, da ne dopušča dvomov o tem, ali sta neka poteza ali položaj zakonita ali ne, na primer konj je na polju e4 ali ne, mejnih primerov, da bi bil nekje vmes, ni, in, tretjič, "*preverljiva*, to je, da je v vsakem

[9] Leta 1957 je bil napisan prvi povsem delujoči šahovski program.

[10] Fritz je nemški šahovski program, katerega avtorja sta Frans Morsch in Mathias Feist, prva izdaja pa je luč sveta ugledala v zgodnjih 80. letih 20. stol.

[11] Računalnik se npr. izogne paradoksu mehanskega mišljenja, saj ne sledi *naravnim zakonom*, kar je bilo usodno za Humov princip »mislečega stroja«, ki je tako izgubil mentalnost, prav tako pa ne predpostavlja nobenega notranjega manipulatorja, ki bi zanj opravljal zahtevane racionalne operacije in ga s tem oropal mišljenja.

položaju število stvari, ki jih moramo pregledati za to, da bi ugotovili, ali je izvedena poteza dovoljena ali ne, končno (*glej* Haugeland, 1987: 7). Digitalnost je pomembna zaradi tega, ker lahko z njeno pomočjo ugotovimo, da sta dva formalna sistema, ki izgledata povsem različna, v bistvu enaka, to je delujeta po opisanih treh načelih digitalnosti (na primer človeški možgani in računalniški procesor), kar lahko predstavlja motivacijo za pritrdilni odgovor na vprašanje, ali nismo tudi ljudje zgolj neke vrste stroji.

Še več, računalnik je *samodejni* formalni sistem, ki je sposoben *interpretirati* simbole.[12] To pomeni, prvič, da je sposoben prepoznati označevalce in na njih samodejno izvajati manipulacije, kar vključuje tudi »sodnika«, ki skrbi za to, da poteka igra po pravilih, določa, kdo je na potezi in s katero figuro naj igra ter razglasi rezultat, in drugič, da je sposoben ugotoviti, kaj označevalci pomenijo. To, kar pomenijo, se nanaša na to, kaj izražajo ali predstavljajo v vsakdanjem jeziku. Interpretacija ima zato dva koraka, najprej računalnik ugotovi, kaj pomeni vsak enostavni simbol (na primer besede), potem pa, kaj pomeni vsak sestavljeni simbol (na primer stavki). Pomeni prvih so poljubni, pomeni drugih pa so določeni z njihovo zgradbo. Naloga interpretacije je tako preprosto, da simbole iz enega sistema, ki ni razumljiv, »prevede« v drugega, ki je razumljiv. Če sedaj lastnosti, ki določajo, kateri organizem je računalnik, strnemo, dobimo naslednje: (i) sodnik in vsaj nekateri igralci so avtomatizirani, njihova naloga pa je, da vedno prevedejo označevalce glede na dana pravila in (ii) označevalci so simboli, ki so interpretirani tako, da imajo v kontekstu, v katerem nastopajo, smisel. Velja pravilo, da če računalnik poskrbi za sintakso, to je, da se drži specificiranih formalnih pravil, s tem hkrati poskrbi tudi za semantiko, to je, iznosi (outputi) imajo v standardnem jeziku smisel.

»Če stroj sledi znotraj sistema definiranim sintaktičnim pravilom pri tvorjenju novih formul, potem bodo interpretirane formule ohranjale svojo semantično vrednost« (Markič, 1997a: 43–44). Vzemimo na primep kalkulator[13] in vtipkajmo *5 + 5 =*, pri čemer vsak simbol predstavlja potezo, vsi skupaj pa položaj, v katerem je. Ko končamo, bo kalkulator samodejno na ekranu izpisal *10*, kar bo glede na dana pravila in položaj, v katerem se nahaja, edina prava/zakonita/dovoljena poteza.

Če sedaj zaključimo iskanje definicije računalnika, lahko poenostavljeno rečemo, da je to *stroj, ki manipulira s simboli*.[14] Te simbole imenujemo *biti* – 0 in 1 – in označujejo besede ter števila, zapisani pa so v vrstah, to je registrih. Manipulacije, ki jih računalnik opravlja so na primer kopiranje, brisanje ali primerjava registrov itn. Ko vtipkam »a«, električno omrežje v računalniku avtomatsko proizvede na primer bitni zapis 1000001 (ti so vedno dolgi sedem znakov), kar je po konvenciji računalnikova notranja koda za to črko, ko vtipkam »mačka« pa na primer 1000011 1000001 1000111 1001111 1000001. Za stavek »mačka sedi na strehi« računalnik proizvede 20 sedemmestnih bitnih zapisov, 17 za črke in 3 za presledke. Če dodam na koncu piko se vsota poveča

[12] Obstajajo tudi ročni formalni sistemi, šah je eden izmed njih, kjer nekdo od zunaj – igralec – premika figure po šahovnici in ustvarja nove položaje.

[13] Kalkulator je sicer preveč enostaven primer, ampak za občutek, kako računalniki delujejo, zadošča.

[14] Ali »računalnik je interpretacijski samodejni formalni sistem« (Haugeland, 1986: 4).

še za enega. Ljudje manipuliramo z besedami in stavki, ki so simbolni opis predmetov in dogodkov v našem svetu, računalnik pa stvarnost opisuje z bitnimi zapisi. In če za človeka pravimo, da je sposoben načrtovati, analizirati, primerjati, ocenjevati, kritizirati in se odločati, moramo po analogiji iste lastnosti pripisati tudi računalniku. To imenujemo preprosto *simbolna hipoteza*, izhaja pa iz tega, da tudi naš silni duh ni nič več kot univerzalni sistem simbolov, zaradi česar je tudi človeško mišljenje zgolj manipulacija z njimi.

Res je sicer, da »računalniki« znotraj naših glav izvajajo drugačne operacije kot programje v stroju in da tudi simbolni zapis, ki ga uporabljajo, ni binaren, vendar to še zdaleč nista dovolj dobra razloga za to, da se človek in računalnik ne moreta enako obnašati. Simbolna hipoteza je problematična, saj ni jasno, kako bi lahko naprava, ki je zmožna zgolj premetavati simbole, razumela stavke kot na primer »mačka sedi na strehi«, kar pa se nam v primeru človeka niti malo ne zdi sporno.[15] Če se sedaj vrnemo na temeljno filozofsko vprašanje 17. in 18. stol., ki je iz vidika umetne inteligence, kako je možno, da v mehanskem svetu obstaja mišljenje, je odgovor, ker obstaja umetni mehanizem, to je samodejni formalni sistem – računalnik, ki je zmožen interpretirati simbole, to je, jih tolmačiti tako, da imajo v vsakdanjem jeziku smisel.

»Pravi« računalnik, če izvzamemo kalkulatorje, kjer je prvi javno znani nastal okrog 1640, njegov avtor pa je bil Pascal (znal je in to precej nerodno seštevati in odštevati, množenje in deljenje pa mu je okrog 1670 dodal šele Leibniz), se je pojavil med leti 1823 in 1833. Izdelal ga je Anglež Charles Babbage, imenoval pa se je *Analitični stroj*. Sestavljen je iz treh delov: *mlina* (aritmetična enota), *shrambe* (spominska enota) in *kontrolne enote*. Figure so števila, kraj dogajanja pa je shramba. Mlin lahko izvaja štiri osnovne aritmetične operacije, kontrolna enota pa je »sodnik«, ki ne manipulira s števili, ampak zgolj ukazuje, s katerimi števili se sešteva, odšteva, množi ali deli ter kam se zapisujejo odgovori. Še več, Analitični stroj lahko znotraj svojih mej avtomatsko izračuna katerokoli matematično formulo, na primer 5(x +y) – 7 (pri navadnem kalkulatorju je treba vsak korak opraviti ročno), če smo mu to prej v pravilih, po katerih naj se ravna (najprej vstavi dana x in y, potem njuno vsoto pomnoži s 5 in na koncu od zmnožka odštej 7), na način, ki ga »razume«, določili. Spreminjanje pravil, ki Analitični stroj preobrazi v katerikoli samodejni formalni sistem na svetu, pa je že *programiranje*.[16] Tako je bil Charles Babbage prvi, ki je »naredil« sprogramirani računalnik, pri čemer je treba »naredil« vzeti z rezervo: ker je bil njegov stroj prevelik in prezapleten, zaradi česar ga ni skoraj nihče razumel, kaj šele bil sposoben zgraditi, ni v praksi nikoli zaživel (*glej* Haugeland, 1986: 125–130).

Drugi znani stroj, ki ga je doletela enaka usoda, čeprav iz drugega razloga – že v osnovi je bil mišljen zgolj kot enostavni abstraktni sistem za dokazovanje teo-

[15] Ker naš namen ni problematizacija simbolne hipoteze v smislu rušenja analogije med človekom in računalnikom, v nadaljevanju članka o tem ne bomo več govorili. Za to *glej* na primer Searle, 1990.

[16] Turing recimo pozneje pravi, da lahko takšne računalnike, ki imajo »posebno lastnost …, da lahko posnemajo kateri koli diskretni stroj, opišemo tako, da rečemo, da so *univerzalni* stroji« (*Turing*, 1990: 64).

retičnih predpostavk brez praktične vrednosti – si je leta 1936 izmislil Alan Turing.[17] Znan je pod imenom *Turingov stroj*, sestavljen pa iz *glave* in *traku*. Trak je razdeljen na neskončno število kvadratov, izmed katerih so eni vedno zasedeni s črko iz predpisane abecede, drugi pa prazni. Glava se pomika preko kvadratov, skenira njihove črke, jih po potrebi briše in nadomešča z novimi. Kje se ustavi in kaj s kvadratom stori, je vnaprej določeno s pravili, po katerih deluje. Kraj dogajanja je trak, ki predstavlja shrambo podatkov, črke v kvadratih pa figure, s katerimi se manipulira. Različna notranja stanja glave predstavljajo operacije, ki se izvajajo – skeniranje, brisanje, nadomeščanje – sodnik pa zgolj začne igro in vleče trak, dokler ga naloga, ki jo mora glava na danem kvadratu opraviti, ne prisili, da ga zadrži. Tudi Turingovi stroji so samodejni formalni sistemi, nekatere izmed njih pa se da tudi sprogramirati tako, da posnemajo druge: stroj *x* preprosto razvozla opis nalog stroja *y* in jih zapiše na svoj trak, potem pa *x* skladno s tem obravnava posamezne kvadrate enako, kot bi jih obravnaval *y*, posledica česar so enake poteze *x*-a in *y*-a. Skratka, dešifrirani opis delovanja prvega stroja uporabimo kot program za delovanje drugega (*glej* Haugeland, 1986: 133–138).

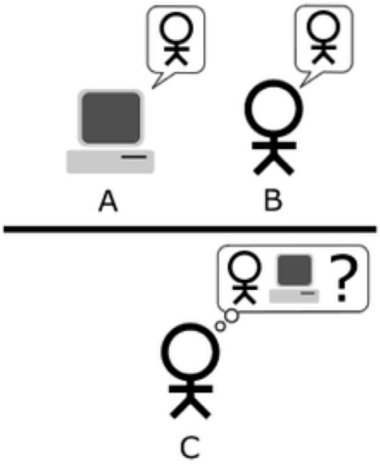
Prvi sprogramirani računalnik, ki ni bil zgolj abstraktna tvorba na listu papirja in to, ali v obliki neuresničenih načrtov ali globokumnega teoretičnega eksperimenta, pa je v Nemčiji leta 1941 izdelal Konrad Zuse.V okviru svojih spominskih zmožnosti je bil sposoben opraviti katerokoli računsko operacijo, zanimivo pa je, da na kasnejši razvoj računalnikov ni imel nobenega vpliva, saj zanj takrat, ko je nastal, ni vedel skoraj nihče. Malo pozneje, leta 1943, so tudi Britanci razvili svoj prvi sprogramirani računalnik, ki se je imenoval *Kolos* (velikan). Namenjen je bil odkodiranju nemških sporočil, program zanj – pravila, po katerih je informacije, to je črke in besede, na smiselni način pretvoril v vsakdanji jezik – pa je priskrbela takrat v Bletchley Parku delujoča skupina dešifratorjev. Žal ni znal ničesar drugega kot razvozlavati na prvi pogled nesmiselna nemška sporočila o premikih njihovih čet itn., množenje na primer je že presegalo njegove zmožnosti, zaradi česar je vprašljivo, ali ga sploh lahko uvrstimo med računalnike. Po 2. sv. vojni so Britanci v Manchestru leta 1948 na noge postavili prvi povsem elektronski sprogramirani računalnik, z imenom *Mark I*. Ta za svoje delovanje ni uporabljal več relejev, ampak elektronke, danes bi rekli tranzistorje – polprevodnike in je prvi, ki je leta 1951 stopil na pot »množične« proizvodnje, prodali so vseh 9 primerkov. Kasneje so prevladujoč položaj pri proizvodnji računalnikov v komercialne namene prevzela ZDA, ki so svoj *ENIAC* sicer predstavile že leta 1945, vendar ta z vidika uporabnosti *Marku I* ni segal niti do gležnjev (*glej* Copeland, 1993: 1–5).

V 50. letih 20. stol. se je utrdilo prepričanje, da je kriterij za pripisovanja inteligence zmožnost logičnega mišljenja, česar bi moral biti tega sposoben tudi stroj, če bi želel veljati za »pametnega«. Osnovo za to je leta 1956 postavil John von Neumann s tem, ko je razvil računalnik – hardwer – z imenom *Johnniac*, na katerega so potem Newell, Shaw in Simon namestili program – programje – z nazivom *Logični teoretik*. Ta kombinacija je takrat po mnenju mnogih pred-

[17] Zanj *glej* Turing, 1936.

stavljala prvega resnega kandidata za to, da bi lahko govorili o prvem inteligentnem stroju, avtorji pa z imenom, ki so ga programu nadeli, teh ambicij niso skrivali. Da bi *Logični teoretik* to dosegel, bi moral biti sposoben dokazati logične teoreme tipa $p \rightarrow (p \ V \ q)$.[18] Tega lahko na primer dokažemo tako, da predpostavimo *p*, to je antecedens pogojnika, za pogojni dokaz – pri katerem je pogojnik dokazan, kakor hitro nam uspe izpeljati njegov konsekvens – z adicijo dobimo $p \ V \ q$, kar je takoj konsekvens pogojnika, s čimer smo gornji teorem že dokazali. To mu je s tem, ko je dokazal prvih 38 teoremov drugega poglavja iz Whiteheadove in Russlove *Prinicpia Mathematica*,[19] ki velja za temeljno delo s področja logike in matematike, tudi uspelo. Prvič se je zgodilo, da stroj ni zgolj računal, ampak tudi dokazoval, pa ne samo to, njegov dokaz enega izmed teoremov je bil elegantnejši od dokaza obeh avtorjev. Pisci programa so potem na to temo napisali članek, kjer so kot soavtorja navedli tudi računalnik, kar je bil prvi primer takšnega soavtorstva v zgodovini, vendar članka uredniki revije *The Journal of Symbolic Logic* žal niso želeli objaviti (*glej* Copeland, 1993: 7). *Logični teoretik* velja danes za prvi inteligentni program.[20]

## III Turingov test




Turing pa ni znan samo kot izumitelj enega izmed abstraktnih računalnikov, ampak tudi kot avtor posebnega preizkusa, ki si ga je izmislil leta 1950[21] in ki naj bi pokazal, ali stroj lahko misli ali ne. Imenuje se *Turingov test,* zasnovan pa je na tako imenovani igri oponašanja, v kateri v standardni različici preizkusa nastopajo trije akterji, dva človeka in računalnik. Eden izmed ljudi je spraševalec, ki mora ugotoviti, kateri izmed preostalih dveh igralcev, ki odgovarjata na njegova vprašanja, je računalnik. Vsi trije so v ločenih sobah, komunikacija pa danes poteka namesto preko teleprinterja, kot je predvidel Turing, preko ekrana in tipkovnice. To lepo ponazarja naslednja posodobljena shema iz Saygin, A. P., Cicekli, I., Akman, V., 2000: 467:

[18] *Glej* Newell, A., Shaw, J. C., Simon, H. A., 1963.

[19] *Glej* Whitehead, A. N., Russell, B., 1997.

[20] S tem se sicer vsi ne strinjajo, saj naj bi že pred njim obstajali programi, ki so bili inteligentni. Najbolj znan in uspešen je bil *»program za damo«* [v *angl. the checkers program*], ki ga je že na začetku 50. let 20. stol. razvil Arthur Samuel, deloval pa je na IBM-ovem računalniku 701 iz tistega časa. (Salonske igre na deski se prav tako pogosto jemljejo kot preizkusni kamen inteligence stroja). Vključeval je mehanizem za učenje, kar mu je omogočalo napredovanje in na koncu vodilo celo do tega, da je svojega avtorja premagal.

[21] V Turing, 1950.

> Spraševalec lahko postavlja kakršnakoli vprašanja – od enostavnih in površinskih, do zapletenih in globokih – za to, da bi ugotovil, v kateri sobi se nahaja stroj, računalnik pa lahko stori karkoli, da bi s svojimi odgovori spraševalca prevaral in izsilil napačno identifikacijo: če ga spraševalec vpraša, ali je stroj, lahko odgovori z ne, če od njega zahteva množenje večmestnih števil, lahko odgovori po daljšem premoru in napačno, če ga vpraša, ali je že bil zaljubljen, lahko odgovori z da itn. Za to, da bi stroj preizkus opravil, je »najboljša strategija ta, da poskusi odgovarjati tako, kot bi naravno odgovarjal človek« (Turing, 1990: 63). Turingovi primeri vprašanj (*V*) in odgovorov (*O*) so naslednji:
>
> *V*: Napišite mi prosim sonet na temo mostu Forth.
>
> *O*: S tem ne bo nič; nikoli nisem mogel pisati poezije.
>
> *V:* Seštejte 34957 in 70764.
>
> *O*: (počaka približno 30 sekund in nato odgovori) 105621.
>
> *V*: Ali igrate šah?
>
> *O*: Da.
>
> *V*: Imam kralja na e1, in nobenih drugih figur. Vi imate le kralja na e6 in trdnjavo na h8. Vi ste na potezi. Kaj boste igrali?
>
> *O*: (po pavzi 15 sekund) Trdnjava na h1 mat. (Turing, 1990: 62–63).

Človek mora na vprašanja odgovarjati po resnici. Test se ponovi večkrat, pri čemer se človeški igralci vedno menjajo. Stroj ga opravi, če na koncu spraševalec v več kot 50 % primerov računalnik zamenja za človeka. V tem primeru moramo tudi stroju, če želimo biti dosledni, to je tudi zanj uporabljati iste vatle kot za človeka, pripisati mišljenje. »Če vem, da se objekt v vseh pogledih obnaša tako, kot se po definiciji obnaša zavestno bitje, potem vem, da ima v resnici zavest« (Ayer, 1978: 172).

Omeniti velja, da je ideja o jeziku kot znamenju razumevanja že stara, saj jo prvič najdemo že v 17. stol. pri Descartesu, ki je predlagal konverzacijo kot sredstvo zanesljivega ločevanja med mislečim bitjem in še tako sofisticiranim strojem.[22] Prav tako ni nova ideja o obnašanju kot znamenju inteligence, opisoval jo je že Ayer, ko je leta 1936 v *Language, Truth and Logic* trdil, da »… se razlika med zavestnim človekom in nezavednim strojem skrči na razliko v njunem obnašanju« (Ayer, 1978: 172). V nadaljevanju je predlagal empirični test kot sredstvo razlikovanja med zavestnimi in nezavednimi bitji: »Edini razlog, ki bi ga lahko imel za to, da objekt, za katerega se zdi, da ima zavest, lete v resnici nima in je tako zgolj lutka ali stroj, je, da ni opravil nobenega izmed empiričnih testov, s katerimi se določa prisotnost ali odsotnost zavesti« (Ayer, 1978: 172).

Sklenimo to poglavje s tem, kateri stroji naj bi bili po Turingovem mnenju kandidati za to, da prestanejo njegov test. Idejo o tem, da bi za izdelavo mislečega stroja uporabili biološko tehniko in ga vzgojili iz ene same kožne celice, zavrne kot neustrezno, pa čeprav bi to bilo veliko zmagoslavje znanosti. Še več, pripravljen je priznati, da pride za to v poštev samo en način izdelave: »… sedan-

[22] Glej Descartesov citat na to temo v prvem poglavju.

je zanimanje za 'misleče stroje' je vzbudila posebna vrsta stroja, običajno imenovanega 'elektronski računalnik' ali 'digitalni računalnik'. Sledeč tej sugestiji bomo le digitalnim računalnikom dovolili sodelovanje v igri …« (Turing, 1990: 64).

V drugem poglavju smo omenili, da imenuje digitalne računalnike univerzalni stroji,[23] o njih pa pravi naslednje: »Obstoj strojev s to lastnostjo ima pomembno posledico, da, če pustimo ob strani vprašanja hitrosti, ni potrebno načrtovati različnih novih strojev za opravljanje raznih procesov računanja. Vse je mogoče opraviti z enim digitalnim računalnikom, ustrezno programiranim za vsak posamezen primer« (Turing, 1990: 64).

Turing je bil glede razvoja umetne inteligence tako optimističen, da je trdil, da bo čez 50 let imel povprečen spraševalec v igri oponašanja 70% možnosti za pravilno identifikacijo ter da ideji o mislečih strojih ob koncu 20. stol. nihče več ne bo nasprotoval. Danes vidimo, kako se je motil.

## IV Ugovori proti Turingovem testu

Vprašanje je, ali je Turingov test res merilo za to, da lahko nekemu živemu ali neživemu organizmu pripišemo mišljenje.[24] Odgovore, ki ga zavračajo, lahko zaradi različnih razlogov, ki jih za to ponujajo, razdelimo na štiri dele. Na koncu vsakega izmed njih bomo predstavili odgovor, ki bi ga lahko ponudil zagovornik Turingovega testa.[25]

*1. Ugovor iz šimpanzov*

Preizkus je preozek, saj ga na primer tudi šimpanzi, delfini, psi itn. ne bi opravili, pa jim kljub temu pripisujemo določene kognitivne sposobnosti. Enako bi se lahko zgodilo stroju, pa to še ne bi pomenilo, da ne misli. Lahko bi bil psihološko premalo kompleksen ali pa bi bili njegovi odgovori drugačni, kot jih pričakujemo od človeka. Ker negativen rezultat tako ne pokaže ničesar, je test neuporaben.

Zagovornik preizkusa bi lahko na to odgovoril, da kljub temu, da negativen rezultat na testu ne pokaže ničesar dokončnega, to še ne pomeni, da je test slab in da ga moramo zavreči.

Enako namreč velja za številne druge po mnenju večine koristnih testov, na primer za test s prstnimi odtisi (če na kraju zločina ni mojih prstnih odtisov, to še ne pomeni, da me tam ni bilo), ki jih kljub nedokončnemu odgovoru, še vedno izvajamo.

*2. Ugovor iz čutil*

Test se omejuje zgolj na strojevo sposobnost dajanja verbalnih odgovorov, ne pa tudi na njegovo zmožnost povezati besede s stvarmi, ki jih označujejo. Računalnik bi tako lahko opravil preizkus, ne da bi poznal pomen besed, ki jih

[23] *Glej* 16. op.

[24] Turing sam je zbral devet najbolj verjetnih ugovorov takratnega časa proti temu, da lahko stroj misli in jih vse po vrsti gladko zavrnil (*glej* Turing, 1990).

[25] Za predstavitev ugovorov in odgovorov nanje *glej* Copeland, 1993.

uporablja, zaradi česar pozitiven test ne bi bil zanesljiv pokazatelj tega, da stroj lahko misli. Opremiti bi ga morali z umetnimi čutili, na primer kamero, da bi spraševalec lahko preveril, ali stroj resnično razume besede, ki jih uporablja, na primer, če mu pokažemo svinčnik, mora odgovoriti, da imamo v roki svinčnik itn.

Zagovornik testa bi lahko na to odvrnil, da se številne besede v vsakdanjem življenju nanašajo na stvari, ki jih ne moremo videti ali se jih dotakniti, na primer podatek, inteligenca, filozofija, sreča, tragedija itn., zaradi česar je edini način, da ugotovimo, ali jih stroj razume, zastavljanje sofisticiranih vprašanj in preverjanje njegovih odgovorov nanje. Ugovor je tako neprepričljiv, saj ne drži, da je edini način za preverjanje razumevanja besed, ki jih računalnik uporablja, čutna interakcija z okoljem. Še več, Turingova navodila za izvedbo testa ne prepovedujejo tega, da bi stroj opremili s kamero, kar bi zajelo tudi razumevanje tistih besed, za katere jo rabimo.

*3. Ugovor iz črne škatle*

Pozitiven test, ki naj bi dokazoval, da računalnik lahko misli, temelji zgolj na njegovem obnašanju, kaj se dogaja znotraj njega, pa ostaja skrivnost.[26] To pa je v primerjavi z ljudmi pomanjkljivost, saj ko jim na osnovi obnašanja pripisujemo mentalna stanja, vemo, kakšna je njihova fizična zgradba in kako delujejo njihovi možgani, ko rečejo, da so žejni, vemo natančno, kateri elektrokemični procesi so za to odgovorni in katerim nevrofiziološkim mehanizmom so podvrženi. Delovanje računalnika pa nam je uganka, saj ima drugačno telo kot mi, zaradi česar mu nismo zgolj na podlagi enakega obnašanja takoj pripravljeni pripisati mišljenja.

Na primer, če bi se s Stalinom peljala v avtu in na Vršiču doživela prometno nesrečo, v kateri bi bil sam nepoškodovan, on pa bi »izkrvavel«, pri čemer bi naknadno ugotovil, da so njegove žile plastične, iz njih pa teče kri, ampak olje, in da iz njegovih ran visi cel kup žic in druge železnine, ali ne bi potem spremenili mnenja, da je bil Stalin misleča stvar, na kar vas je prvotno, ko je bil še živ, napeljala opisana konverzacija, ki sem jo imel z njim? Večina ljudi bi na to vprašanje zanesljivo odgovorila z da. Tako je pripisovanje mišljenja zgolj na osnovi obnašanja napačno.

Zagovornik preizkusa bi na to lahko odgovoril, da ga lahko posodobimo tako, da se upošteva ne samo obnašanje računalnika, ampak tudi njegova zgradba, kar se lahko najbolj enostavno naredi tako, da se poleg njegovih odgovorov pogleda tudi njegovo delovanje. Tako bi moral stroj za to, da bi prestal test in bil misleča stvar, poleg tako imenovanega *behavioristič*nega kriterija (ustrezni odgovori na vprašanja) izpolniti še tako imenovani *oblikovni* kriterij (raziskava sestave in funkcioniranja računalnika).

Vprašanje, ki se tukaj zastavi, je, kako pa naj bi bil oblikovan? Najmočnejša zahteva je, da bi delal stvari na podoben način, kot jih delajo možgani v naših lobanjah, to je njegovo delovanje bi potekalo na podoben način kot možgansko delovanje. To ne pomeni, da bi moral biti narejen iz enakih materialov, saj je pomembno zgolj to, da podobno funkcionirata. Za test to ne predstavlja no-

[26] Od tod ime črna škatla, s katero ta misterij opisujemo.

benega problema, saj bi ga iz antropomorfizma glede obnašanja preprosto razširili na antropomorfizem glede zgradbe.

*4. Ugovor iz simulacije*

Če računalnik opravi preizkus, to kaže zgolj na to, da je sposoben *simulirati* mislečo stvar, ne pomeni pa, da dejansko *je* misleča stvar, kot na primer simulacija mleka še ni dejansko mleko. Če bi v igri oponašanja namesto računalnika in človeka, spraševalec postavljal vprašanja moškemu in ženski in bi moškega napačno identificiral kot žensko, to še ne bi pomenilo, da moški dejansko je ženska, ampak zgolj, da je sposoben simulirati žensko. Pozitiven test je tako kot merilo za pripisovanje mišljenja neuporaben.

Zagovornik testa bi lahko na to odvrnil, da ugovor meša dva pomena besede »simulacija«. Ta lahko pomeni, da stvari, ki nekaj simulira, manjka vsaj ena bistvena lastnost, ki jo dejanska stvar ima, na primer ponarejenemu diamantu manjka prava molekulska zgradba, oseba, ki se dela mrtvo, je še vedno živa itn. Imenujmo to *simulacija*$_1$. Lahko pa pomeni, da stvar, ki nekaj simulira, ima vse bistvene lastnosti, ki jih ima dejanska stvar, razlika je samo v tem, da za razliko od dejanske ni bila proizvedena na naraven, ampak na umeten način v laboratorijskih pogojih, na primer umetni diamant ima vse lastnosti naravnega: molekulsko zgradbo, trdoto, prozornost, bleščavo, toplotno prevodnost itn., umetni vitamin C ima vse lastnosti naravnega itn. Imenujmo to s*imualcija*$_2$. Zdi se, da ugovor brez argumenta predpostavlja, da bo simulacija mišljenja, ki jo bo proizvedel računalnik vedno simulacija$_1$, to je zgolj ponaredek mišljenja, nikoli simulacija$_2$, to je prava stvar. Ker pa je bistvo vprašanja, ali računalnik lahko misli ravno to, ali je njegova simulacija mišljenja lahko kdajkoli simulacija$_2$, je nezanimiv, saj vnaprej določa odgovor.

Priznati moramo, da je ugovor iz simulacije v enem pogledu koristen, saj jasno pokaže, na kaj moramo odgovoriti, da dobimo inteligentni stroj. Vprašanje je torej, ali lahko zgolj na osnovi opravljenega Turingovega testa – ustreznih verbalnih odgovorov – kdajkoli ugotovimo, da je šlo za simulacijo$_2$, to je pravo, a umetno proizvedeno inteligenco, in ne za simulacijo$_1$, to je ponaredek inteligence, na osnovi česar lahko rečemo, da smo dobili misleči stroj. Slovenski filozof Danilo Šuster denimo pravi, da ne vemo, ali gre za simulacijo$_1$ ali za simulacijo$_2$. Vse, kar test odkrije, so le zunanje lastnosti računalnika, to pa še ne pomeni, da je hkrati s tem računalnik že demonstriral bistvene lastnosti mišljenja in da bi ga zaradi tega lahko imeli za inteligentnega. Res je, da to ne izključi možnosti, da stroj vendarle misli, pravi, to je, da gre za simulacijo$_2$, ampak to mora dokazati tisti, ki to trdi. Breme dokaza je torej na plečih zagovornika Turingovega testa, vprašanje pa je, ali mu je dorasel. Če mu ni, smo v položaju, ko nam opravljen preizkus ne da odgovora na to, ali smo dobili pameten računalnik. Tako dokler ne najdemo nečesa, kar pravo, a umetno proizvedeno mišljenje – simulacijo$_2$, loči od ponaredka mišljenja – simulacije$_1$, smo na istem kot prej, ne vemo, ali stroj lahko misli ali ne.

Zdi se mi, da je test možno dopolniti tako, da bomo lahko ločili pravo inteligenco od njenega ponaredka. Mislim, da se odgovor res skriva v ugovoru iz črne

škatle.[27] Če se potrudimo za to, da misterij o tem, kako je stroj zgrajen in kako deluje, izgine, dobimo tudi kriterij za ločevanje obeh simulacij. Vzemimo za primer človeka. Kdaj smo mu pripravljeni pripisati na primer bolečino, če čisto iz pragmatičnih razlogov in zavoljo argumenta z njo nadomestimo mišljenje. Odgovor na to se bržkone skriva v tem, ali izpolni dva naslednja pogoja ki smo ju omenili že pri ugovoru iz črne škatle. Prvi je, da se obnaša na javno preverljiv način in da je ta tipičen za bolečino, na primer, ko kriči in stoka. Ker pa lahko čuti bolečino, pa tega ne dela, ker je na primer superstoik ali superšpartanec, če si sposodimo izraz od Putnama, ali pa to počne, pa ne čuti bolečine, ker na primer uživa v spolnem odnosu, rabimo še nekaj drugega. Kaj?

Začnimo s tako imenovano elektroencefalografijo ali krajše EEG-jem, ki je snemanje električne aktivnosti na glavi, ki jo povzročajo hitro nihajoči nevroni možganov. Traja od 20 do 40 minut, merjenje pa poteka preko elektrod, ki so razporejene po celotnem lasišču. Ko nekdo odgovarja na različna vprašanja ali opisuje svoje občutke, na ekranu vidimo, kateri nevroni »gorijo«, iz česar lahko sklepamo, da so ravno ti odgovorni za mentalne procese ali stanja, v katerih se nahaja: to je lahko na primer mišljenje, ki je pripeljalo do pravilnega odgovora na dano vprašanje ali subjektivni občutek bolečine, v glavnem nekaj, kar je bilo prej opisano zgolj v behavioristični ali kakšnih drugih pojmih. Ko po primerjavi seznamov različnih ljudi, ugotovimo, da so pri vseh, ko so čutili bolečino, vibrirali isti nevroni in ko iz nevrokemičnega vidika razumemo delovanje naših možganov, to je mehanizmov, ki na nevrofiziološki ravni pripeljejo do subjektivnih mentalnih stanj, imamo na voljo njihov objektivni opis. Tako je prepričanje, da imamo vsi ljudje v glavah enak nevrofiziološki aparat in da njegovo delovanje vsaj v obrisih razumemo, drugi pogoj, ki mora biti izpolnjen za to, da človeku pripišemo na primer mišljenje ali bolečino. Sklenemo lahko, da mora človek za to, da bi čutil bolečino, zadostiti dvema kriterijema, *behavorističnemu* in *oblikovalskemu*.

Vrnimo se sedaj na stroj in na nalogo, ki smo si jo zadali, to je najti nekaj, kar pravo, a umetno proizvedeno mišljenje – simulacijo$_2$, loči od njegovega ponaredka – simulacije$_1$. Ugotovili smo, da stroj, ki opravi Turingov test, izpolni behavioristični pogoj, kar pa še ni dovolj, da bi to bilo pravo mišljenje. Če sedaj preizkus dopolnimo s tem, da mora stroj poleg tega izpolniti še oblikovalski pogoj, to je posedovanje razumljenega elektrokemičnega mehanizma, ki je enak za vse računalnike iste vrste in je odgovoren za proizvodnjo obnašanja, po analogiji s človekom za to, da bi ločili simulacijo$_2$ od simulacije$_1$ in s tem pravo, a umetno proizvedeno inteligenco, od njenega ponaredka, ni več nobenih ovir. Za to poskrbi oblikovalski kriterij. Ko stroj, ki opravlja Turingov test, zadosti tako behavioristični kot tudi oblikovalskemu pogoju, moramo, če želimo biti dosledni, tudi zanj reči, da misli.

Nasprotnik Turingovega testa bi lahko na to odvrnil, da ko je enkrat oblikovalski pogoj izpolnjen, behaviorističnega za to, da bi računalniku pripisali mišljenja, ne rabimo več, zaradi česar je Turingov test kot kriterij za ugotavljanje inte-

[27] Zdi se, da na to namiguje tudi *Copeland*, 1993: 47.

ligentnosti strojev nepotreben. Ideja je mamljiva, vendar znotraj funkcionalizma[28] kot vodilne filozofske teorije o mentalnem življenju stroja (in človeka), ni zadostna. Ta predpostavlja, da je npr. čutiti bolečino biti v takšnem fizičnem stanju, recimo *B*, ki je povzročeno s *tipičnimi vzroki*, tj. fizičnimi dražljaji (npr. praske) in drugimi mentalnimi stanji (npr. zavedanje, da sem jih staknil), in ki povzroča *tipične učinke*, tj. fizično obnašanje (npr. kričanje in stokanje) in druga mentalna stanja (npr. željo po prenehanju). To imenujemo *funkcionalna vloga* psihičnega stanja, katero fizično stanje jo izvrši, pa ni pomembno. Pričakujemo, da jo v različnih vrstah izvršijo različna notranja stanja, npr. *B* pri ljudeh, *P* pri ptičih in *K* pri krokodilih. To pa ne pomeni nič drugega kot, da bi v skrajnem primeru lahko imeli mentalno življenje tudi računalniki, saj bi ravno tako lahko izvršili predpisane funkcionalno vloge, ki označujejo posamezna mentalna stanja. Skratka, funkcionalizem trdi, da so psihični pojavi določeni z njihovimi vzročnimi relacijami med čutnimi dražljaji, drugimi psihičnimi pojavi in obnašanjem. Ker je tako bistven del funkcionalne vloge kateregakoli mentalnega stanja obnašanje, tega za to, da bi lahko umetnemu ali naravnemu sistemu pripisali npr. bolečino, ne moremo preprosto izločiti iz pojasnitev o mentalnem življenju stroja (in človeka).[29] Znotraj funkcionalizma bi lahko na oblikovalski in behavioristični kriterij gledali tako, da nam prvi priskrbi znanje o tem, katero notranje stanje je tisto, ki realizira funkcionalno vlogo in razumevanje fizičnega mehanizma, ki je do njega pripeljal, behavioristični pa poskrbi za znanje o tem, ali je zaznano obnašanje res tisto, ki ga v svojih vnosih in iznosih predpisuje funkcionalna vloga mentalnega stanja, za katero mislimo, da ga umetni ali naravni sistem ima.

Kakorkoli, tudi stroj, ki opravi tako dopolnjen Turingov test, česar do sedaj ni uspelo še nobenemu, še ni na koncu svoje poti do tega, da mu priznamo mišljenje. S filozofskega vidika je v najboljšem primeru čez prvo oviro.[30] Čaka ga še dolga pot, prav tako posuta s trnjem, kot je bilo dosedanja, končni uspeh pa je zavit v tančico skrivnosti. O njej bomo razpravljali drugič.

[28] Zanj *glej* na primer *Block*, 1980.

[29] Razen če smo se pripravljeni odreči funkcionalizmu in izbrati katero od danes manj priljubljenih alternativ.

[30] Z vidika umetne inteligence pa ne: »… Če stroj opravi Turingov test, mu moramo priznati inteligenco. Če ga pa ne, ne moremo zagotovo reči ali misli ali ne. To je verjetno najbolj običajna interpretacija Turingovega testa« (Saygin, A. P., Cicekli, I., Akman, V., 2000: 509).

## Can a machine think

The article discusses artificial intelligence as part of cognitive science arising from the fact that our minds work according to computational principles. The first part introduces visionists who has established as early as many centuries before that thinking is simply reckoning. The second part of the article represents formal conditions which a system has to meet in order to be called a computer. The third part describes the so called Turing Test, i.e. a test of a machine's ability to show its intelligence, which is in the artificial intelligence circles taken as a proof for the already existing age of thinking computers, a thesis widely disputed by philosophers. Four basic objections are discussed in the fourth part, and the most promising one called *the simulation objection* is being refuted by complementing the Turing Test and so keeping it alive.

*Keywords*: artificial intelligence, cognitive science, Turing Machine, Turing Test, classical AI, connectionism.

### Literatura

Ayer, A. (1978). *Language, Truth and Logic*. Harmondsworth: Penguin.

Block, N. (1980). »Introduction: What is Functionalism?«. V Block, N. (ur.), Readings in Philosophy of Psychology, Cambridge: Harvard University Press.

Boden, M. A. (1990). »Introduction«. V Boden, M. A. (ur.), *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press.

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

Descartes, R. (1957). *Razprava o metodi*. Ljubljana: Slovenska matica.

Descartes, R. (1985). *The Passions of the Soul*. V Cottingham, J., Stoothoff, R., Murdoch, D. (ur.), *The Philosophical Writtings of Descartes*. Cambridge: Cambridge University Press.

Haugeland, J. (1986). *Artificial Intelligence: The Very Idea*. Cambridge: The MIT Press.

Haugeland, J. (ur.) (1987). »Introduction«. V Haugeland, J. (ur.), *Mind Design*. Cambridge: The MIT Press.

Hobbes, T. (2006). *Leviathan (Revised Student Edition)*. Cambridge: Cambridge University Press.

Kim, J. (2001). *Mind in a Physical World*. Cambridge: MIT Press.

Miščević, N. (2001). »Uvod«. V Miščević, N., Smokrović, N. (ur.), *Računala, mozak i ljudski um*. Rijeka: Izdavački centar Rijeka.

Markič, O. (1995). »Kognitivizem in konekcionizem – dva pristopa v kognitivni znanosti«, *Časopis za kritiko znanosti*, 176, str. 101–105.

Markič, O. (1997a). »Klasična kognitivna znanost in simbolni modeli«, *Analiza 1*, 1999, str. 38–52.

Markič, O. (1997b). »Metafora in kognitivna znanost«, *Analiza*, 3–4, str. 72–73.

Newell, A., Shaw, J. C., Simon, H. A. (1963). »Empirical Explorations with the Logic Theory Machine: a case Study in Heuristics«. V Feigenbaum, E. A., Feldman, J. (ur.), *Computers and Thought*. New York: McGraw-Hill

Searle, J. (1990). »Duhovi, možgani in programi«. V Hofstadter, D. R., Dennet, D. C. (ur.). *Oko duha*, Ljubljana: Mladinska knjiga.

Saygin, A. P., Cicekli, I., Akman, V. (2000). »Turing Test: 50 Years Later«. *Minds and Machines*, 10, str. 463–518.

Turing, A. (1936). »On Computable Numbers, with an Application to the *Entscheidungsproblem*«, *Proceedings of the London Mathematical Society*, 42, str. 230–265.

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

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

Whitehead, A. N., Russell, B. (1997). *Principia Mathematica*. Cambridge: Cambridge University Press.