Matjaž Gams

Ljudje, stroji in inteligentne agentke

Ali je v ljudeh res nekaj več, nekaj, česar dosedanja znanost ne zna opisati? Čedalje več resnih znanstvenikov po svetu in Sloveniji dopušča to možnost.

Ker ljudje pri razmišljanju obdelujejo informacije, se je že pred šestdesetimi leti postavilo vprašanje, ali bodo računalniki znali početi vse to kar ljudje. Prvi veliki vizionar **Turing** (1936) je bil mnenja, da se bo to zgodilo okoli leta 2000. Turing je poleg dekodiranja nemškega kodirnega strojčka Enigme (2. svetovna vojna), za katerega so menili, da je popolnoma varen, postavil nekaj ključnih konceptov obdelovanja informacij in inteligence. Zasnoval je *Turingov stroj*, ki je odličen matematičen model za današnje računalnike. Inteligenco strojev naj bi preverjali s t.i. *Turingovim testom*, po katerem je v eni sobi človek, v drugi stroj (računalnik), komunikacija pa poteka prek omrežja. Če ne moremo odkriti, kje je človek, moramo stroju priznati inteligenco.

Neodvisno so tako **Turing** kot **Church**, **Post** in drugi dokazali, da so določene funkcije izračunljive in druge ne. *Izračunljive funkcije* (probleme) lahko izračuna *Turingov stroj*, neizračunljive pa so v resnici neizračunljive. Temu rečemo *Church-Turingova teza*, kot že ime samo pove pa gre le za tezo, ne za formalno dokazljivi izrek.

Razumevanje Church-Turingove teze je zapletena zadeva, povezana s pojmoma *simbolnega opisa* in izračunljivostjo. Poskusimo takole: Imamo dve hruški, dve luni, dva avtomobila. Povezovalna nit je koncept dvojke. Če ta koncept dvojke

priznamo v svojih mislih, na papirju, na Turingovem stroju, na računalniku, v fizikalnem svetu, potem smo pristali na to, da je vse moč zapisati s simboli (pisanimi, govornimi ...). Pisna "ljubezen" pomeni tudi dejansko ljubezen. Pri izračunljivosti pa gre v bistvu za rešljivost. Izraz "*računanje*" v resnici pomeni "*reševanje*", dvoumnost pa se pojavlja zaradi prevoda iz angleščine.

Ob priznavanju temeljnih kamnov računalniških znanosti (Turingovega stroja in Church-Turingove teze) so računalniki teoretično tako zmogljivi kot ljudje ali katerokoli drugo informacijsko bitje ali stroj. Ker se signali po žicah računalnikov premikajo s hitrostjo svetlobe, po človeških nevronih pa s hitrostjo zvoka, je potemtakem le vprašanje razvoja minituarizacije in tehnike, pa bo nekoč računalnik prehitel ljudi.

Ta optimistični pristop je doživel več razočaranj. Med najbolj znanimi kritiki so **Dreyfus** (1979), **Searle** (1982) in **Winograd** (1991). Raziskovalci so obljubljali gradove v oblakih, po nekaj letih financiranja pa niso bili nič bliže pravi inteligentnosti. Zanimanje in sofinanciranje se je zmanjšalo in pa se ponovno pojavilo z novim podpodročjem umetne inteligence. Valovi optimizma so gradili na nevronskih mrežah, ekspertnih sistemih itd. (Ne smemo pa pozabiti, da je ravno umetna inteligenca kljub neuspehom pri razvoju resnično inteligentnih sistemov eden od generatorjev razvoja klasičnega računalništva.)

Zadnja pomembna prelomnica se je dogajala nekaj let nazaj. **Pollock** (1989) piše: "Umetna inteligenca je imela sanje (op.p. o snovanju inteligentnih strojev) od svojega nastanka dalje, vendar te sanje izginevajo. Zato ker doseženi rezultati daleč zaostajajo za sanjami."

Minsky (1987; 1991) pravi: "Snovanje uma (angl. mind design) bo v bodočnosti bistveno drugačno kot dosedaj."

Enega najpomembnejših ugovorov je predstavil **Roger Penrose** (1989; 1994). Po njegovem so ljudje ali teoretično močnejši ali vsaj praktično drugače narejeni kot računalniki, da so (skoraj) nepremostljive razlike med njimi. V človeških glavah naj bi potekali *neizračunljivi kvantni procesi*. Konkretno naj bi se to dogajalo v *mikrotubulih*. Za dokaz svoje trditve Penrose vzame ***Gödlov teorem***. Noben formalen sistem namreč ne more videti resničnosti stavka: *Jaz (stavek) sem nedokazljiv*. Čeprav formalni sistemi lahko dokažejo, da ta stavek drži (obstajajo programi, ki so to v resnici naredili), ne morejo narediti sklepa, da je stavek tudi vsebinsko v resnici nedokazljiv. Seveda se vsakemu normalnemu človeku takoj zdi jasno, da je tak stavek potem res nedokazljiv, vendar formalni zagovorniki trde matematike hitro pokažejo, da je bistvena razlika med "zdi se mi" in "dokazal sem".

Je razlaga razlik med računalniki in ljudmi v kvantnem računanju? Pokazalo se je, da Turingovi stroji ne pokrivajo v celoti računanja v vsej naši naravi (vesolju). Turingov stroj temelji na

bitih, enotah informacije z dvema vrednostima: 0 in 1. To zajema vso naravo, kot jo vidimo ljudje, ne zajema pa kvantnih pojavov. Kvantni pojavi zajemajo linearno transponirane delce, torej se lahko nek atomski delec nahaja na dveh mestih naenkrat. Paradoksalno, ampak naše ocenjevanje je pač le sad izkušenj življenja v svetu velikih delcev. Prav tako kot pri kvantni fiziki pa se tudi s *kvantnimi stroji* da povsem lepo računati. Tako je **Benioff** leta 1982 opisal načrt za izgradnjo klasičnega računalnika iz kvantnih komponent. Leta 1985 je **Deutsch** definiral *univerzalni kvantni računalnik*, tj. kvantno verzijo univerzalnega Turingovega stroja. V zadnjih letih se kopica posameznikov (npr. **Lloyd**, **Shor**, **Kimble, Wineland**) ukvarja z izgradno modulov kvantnih računalnikov v praksi. Rezultati so daleč od industrijske uporabnosti, hkrati pa predstavljajo velik napredek v primerjavi s stanjem nekaj let nazaj.

Kaj lahko računajo kvantni stroji?

– Pri izračunu "klasične" funkcije, npr. množenju celih števil, kvantni računalniki niso bistveno hitrejši ali počasnejši od Turingovih strojev.
– Pri reševanju nekaterih problemov, npr. iskanju faktorja danega števila, so kvantni računalniki lahko bistveno hitrejši.
– Pri dokazovanju izrekov lahko kvantni računalniki včasih kaj dokažejo brez dejansko izvedenega dokaza.
– Pri nekaterih simulacijah, npr. simulaciji potresa, so lahko bistveno hitrejši.
– Omogočajo bolj zanesljive načine kodiranja in odpirajo nove možnosti v umetni inteligenci. Zanimive so predvsem misli v zvezi s svobodno voljo, saj pravi kvantni pojavi omogočajo povsem nedeterminirano možnost odločanja.

Kako lahko kvantni stroji omogočajo drugačne načine računanja? Po **Deutchu** (1992) se to dogaja zaradi t.j. *teorije mnogoterih svetov*, najbolj široke izmed interpretacij kvantne fizike. Računanje poteka v vzporednih svetovih, ki sicer niso direktno dostopni, vendar računsko omogočajo nekatere zaključke. V realnem življenju se teh vzporednih svetov ne zavedamo, saj naše življenje poteka le v enem izmed veliko možnih. Ta teorija omogoča potovanje v času; omogoča celo to, da se človek sreča s svojo nekaj starejšo verzijo, ki je šla potovat v času. Pri tem potovanju pa vsak posameznik nosi svoj "čas" s seboj in ga pri potovanju tudi troši.

Čeprav kvantni računalniki znajo izračunati nekaj nalog, ki so trd oreh za Turingove stroje, pa so načeloma enako zmogljivi. Oboji stroji lahko rešijo vse rešljive naloge in nobeni nobene nerešljive.

Naslednja misel bo po vsem napisanem zvenela že dokaj skromno: Načeloma je vseeno, ali računamo s Turingovim

strojem, PC-jem, superračunalnikom, kvantnim ali mehanskim računalnikom. Vsi računalniki bodo znali izračunati iste naloge, eni hitreje, drugi počasneje. Torej je vprašljivo, ali kvantno računanje res omogoča tisto več, kar naj bi bilo v naših glavah.

Kako je potem z računalniki? Jasno je, da se razvijajo bistveno hitreje kot ljudje. Kjer se da problema lotiti z računanjem in algoritmi, slej ko prej prehitijo ljudi. **Deep Blue**, IBM-ov šahovski program, je nadigral **Kasparova** z vrhunsko igro, ki je temeljila na izredno hitrem računalniku in zapletenih algoritmih.

Pa vendar človek nehote dobi vtis, da je Deep Blue popolnoma neinteligenten. Saj le izredno hitro nekaj izračuna, nikjer ni govora o kakršnemkoli razumevanju, kakršnikoli vsebini. Računalniki so izredno hitri pri računanju in shranjevanju golih informacij, vendar vse to počnejo na drugačen, manj sposoben način kot "počasni" ljudje. Težko je verjeti, da bi narava potrošila toliko let revolucije za "uboge" človeške možgane, ki računajo milijonkrat počasneje kot računalniki. Če bi bilo računanje res tako pomembno, bi ljudje že davno imeli v možganih tudi računalniška vezja, saj silicijevi atomi niso nič redkega v naravi.

Srž problema si lahko ogledamo še na nekaj primerih:

Po **Wilkesu** v zadnjih desetletjih kljub neverjetni rasti sposobnosti računalniki niso postali vsebinsko nič dejansko pametnejši. **Sloman** opozarja na problem Einsteinove knjige, kjer celotne Einsteinove možgane prepišemo na knjigo. Ta knjiga sicer vsebuje vse informacije, nima pa mehanizma izvajanja, zato ne more nič narediti. Očitno je potrebno nekaj več kot samo informacije. **Searle** opozori na problem *Kitajske sobe,* kjer imamo poleg knjige vseh besed in pravil prevajanja tudi človeka, ki izvaja ta pravila, ne da bi jih razumel. Čeprav je sposoben brezhibno prevesti iz kitajščine v slovenščino, ni nobenega razumevanja. Po slovitem *Turingovem testu* bi taka kitajska soba dajala vtis, da je v sobi dejansko neko razumevanje, pa ga v resnici ni. (Mimogrede – na internetu dobimo programe, ki prevajajo med večino velikih evropskih jezikov.) Seveda je primer s kitajsko sobo le nazorni prikaz, da bi računalniški program, ki bi po svojih pravilih izvajal nalogo, lahko odlično rešil nalogo, pa vseeno ne bi nič razumel.

In tako danes tudi je. V veliko domenah so računalniki uspešnejši kot ljudje, vendar se ljudem vseeno zdijo popolnoma brez inteligence ali zavesti. Kot da bi skušali s helikopterjem priti na Luno. Gremo višje in višje, a od Zemlje se ne moremo odlepiti. Kaj manjka računalniškim sistemom, da bi postali vsaj malce inteligentni? Gotovo nekaj več kot nekajkrat večja hitrost.

Še leta 1994 je **Abrahamson** za podobne misli obsojal **Penrosa** kot "enega izmed mnogih, ki v imenu kdo ve katerega boga poskušajo uničiti racionalnost in znanost. Posebej moramo biti pozorni na resne znanstvenike, ki skušajo razumno sklepanje in logiko zamenjati s spiritualizmom."

Nekaj let nazaj pa se je zgodila omenjena pomembna prelomnica. V odgovor **Abrahamsonu** je **Cronin** (1994) zapisal: "(staromodna, stroga) umetna inteligenca je postala vase zaprta disciplina, ki se ukvarja sama s seboj." **Angell** (1993, str.15) piše: "Ali res mislijo, da je mogoče zajeti pomen z logično-matematično analizo?"

Zdravi razum nam pove tole: če je teorija še tako lepa, pa v praksi ne uspe, potem je nekaj narobe s teorijo.

Nova (šibka) umetna inteligenca in *kognitivne znanosti* so drugačne, daleč bolj odprte in daleč bolj interdisciplinarne. V svojih vrstah imajo vrsto najbolj znanih svetovnih znanstvenikov:

- **Francis Crick** je dobil Nobelovo nagrado za odkritje strikture DNA. Uvedel je raziskave zavesti med regularne znanstvene discipline. Zavesti se loteva predvsem z nivoja nevronske strukture možganov.
- **Gerald M. Edelman** je dobil Nobelovo nagrado za raziskave človeškega imunskega sistema. Kot osnovni mehanizem zavesti predlaga *nevronski darvinizem*, kjer skupine nevronov tekmujejo v populacijski evoluciji.
- **Brian D. Josephson** je dobil Nobelovo nagrado za študij kvantnih efektov pri Josephsonovem spoju. Predlaga združitev fizičnih in psihičnih izkušenj.
- **Maurice W. Wilkes** je eden izmed pionirjev umetne inteligence, ki na osnovi empiričnih rezultatov trdi, da s klasičnim pristopom ni mogoče doseči inteligence na Turingovih strojih.

Staromodna umetna inteligenca je v bistvu podoživela trende formalnih znanosti: osnovne zakonitosti so odkrite, novih spoznanj je malo. *Objavi ali propadi* (angl. publish or perish) je geslo, ki bolj spominja na modo, kot pa na pravo znanost. V pomanjkanju pravih odkritij so se znanstveniki začeli zatekati v preštevanje števila objav in citatov. Prava teža novih spoznanj marsikje ni med osnovnimi znanstvenimi kriteriji, tudi v Sloveniji pogosto ne.

Problemi v znanosti so pripeljali do večje dojemljivosti za popularne nove, čeprav včasih nepreverjene teorije, in v večjo usmerjenost v praktične probleme. Nova umetna inteligenca in kognitivne znanosti so s tem dobile legitimnost. Hkrati pa so se pojavili novi problemi. V širini in odprtosti gibanja se je pojavilo veliko nepreverjenih in malo verjetnih novih teorij. Medtem ko je stara, toga znanost omogočala le razvoj zelo podrobnim novim teorijam in zavračala premalo verjetne, nove struje omogočajo poplavo v povprečju neveljavnih teorij. Še več, med resne znanstvenike nove dobe se je pomešala množica mistikov, spiritualistov in fantazerjev. Zapornice, ki jih je previsoko dvigal **Abrahamson** leta 1994, so že leta 1998 prenizko spuščene.

Drug problem novih struj je premajhno število pomembnih splošno sprejetih novih spoznanj. Za to morda obstajajo objektivni razlogi. Tako morda ni možno zapisati teorije o zavesti ali inteligenci, ker sta morebiti neizračunljivi ali simbolno neopisljivi. Morda so naši človeški možgani nezmožni zapisati znanja o samem sebi, tako kot so računalniki nezmožni rešiti veljavo Gödlovega stavka (izjava o samem sebi).

Vseeno je poleg kvantne kar nekaj zanimivih teorij. Po *teoriji mnogoterosti* (**Gams** 1997) je človeško znanje mnogotero, sestavljeno iz več modelov problema in rešitve. Možgani dinamično kombinirajo te modele in jih integrirajo. Torej resnica ni več ena sama; algoritem ni več en sam kot pri Turingovem stroju. (Paralelnost le pospeši hitrost izvajanja, mnogoterost je vsebinsko drugačen koncept.) Blizu te teorije sta Edelmanova teorija tekmovalnih procesov v nevronskih strukturah in Deutchova teorija več svetov. Daleč bolj kot število nevronov je pomembno število povezav, saj le-te omogočajo večkratne poglede na isti problem.

Zelo pomembno teoretsko novost je objavil **Peter Wegner** leta 1997 v ACM. Opozoril je, da *Turingov stroj z odprtim trakom* ni enako močan kot *interakcijski Turingov stroj*. Torej so programi močnejši od enačb in ljudje močnejši kot sedanji računalniki. Da je v praksi razlika med obema računskima konceptoma, se hitro prepričamo, če primerjamo moč samostojnega PC-ja in PC-ja s priključkom na internet. Dodatna moč pride iz dinamične odprte komunikacije (interakcije) z okoljem. Formalna znanost je to razliko dosedaj zanemarjala iz praktičnih razlogov, saj je samostojni PC lepo opisljiv z matematično-formalnim mehanizmom kot Turingov stroj, medtem ko interakcijski Turingov stroj ni več lepo formalno opisljiv. To je tudi intuitivno razumljivo, saj formalno ne moremo opisati vseh informacij, ki se nahajajo na internetu, in interakcije z drugimi, nepoznanimi akterji prek interneta.

Praktična izvedba interakcijskega Turingovega stroja je torej preprosta – npr. že omenjeni PC s povezavo z internetom. Bolj znani so *inteligentni agenti*. (Pozor – ne interakcijskost ne mnogoterost ne zadoščata za inteligenco, sta le dve izmed več potrebnih lastnosti, ki jih morajo imeti resnično inteligentni sistemi.)

Inteligentni agenti so računalniški sistemi, ki simulirajo obnašanje tipičnega agenta, npr. zavarovalniškega agenta. Ta nam skuša prijazno ponuditi nekaj možnih zavarovanj, glede na naše odzive pa prilagodi ponudbo. Drugače kot *ekspertni sistemi,* ki skušajo posnemati vrhunske človeške strokovnjake, inteligentni agenti posnemajo prijazne človeške agente (včasih tudi birokrate). Ker inteligentni agenti lahko delajo 24 ur na dan 365 dni na leto približno tako kvalitetno kot birokrati, vendar s po 1000 strankami naenkrat, imajo zajamčeno bleščečo prihodnost. Ključna lastnost agentov je *avtonomnost*. Do nje je prišlo, ker ljudje ne moremo

več obvladovati interneta tako, kot smo prejšnje računalniške sisteme. Internetsko okolje je preveč hitro spreminjajoče se, da bi togi programi po strogih navodilih še lahko počeli kaj koristnega. Zato smo nezavedno agentom dali določeno svobodo odločanja. Tako se je pojavila primitivna *omrežna inteligenca.* Ta inteligenca je velik korak naprej proti pravi inteligenci, vendar je še vedno bolj podobna posebnim zvrstem inteligence, npr. motorični inteligenci. Omrežna inteligenca pomeni sposobnost obvladovanja globalnih iformacij.

Inteligentni agenti so med najbolj pogosto uporabljanimi programi na internetu in delno tudi na PC-jih. Tak je npr. Office Assistant, pomočnik v obliki sponke v Microsoft Officeu Med prvimi preprostimi agenti na internetu najdemo brkljalnike. Ti danes že veljajo za klasične aplikacije in nič več za inteligentne agente. Zgodba umetne inteligence se ponavlja – ko kak produkt postane splošno uporaben in zanimiv, se začne šteti za aplikacijo. Znanstvene objave so se medtem že zapičile v pogosto (pre)zapletene algoritme brez uporabne veljave. Drug problem agentov je v tem, da so s sočnostjo imena (angl. buzzword) raznovrstni prodajalci začeli prodajati vse mogoče sisteme pod rubriko inteligentni sistemi in inteligentni agenti.

Ne glede na probleme je večina revij s področja računalništva in informatike objavila kopico člankov o agentih in najpogosteje tudi posebne namenske številke. Taka sta npr. AI magazine ali IEEE Internet Computing poleti leta 1997. Inteligentni agenti (termin agentke v naslovu članka je bil izbran zaradi prodora žensk v vse pore moderne družbe) so nosilci nove generacije računalniških sistemov, tesno povezanih s prodorom interneta. To je *informacijska doba.*

Inteligentne agente imamo tudi v Sloveniji. Na http://www2.ijs.si//mezi/agents.html je zbirka 60 najbolj zanimivih svetovnih agentov. Na http://www-ai.ijs.si//ema/welcome-s.html je inteligentna agentka Ema, ki nudi zaposlovalne informacije. Ema je najstarejša agentka v Sloveniji in z okoli 40.000 obiski mesečno tudi najpogosteje obiskan sistem umetne inteligence (najpogosteje obiskana je Elisa, preprost program za klepet v naravnem jeziku). Ema tekoče govori angleško in slovensko, pošlje vam elektronsko pošto, če se kje pojavi zanimiva zaposlovalna informacija. Ema je nekaj časa nudila relativno največ zaposlovalnih informacij na državo. Glede tega je bila Slovenija prva v svetu, Republiški zavod za zaposlovanje pa med prvimi evropskimi zavodi.

Inteligentne agentke so tu, še več jih prihaja. So najbolj obetavna nova smer umetne inteligence, med najpogosteje uporabljanimi sistemi na internetu. So prva generacija računalniških sistemov z določeno avtonomijo, svobodo izbire (svobodo volje?). So sestavni del informacijske dobe, interneta in omrežne inteligence.

Resnično inteligentni sistemi se skrivajo v daljnji prihodnosti. Brez dvoma jih bomo nekoč ustvarili in takrat bomo Stvaritelji v dobrem in slabem pomenu besede. Do tedaj pa nas nosi želja, da bi odkrili največjo skrivnost vesolja, najbolj zapletenega samostojnega sistema – človeških možgan in uma.

LITERATURA

I. O. ANGELL (1993), Intelligence: Logical or Biological, Viewpoint, Communications of the ACM 36, pp. 15-16.

J. R. ABRAHAMSON (1994), Mind, Evolution, and Computers, AI magazine, Spring 1994, pp. 19-22.

F. CRICK (1994), The Astonishing Hypothesis, The Scientific Search for the Soul, New York.

M. R. CRONIN (1994), A Reply to Mind, Evolution, and Computers, AI magazine, Summer 1994, p. 6.

D. DEUTCH (1985), Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer, Proceedings of Royal Society, pp. 97-117.

D. DEUTCH (1992), Quantum Computation, Physics Word, pp. 57-61.

H. L. DREYFUS (1979), What Computers Can't Do, Harper and Row.

G. EDELMAN (1992), Bright Air, Brilliant Fire, On the Matter of the Mind, Penguin Books.

M. GAMS: Is Weak AI Stronger than Strong AI? in Minds Versus Computer, IOS Pess, (eds.) M. Gams, M. Paprzycki, X. Wu, pp. 30-45, 1997.

M. MINSKY (1987), The Society of Mind, New York: Simon and Schuster.

M. MINSKY (1991), Society of Mind: A Response to Four Reviews, Artificial Intelligence 8, pp. 371-396.

R. PENROSE (1989), The Emperor's New Mind: Concerning computers, minds, and the laws of physics, Oxford University Press.

R. PENROSE (1994), Shadows of the Mind, A Search for the Missing Science of Consciousness, Oxford University Press.

J. L. POLLOCK (1989), How to Build a Person: A Prolegomenon, MIT Press.

J. R. SEARLE (1982), The Chinese Room Revisited, Behavioral and Brain Sciences 8, pp. 345-348.

A. SLOMAN (1992), The Emperor's Real Mind: Review of the Roger Penrose's The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics, Artificial Intelligence 56, pp. 335-396.

A. M. TURING (1936), On Computable Numbers with an Application to the Entscheidungsproblem, Proc. London Math. Soc. 2, pp. 230-265.

P. WEGNER (1997), Why Interaction is More Powevful than Computing, Communications of the ACM, May 1997/Vol. 40, No. 5, pp. 81-91.

M. W. WILKES (1992), Artificial Intelligence as the Year 2000 Approaches, Communications of the ACM 35, pp. 17-20.

T. WINOGRAD (1991), Thinking Machines: Can there be? Are We?, The Boundaries of Humanity: Humans, Animals, Machines, Berkeley, University of California press, pp. 198-223, (ed.) J. Sheehan, M. Sosna.