Háskóli Íslands

Hugvísindasvið

Heimspeki

Eiga tölvur að fá mannréttindi?

Hugleiðingar um hugsun, gervigreind, siðfræði og frumspeki

Ritgerð til BA-prófs í heimspeki

Helgi Jónsson

Kt.: 130893-2609

Leiðbeinandi: Eyja Margrét Brynjarsdóttir

Maí 2017

Efnisyfirlit

Inngangur

„Allir menn eiga rétt til lífs, frelsis og mannhelgi.“
Svo hljóðar 3.grein Mannréttindayfirlýsingar Sameinuðu þjóðanna, sem samþykkt var í París 10.desember árið 1948. Yfirlýsingin setur skýrar reglur um frelsi og mannhelgi einstaklingsins sem eru í hávegum hafðar í flestum samfélögum nútímans, og brot á þeim eru litin mjög alvarlegum augum.
En nú er önnur öld runnin upp, öld stórkostlegra tækniframfara. Á síðustu áratugum hafa vísindamenn unnið að því að þróa svokallaða gervigreind (e. *artificial intelligence*) í tölvum, tölvur sem virðast ‘hugsa’. Tölvurnar m.a. meðtaka og vista upplýsingar, vinna úr þeim, gefa svör og spá fram í tímann og slíkt ferli líkist mannlegri hugsun í okkar augum. Frægt er orðið þegar vitvélin IBM Deep Blue sigraði skákmeistarann Garry Kasparov í skákeinvígi árið 1997. Kasparov taldi að svindlað hefði verið á sér og krafðist þess að fá að endurtaka leikinn. Forsprakkar IBM neituðu ásökunum Kasparovs og sögðu að vitvélin hefði virkilega sigrað meistarann. Deep Blue hefði færni til að reikna út 200 milljón mögulega skákleiki á sekúndu og hefði því stórkostlega skákhæfileika.[1]
Þótt Deep Blue hafi verið byltingarkennd fyrir sinn tíma og haft undraverðan skilning á skák, þá munum við hér ræða um gríðarlega háþróaða gervigreind sem er ekki enn til og gæti virst mörgum vera tómur vísindaskáldskapur. Tölvur sem virðast okkur hafa mannlegar tilfinningar, innsæi og skilning. Við munum ekki fást hér við spurninguna hvort tölvur með slík forrit muni nokkurn tímann verða til, heldur höfum við það okkur að forsendu að þær verði einhverntímann til, rökræðunnar vegna (hvort sem við séum endilega sannfærð um það eða ekki).
Ef við lítum á mannréttindayfirlýsingu Sameinuðu þjóðanna liggur í augum uppi að sumum atriðum sáttmálans þyrfti að breyta ef um tölvuforrit væri að ræða. Sum atriðin virðast miðast beint að því hvaða líkamlegar þarfir manneskja klædd holdi og blóði hefur. Til dæmis stendur í 25.grein að allir menn eigi kröfu til lífskjara eins og matar, klæðnaðar, húsnæðis o.s.frv. Ekkert þessara atriða virðist eiga við þegar kemur að tölvuforritum, enda hafa tölvuforrit ekki „líkama” í hinum mannlega líffræðilega skilningi. Því þyrfti hugsanlega að búa til nýja yfirlýsingu og nýtt hugtak, vélréttindi,

[1] „Deep Blue“, *IBM* http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/.

sem miðaðist meira að tölvuforritum og vitvélum. Slíkar hugleiðingar munu þó ekki vera í forgrunni hér.

Spurningin sem við eigum við hér snýst ekki um smávægileg atriði sem þessi, heldur er hún af mun frumspekilegri toga. Grundvallaratriði þess að hafa tilkall til mannréttinda hefur hingað til verið það að vera maður; lifandi spendýr af gerðinni Homo Sapiens. Að vísu hafa slík réttindi ekki alltaf verið virt og eru það ekki sumstaðar í veröldinni, t.a.m. í Norður-Kóreu þar sem fólk er pyntað og jafnvel tekið af lífi fyrir það eitt að hafa skoðanir sem þóknast ekki stjórnvöldum.[2] Engu að síður stendur í Mannréttindayfirlýsingunni að allir menn *eigi* að vera fæddir frjálsir og búa við ákveðin grundvallarréttindi. Skilgreiningin á manneskju hefur verið tiltölulega einföld fram að þessu, en nú á öld vitvéla er okkur unnt um að spyrja hvort ákveðin tölvuforrit, sem virðast að öllu leyti vera mennsk, ættu að hafa nokkurn veginn sömu réttindi og menn. Ég geng hér út frá því að mannréttindi séu manngert fyrirbæri og því séu engin náttúrulögmál sem fá því ráðið hverjir ættu að fá þau og hver ekki, heldur eru teknar siðferðilegar ákvarðanir um slíkt. Hér erum við að velta fyrir okkur hvort það sé siðferðilega rétt að veita slíkum forritum réttindi, eða hvort það sé siðferðilega rangt að neita þeim um slík réttindi. Allt er þetta gert undir þeirri (háfleygu) forsendu að háþróaðar gervigreindartölvur muni að lokum búa í mannlegu samfélagi og hafa virk samskipti við mannfólk sem og við hvorar aðrar. Spurningin fjallar bæði um hvernig mannfólk eigi að koma fram við tölvur og hvernig tölvur eigi að koma fram við hverja aðra.

Forsendur mannréttinda

Það kann að vera að margir hafi ranghvolft augunum þegar þeir lásu titilinn á þessari ritgerð. Hugsanlega hefur það fólk yppt öxlum og spurt hneykslað „hvers vegna" ? Hvers vegna ættum við að vera að ræða réttindi tölva sem eru ekki ennþá til, og ekki einu sinni víst að þær verði nokkurn tímann til? Ég tel aftur á móti að einmitt nú sé rétti tíminn til þess að byrja að ræða þessi málefni, og helst ekki seinna en nú. Í þúsundir ára var þrælahald réttlætt með rökum á við að þrælarnir væru „óæðri verur"

[2] „Víðtæk mannréttindabrot í Norður-Kóreu," Rúv http://www.ruv.is/frett/vidtaek-mannrettindabrot-i-n-koreu

„og ekki mennskir“. Nú er það hins vegar viðurkennt viðhorf í flestum samfélögum að menn skuli vera jafnir óháð litarhafti eða trúarskoðunum (sbr. Mannréttindayfirlýsingu Sameinuðu þjóðanna). Aftur á móti er ekkert sem hindrar okkur að fara aftur í gamla farið þegar kemur að nýjum vitrænum verum sem við skiljum ekki fullkomlega. Sannfærandi gervigreind gæti verið nytsamleg í ýmislegt og hægt er að ímynda sér að aðilar eins og t.d. fyrirtækjaeigendur réðu vélmenni vædd tölvubúnaði með gervigreind í erfiða líkamlega vinnu og forrituðu þau til að finna til sársauka ef þau hlýddu ekki fyrirmælum. Þess vegna er mikilvægt, ekki seinna en nú, að fara að ræða spurninguna hvort tölvur ættu að hafa réttindi.

Hvað er að vera persóna?

Við skulum hér byrja á að ræða spurninguna hvort tölvur sem geta hugsað ættu að hafa einhverskonar réttindi, og geyma spurninguna hvort þær ættu að hafa réttindi sem jafnast á við réttindi manna. Við gætum hér á þessum tímapunkti velt fyrir okkur hvað það þýði að vera persóna. Ég tel að það sé afar erfitt að ætla sér að svara slíkri spurningu hér, enda er sú spurning gríðarlega stór og margþætt og gæti verið rannsóknarefni út af fyrir sig. Sumir myndu t.d. segja að dýr séu persónur og hægt er að koma með góð rök fyrir því. Dýr eru fær um sjálfstæða hugsun og hvert dýr virðist hafa sinn eigin sérstaka persónuleika. Myndir þú ekki segja að fjölskylduhundurinn úr æsku þinni hafi haft sinn eigin persónuleika og skapgerð? Þó margir myndu skilgreina hunda og ketti sem persónur er samt enn loðið hvar skilgreiningin liggur á persónuhugtakinu. Við gefum þeim t.d. ekki jöfn réttindi á við menn. Að vísu er mikið af fólki sem berst fyrir réttindum dýra, en þau réttindi beinast aðallega að því að ekki megi sæta þau pyntingum eða fara illa með þau. Í 2. grein íslenskra dýraverndunarlaga stendur m.a. „Skylt er að fara vel með öll dýr. Óheimilt er að hrekkja dýr eða meiða.” [3] Hins vegar stendur ekkert um að óheimilt sé að aflífa dýr eða þau eigi að hafa kosningarétt og friðhelgi einkalífs, sem eru réttindi sem við veitum okkur sjálfum. Það eru sannarlega ekki margir sem vilja banna mannfólki að leggja eignarhald á dýr og halda þeim í búrum og girðingum (þó Tom Regan hafi

[3] „Lög um dýravernd,“ Alþingi http://www.althingi.is/lagas/119/1994015.html.

verið einn af þeim, en í grein sinni „The Case for Animal Rights"[4] frá 1983 fullyrti hann að það að gefa búfénaði betri aðstöðu og umhirðu réttlætti ekki það að halda þeim föngnum og það eina sem myndi til þess duga væri að leysa upp allan búskap eins og hann leggur sig. Ekki verður þó frekar fjallað um slíkar skoðanir hér). Hins vegar er ljóst að það að halda mannfólki í búrum og girðingum gegn þeirra vilja væri hreint og klárt mannréttindabrot. Málið vandast svo enn frekar þegar hugsað er til allra þeirra ótalmörgu tegunda dýra sem til eru í heiminum. Ef fjölskylduhundurinn er persóna, eru þá ekki flærnar á baki hans líka persónur? Flestir myndu líklega segja að það sé einhverskonar loðinn stigbundinn munur þarna á milli.

Hverjum veitum við réttindi?

Ef við einföldum spurninguna þá getum við spurt að því hverjum við veitum réttindi. Hvað með dauða hluti? Hefur stóri steinninn í garðinum réttindi? Eru til hópar af fólki sem berjast fyrir réttindum grjóthnullunga, reyna að koma á stokk lagaákvæðum um að ekki megi brjóta þá, færa þá úr stað gegn þeirra eigin vilja o.s.frv? Að vísu er hér á landi til fólk sem trúir því að álfar og huldufólk eigi sér bústað í vissum grjóthnullungum. Vegna baráttu þeirra hafa fyrirhugaðir vegir verið sveigðir fram hjá steinum sem fólkið taldi að væri heimili álfa, sem og steinarnir færðir úr stað. Hins vegar má líta svo á að fólkinu sé ekki annt um grjóthnullungana sjálfa, þeim er ekki annt um hvernig steinunum sjálfum *líður*. Þeim er annt um hulduverurnar sem þau telja að búi þar. Að berjast fyrir *réttindum* grjóthnullunga virðist í fljótu bragði algjör fásinna. En hvers vegna? Hvaða eiginleika þarf hluturinn að hafa til að geta haft tilkall til réttinda? Þarf hann að vera lifandi? Þá getum við spurt okkur hvort fíflarnir úti í garði eigi að hafa réttindi. Þeir eru sannarlega lifandi, en er einhver ástæða til að berjast fyrir réttindum fífla? Hvers vegna ekki?

Það er kannski ágætt að rifja upp hvers vegna fyrirbærið „mannréttindi" var búið til til að byrja með. Hugtakið spratt upp vegna nauðsynjar. Hugmyndir um mannréttindi eins og við þekkjum þau í dag spruttu upp á öld Upplýsingarinnar í Evrópu, þar sem þrælahald var leyfilegt og fólk var oft beitt grimmilegum refsingum fyrir að beita sér gegn ráðamönnum. Því getum við litið svo á að mannréttindi hafi sprottið upp vegna

[4] Regan, „The Case for Animal Rights," *Ethics in Practice,* ritstj. Lafollette, Hugh (Sussex: Wiley Blackwell 2014) 192-197.

þjáninga, en hvaða eiginleika þarf einhver að hafa til að þjást? Hvað er það að þjást? Við göngum hér út frá þeirri skilgreiningu að þjáningar séu einhvers konar neikvætt ástand af tilfinningalegum toga. Það hlýtur þá að vera hægt að gera ráð fyrir því að sá sem þjáist hefur getuna til að hafa tilfinningar. En til eru ólíkar gerðir af tilfinningum. Til dæmis finn ég fyrir lyklaborðinu á tölvunni minni, það er tilfinning. En hún er algjörlega líkamleg. Lyklaborðið hefur ekki neins konar áhrif á mína andlegu líðan með því einu að ég finni fyrir því. Á ensku myndum við hér tala um orðið *sensation*, sem við köllum hér skynjanir. Á hinum endanum höfum við huglægar tilfinningar, sem á ensku myndu útleggjast sem *emotions*. Þær eru einungis í huganum og tengjast oftast hugsunum eða skynjun okkar á heiminum. Til dæmis er ég núna spenntur fyrir að útskrifast úr Háskóla Íslands, en sú tilfinning virðist ekki hafa nein líkamleg áhrif á mig, allavega ekki á yfirborðinu. Auðvitað er kjánalegt að flokka tilfinningar niður í þessi tvö hólf; vissulega eru flestar tilfinningar blanda af báðum gerðum, eða allavega virðast huglægar og líkamlegar tilfinningar geta haft áhrif hvor á aðra. Ef einhver myndi kýla mig í magann myndi ég finna til líkamlega, en það myndi einnig hafa huglæg áhrif á mig og ég yrði reiður eða sár. Það væri þá andleg tilfinning sem myndast út frá líkamlegri tilfinningu. Einnig gæti ég verið mjög kvíðinn fyrir próf og fengi magaverk úr streitu, og þá væri það líkamleg tilfinning sem myndast út frá huglægri tilfinningu. Héðan í frá skulum við því ganga út frá þeirri forsendu að bæði líkamlegar og huglægar tilfinningar hafi áhrif á tilkall einstaklingsins til réttinda. Það skipti ekki máli hvort einstaklingur sé beittur líkamlegum eða huglægum þjáningum, bæði flokkist undir brot á mannréttindum. Við skulum svo síðar athuga hvort einstaklingur sem hafi einungis huglægar en ekki líkamlegar tilfinningar, svo sem tölva vædd gervigreind (án þess þó að gera strax ráð fyrir því að slík tölva hafi tilfinningar) hafi jafn mikið tilkall til réttinda og sá sem hefur báðar gerðir, svo sem manneskja. En þá spurningu skulum við ekki ræða frekar hér.

Við getum þá hugsanlega ályktað að til þess að þurfa réttindi þarf viðkomandi að geta haft tilfinningar. En ég vil ganga enn lengra í ályktun minni. Hvað þarf til að geta haft tilfinningar? Það er erfitt að ímynda sér eitthvað sem hefur tilfinningar en getur ekki hugsað, því eru tilfinningar ekki nátengdar hugsunum?

Segjum sem svo að ég sé í matarboði einhvern tímann í framtíðinni, og þar kemur að mér þekktur prófessor sem tekur í höndina á mér og hrósar mér fyrir frábæra bakkalársritgerð í heimspeki. Eyrun mín nema hrósið, hugur minn skynjar orðin,

hugsar um þau, skilur þau sem jákvæð og framkallar jákvæða tilfinningu. Ég finn þessa tilfinningu vegna þess að ég hugsa. Það er nauðsynlegur hluti til þess að finna þessa tilfinningu að ég hugsi. Ef ég hrósa steininum úti í garði gerast tveir hlutir: a) steinninn finnur enga tilfinningu vegna þess að hann er ófær um að hugsa, og b) nágrannar mínir halda að ég sé skrýtinn og hætta að bjóða mér í matarboð. Jafnvel líkamlegar tilfinningar tengjast huganum náið. Ef einhver klípur mig fast í handlegginn senda taugaendarnir í handleggnum skilaboð til hugans, sem nemur skilaboðin sem sársauka. Hlýtur það þá ekki að þýða að þeir sem eiga tilkall til réttinda séu færir um hugsun?

Hvað er meðvitund?

E.t.v. vilja margir nú fara að velta fyrir sér hvort tölvur þurfi að hafa meðvitund til að geta talist hugsa. Vandamálið sem slíkar spurningar kunna að vekja er að það veit enginn nákvæmlega hvað meðvitund er, og varhugavert er að reyna að fyllilega skilgreina hugtakið hér. Meðvitund er regnhlífarhugtak yfir margs konar ólík fyrirbæri sem gerast í hugsuninni, en í grein Kristins R. Þórissonar, „Vélvitund, meðvitund og sjálfsvitund í kjötvélum og vélmennum" frá árinu 2009 ræðir hann um kröfur sem gervigreindarforrit þurfi að uppfylla til þess að geta talist hafa meðvitund. Hann segir að meðvitundin sé flókið fyrirbæri sem sé samsett úr mörgum þáttum. Einn sá þáttur hljóti að vera að forritið „skilji sjálft sig" og tengsl sín við umhverfið að einhverju leyti, enda er það mikilvægur þáttur í sjálfstæðri ákvarðanatöku forritsins. Nefnir hann þar gervihnöttinn Deep Space One sem hafði verið forritaður til að skilja „líkama sinn" til þess að geta fengist við bilanir af ýmsum toga.[5] Annar þáttur hljóti að vera skilningur á núinu, því skilningur á því „virðist nátengdur skilningi okkar á okkur sjálfum og tilvist okkar" (bls. 121). Það að upplifa raunveruleikann núna virðist nátengt meðvitund okkar, því við upplifum alltaf allt núna og ef ég myndi ekki upplifa eitthvað núna myndi ekki nein upplifun vera til staðar. Það er ómögulegt að ímynda

[5] Kristinn R. Þórisson, „Vélvitund, meðvitund og sjálfsvitund í kjötvélum og vélmennum", *Veit efnið af andanum?,* ritstj. Steinar Örn Atlason og Þórdís Helgadóttir (Reykjavík: Heimspekistofnun, 2009), 120-121.

sér að upplifa eitthvað án þess að gera það á akkúrat því augnabliki, í því núi. Þess vegna hlýtur skilningur á núinu að vera mjög mikilvægur hluti af meðvitundinni.[6]

Ég er þeirrar skoðunar að meðvitund hljóti að þurfa að vera skilyrði fyrir því að tölva geti átt tilkall til mannréttinda, vegna þess að ég tel meðvitundina vera órjúfanlega hugsunum og tilfinningum. Hvernig getur eitthvað haft meðvitund en ekki hugsað? Það er ómögulegt að hugsa sér einhvern sem hefur eiginleika meðvitundar, t.d. skilning á sjálfum sér og skilning á núinu án þess að vera fær um nokkurs konar hugsun. Það hljóta að vera að minnsta kosti einhverjar grunnhugsanir til staðar eins og „ég er til“ og „ég er núna,“ vegna þess að slíkar hugsanir hljóta að vera nauðsynlegur þáttur af meðvitundinni. Einnig hljóta tilfinningar að vera náskyldar meðvitundinni, því þær eru alltaf bundnar við manns eigin persónulegu upplifun. Ef ég finn fyrir sársauka geri ég það með tilliti til alls konar eiginleika í meðvitundinni, t.d. finn ég hvar sársaukinn er í líkamanum, ég skil að ég er að finna til sársauka, ég veit að ég er að upplifa sársaukann *núna* og svo framvegis. Sársaukinn, líkt og allar aðrar tilfinningar mínar, er náskyldur minni eigin persónulegu upplifun; meðvitund minni. Ef við teljum að hugsanir og tilfinningar séu nauðsynleg skilyrði til að tölva eigi tilkall til mannréttinda, þá hljótum við að fallast á að meðvitund, eins og við höfum skilgreint hana, sé einnig nauðsynlegt skilyrði.

Ekki eru allir sem telja að gervigreind myndi meðvitund eða tilfinningar. Sumir telja að jafnvel þótt við teljum hin huglægu fyrirbæri sem við köllum meðvitund vera órjúfanlega hluta af mannlegri hugsun, þá er ekki þar með sagt að slíkt eigi einnig við um hugsun gervigreindar (ef hugsun skal kalla). Hún gæti virkað á allt annan hátt. Kannski myndar hún ekki sína eigin fyrstu persónu frásögn eins og mannfólk gerir, og tölvan hefur enga sjálfstæða upplifun heldur *virðist* aðeins gera það. Hér á eftir munum við fjalla um slíkar hugmyndir, einkum hugleiðingu John Searle um Kínverska herbergið.

[6] Kristinn R. Þórisson, „Vélvitund, meðvitund og sjálfsvitund í kjötvélum og vélmennum,“ *Veit efnið af andanum?,* 121-122.

Searle og Kínverska herbergið

Bandarískur heimspekingurinn John Searle skrifaði um tölvur og gervigreind í bók sinni *Minds, Brains and Science* árið 1984 þar sem hann fullyrti að tölvur væru ekki færar um að hugsa í rauninni, heldur gætu þær einungis líkt eftir hugsun. Hann setti þá hugmynd fram í hugsanatilraun sinni um Kínverska herbergið[7], en í þeirri grein biður hann lesandann um að ímynda sér sjálfan sig í hlutverki manneskju sem kann einungis ensku og er læst inni í herbergi. Í herberginu er kassi fullur af kínverskum táknum og bók með leiðbeiningum. Þeir sem framkvæma tilraunina senda inn tákn gegnum lúgu í herberginu og manneskjan í herberginu getur lesið leiðbeiningarnar og fundið viðeigandi svör í kassanum til að senda til baka gegnum lúguna. Leiðbeiningabókin er gríðarlega vel hönnuð og manneskjan verður svo góð í að svara spurningunum að fyrir þann sem les einungis svörin virðist hún kunna reiprennandi kínversku. Í rauninni er manneskjan einungis að fylgja leiðbeiningum á borð við: „Taktu þetta tákn sem lítur svona út úr kassa eitt og settu það við hliðina á öðru tákni sem lítur svona út í kassa tvö". Með því að líta á táknin og finna viðeigandi svör án þess að skilja þau er manneskjan, segir Searle, að hegða sér eins og tölva. Hún tekur inn upplýsingar og sendir frá sér upplýsingar nákvæmlega eins og einhver sem kann að tala kínversku, en skilur ekkert í raun og veru. Til þess að skilja tungumál þurfi maður ekki einungis að hafa táknin, heldur þurfi maður að geta tengt táknin við einhverja merkingu. „Tölva notast við setningarfræði, en ekki merkingarfræði" (a computer has syntax, but no semantics)[8], segir hann í bókinni. Tölva geti ekki haft neitt meira og dýpra en táknin og uppröðun þeirra. Setningarfræðin ein og sér sé ekki nóg til að mynda skilning. Því telur Searle að það skipti engu máli hversu snilldarlega tölva hegðar sér, hún muni aldrei geta hugsað í raun og veru. Hún getur hermt eftir en aldrei skilið, hún getur aldrei komist í það hugarástand að skilja (eða nokkuð hugarástand yfir höfuð).[9]

---

[7] John Searle, *Minds, Brains and Science* (London: Penguin Books, 1989), 31-36.
[8] Searle, *Minds, Brains and Science,* 33.
[9] Searle, *Minds, Brains and Science,* 31-36.

Gagnrýni á Searle

Hér mun ég koma með mótrök við þessari hugmynd hjá Searle, þar sem ég sýni fram á að hún gefi ekki nægilega góða mynd af því hvernig tölvur starfa. Ég vísa hér til gagnrýni Tim Crane úr bókinni *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation,* en Crane telur að það að skilgreina manneskjuna í herberginu sem „tölvuna" í þessu samhengi sé hreint út sagt villandi. Tölva er vissulega meira en sá hluti hennar sem þýðir tákn á milli tungumála, heldur er hún líka kóðinn og reglurnar sem hún starfar eftir, hún er líka allur búnaðurinn sem gerir það að verkum að tölvan geti þýtt táknin. Þess vegna telur Crane að þessi samanburður gangi ekki upp. Tölvan sjálf hlýtur að vera manneskjan + herbergið + reglubókin + blöðin með táknunum.[10] Reglubókin er gagnagrunnur upplýsinga, manneskjan er sá hluti tölvunnar sem vinnur með upplýsingarnar. Ég er sammála Crane, ég tel að rangt sé að skilgreina hugsun eftir því hvernig einhver einn ákveðinn hluti hugans skilur viðfangsefnið. Ég tel að hugsun sé samband þessara ákveðinna hluta hugans.

Ef við berum dæmi Searle saman við mannsheila þá virðist það ganga upp. Hver hluti heilans hefur sitt hlutverk við að stjórna virkni líkamans. T.a.m. er ennisblaðið (lobus frontalis) sá hluti heilans sem sendir boð um hreyfingu til vöðva líkamans og stjórnar tali og samræðum. Hversu mikið skilur þessi tiltekni hluti heilans? Skilur hann þau taugaboð sem hann er að senda? Hann sendir skilaboð um hreyfingar og sendir þær áfram en upplýsingarnar sem hann sendir um lærðar hreyfingar eru geymdar í litla heilanum (cerebellum).[11] Hægt er að bera ennisblaðið saman við manneskjuna í kínverska herberginu og litla heilann við reglubókina eða upplýsingabankann, litli heilinn geymir upplýsingarnar um hreyfingar sem ennisblaðið vinnur úr og sendir áfram til útlimanna rétt eins og reglubókin og manneskjan í dæmi Searle. Myndi Searle líta á einstakan hluta heilans eins og ennisblaðið og segja „þessi hluti heilans vinnur úr upplýsingum án þess að skilja þær, þar af leiðandi getur varla verið að heilinn skilji nokkurn skapaðan hlut"? Ég held ekki. Þess vegna er ég hikandi við að samþykkja þessa samlíkingu Searle. Ég tel að það að skilgreina manneskjuna í

[10] Tim Crane, *The Mechanical Mind* (London: Routledge, 2003), 125.
[11] Valtýr Stefánsson Thors, „Hvernig starfar mannsheilinn? Hverjar eru helstu heilastöðvarnar?", Vísindavefurinn http://www.visindavefur.is/svar.php?id=999.

herberginu sem tölvuna sjálfa sé villandi og betra sé að skilgreina tölvuna sem sjálft herbergið og allt þar inni. Í samhengi við mannsheilann virðist fráleitt að líta á einhvern einn ákveðinn hluta hans og dæma þannig heildina.

Tim Crane - Er heili eins og tölva?

Hér er kannski gott að líta á hvernig, eða hvort, mannleg hugsun líkist því hvernig tölva starfar. Í áðurnefndri bók skilgreinir Tim Crane tölvu sem „hlut/tæki sem meðhöndlar tákn samkvæmt reglum" (e. „A computer is a device that processes representations according to rules")[12]. Hvað á hann við með þessu? Ef við ætlum að líta svo á að tölva sé að framkvæma útreikninga þurfa tölurnar sem hún meðhöndlar að vera tákn (e. representations) fyrir eitthvað. Hann nefnir sem dæmi annað hreyfilögmál Newtons, táknuð með formúlunni $F=ma$, (kraftur = massi margfaldaður með hröðun). Þetta lögmál lýsir náttúrulegu fyrirbæri sem er stöðugt að gerast allt í kring um okkur. En þegar náttúrulegir hlutir verka eftir formúlu Newtons þá þýðir það ekki að hlutirnir séu að framkvæma útreikninga, heldur þvert á móti *lýsir* lögmálið því hvernig náttúrulegir hlutir verka. Þegar kraftur, massi og hröðun mætast í náttúrunni er því ekki um útreikning á lögmáli Newtons að ræða, heldur *tilfelli* af lögmáli Newtons. Táknin 'kraftur' (F), 'massi' (m), og 'hröðun'(a) eru því einungis tákn, sem standa fyrir náttúruleg fyrirbæri.

Tölvur vinna með slík tákn eða myndbirtingar, þær vinna eftir forritunartungumáli þar sem hvert orð táknar eitthvað. Þetta er sambærilegt hugmynd Searle um að tölva hafi tákn en ekki merkinguna á bak við þau. Crane er í rauninni alls ekki að andmæla þeirri hugmynd, merking táknanna er vissulega ekki fólgin í táknunum sjálfum. En hann telur samt sem áður að ef maður sem hefur þekkingu á táknunum lætur hinn ytri heim hafa einhvern áhrif á sig þá gæti einhvers konar skilningur eða merking byrjað að myndast.[13]

Nú byrjar Crane að athuga hvernig mannshugurinn er eins og tölva. Hann hefur á þessum stað í bókinni fjallað um hina vélrænu kenningu um hugann (e. *the mechanical view of mind)*, en samkvæmt kenningu þeirri er hugurinn hluti af

[12] Crane, *The Mechanical Mind,* 121.
[13] Crane, *The Mechanical Mind,* 128.

náttúrunni og starfsemi hans byggir á orsökum og afleiðingum, líkt og allir aðrir hlutir. Hins vegar sé annað og meira að fullyrða að hugurinn starfi eins og tölva, að hugsun sé ekkert nema útreikningur. Crane telur þó að a.m.k. einhver hluti af mannlegri hugsun sé útreikningur af svipuðum toga og tölvur framkvæma; meðhöndlun á táknum samkvæmt reglum. Þetta viðhorf er frábrugðið því viðhorfi að eitthvað geti hugsað bara með því að vera einhvers konar tölva. Crane telur að jafnvel þótt við neituðum því alfarið að eitthvað geti hugsað með því einu að framkvæma útreikninga gætum við samt trúað því að hugsanir okkar hafi nokkurs konar stærðfræðilegan grundvöll.[14] Crane er þó ekki fullkomlega sannfærður að svo sé. Hann telur að ef við ætlum að líta svo á að heilinn starfi eins og tölva þá verðum við einnig að trúa því að heilinn notist við tákn (e. *representations*) líkt og tölvur gera. Hann fjallar um nokkrar tilgátur um mannshugann, t.d. tilgátu Jerry Fodor að heilinn hafi sitt eigið tungumál, Mentalese (sem við getum hér kallað *huglensku)*, og að hvert orð á því tungumáli sé tákn líkt og táknin sem tölvur fást við.[15] Hins vegar telur hann að slíkar kenningar verði aldrei sannreyndar með heimspekinni einni saman, heldur verði vísindin að útkljá hvernig mannsheilinn raunverulega starfar. Við verðum þó að bíða í langan tíma eftir að þess háttar vísindi líti dagsins ljós.[16]

Ég legg til að hér gerum við hlé á spurningum um hvort heili starfi eins og tölva eða tölva starfi eins og heili. Lokaniðurstaða Crane í bókinni er óákveðin. Á síðustu síðunni segir hann að mögulega ættum við að sætta okkur við að við getum ekki haft vísindalega þekkingu á hugtökum eins og meðvitund og hugsun, þótt við getum skilið þau á þeirra eigin grundvelli.[17] Ég tel að á meðan við höfum ekki vísindalegan skilning á þessum hugtökum þá getum við ekki sannað að tölva sé eins og heili og sé raunverulega fær um hugsun. Því ætla ég nú að fjalla í staðinn um aðra mælistiku sem við höfum til að mæla hvort einhver eða eitthvað geti hugsað. Sú mælistika verður til með því að vera mennskur og búa í mannlegu samfélagi. Í staðinn fyrir að reyna að komast að því hvort það sé einhver *vísindalegur* grundvöllur fyrir því að tölvur geti hugsað en ekki einungis líkt eftir hugsun, ættum við ef til vill að skoða hvort okkur

[14] Crane, *The Mechanical Mind,* 128-130.
[15] Crane, *The Mechanical Mind,* 134-148.
[16] Crane, *The Mechanical Mind,* 167.
[17] Crane, *The Mechanical Mind,* 231.

finnist tölvan vera *sannfærandi* sem persóna. Það er einmitt grundvöllur kenningar Alans Turing um eftirhermuleikinn, sem verður viðfangsefni næsta kafla.

Turing og eftirhermuleikurinn

Englendingurinn Alan Turing skrifaði árið 1950 áhrifaríka grein, „Computing Machinery and Intelligence"[18], þar sem hann fullyrti að það væri tilgangslaust að ræða hvort vélar gætu hugsað eða ekki. Hans afstaða er sprottin vegna þess að til þess að svara slíkri spurningu þyrfti að skilgreina hugtökin vélar og hugsun, en slíkt væri ómögulegt. Í staðinn lagði hann til svokallaðan „eftirhermuleik" (e. *imitation game*) sem gæti gefið okkur einhverja hugmynd um greind tölvuforrita. Leikurinn er spilaður af manneskju, tölvu og spyrli. Spyrillinn er í öðru herbergi en hin manneskjan og tölvan, og á að reyna að giska hvor aðilinn er tölva og hvor er manneskja. Hann sér einstaklingana ekki og heyrir ekki raddir þeirra, þekkir þau aðeins með nöfnunum X og Y og spyr þá spurninga. Markmið tölvunnar er að fá spyrilinn til að halda að hún sé sjálf manneskjan, og markmið manneskjunnar er að fá spyrilinn til að komast að réttri niðurstöðu. Þessi eftirhermuleikur, sem er oft kallaður „Turing-prófið," er gagnlegur vegna þess að hann dregur skýr mörk á milli greindar og líkamlegrar getu mannsins. Spurningarnar eru af margs konar toga þar sem spyrlinum leyfist t.d. að spyrja um líkama beggja aðila. Hins vegar mega aðilarnir ljúga í svörum sínum og svara spurningunum á afar frjálslegan máta. Tölvan má t.d. fullyrða að hún sé raunverulega manneskjan en slíkt skilar litlum árangri þar sem manneskjan getur einnig komið með slíkar fullyrðingar. Spyrillinn getur t.d. spurt báða aðila hversu sítt hár þeirra sé, og tölvan getur logið til um svar, en spyrillinn getur ekki krafist líkamlegrar staðfestingar (t.d. að fá að sjá hárið á þátttakendunum o.s.frv.). Turing var sjálfur sannfærður um það að í náinni framtíð yrði til tölvuforrit sem sem gæti staðist prófið, en hann skrifar sjálfur í grein sinni: „Ég tel að eftir um fimmtíu ár verði mögulegt að forrita tölvur [...] til að láta þær spila eftirhermuleikinn svo vel að meðal

[18] Alan Turing, „Computing Machinery and Intelligence," *The Mind's I,* ritstj. Hofstadter, Douglas R. og Dennett, Daniel C, 57.

spyrill mun ekki hafa meira en 70 prósent líkur á að bera rétt kennsl eftir fimm mínútna yfirheyrslu".[19]

Síðar í greininni fer Turing að fjalla um möguleg mótrök gegn kenningu sinni. Ein þau sterkustu eru meðvitundarrökin, sem lýsa sér best í tilvitnun sem Turing sjálfur vitnar í frá Geoffrey Jefferson:

> Meðan vél getur ekki skrifað sonnettu eða samið konsertverk vegna hugsana og tilfinninga, en ekki vegna tilviljanakenndrar röðunar á táknum, getum við ekki fallist á að vél samsvari heila – það er að segja, ekki einungis samið það heldur vita að það hafið samið það. Enginn vélbúnaður gæti fundið fyrir (ekki einungis líkt eftir, sem er auðvelt) ánægju þegar hann nær árangri, harmi þegar lampar hans springa, glaðst yfir hrósi, hryggst yfir mistökum, heillast af kynlífi, orðið reiður eða dapur þegar hann fær ekki það sem hann vill.[20]

Í þessum texta má finna margar ólíkar hugmyndir, og sú fyrsta sem Turing fjallar um er að röksemdafærslur á þessa leið eru afneitun á gildi Turing-prófsins. Hugmyndin virðist vera á þá leið að tölvur geti ekki gert þá hluti sem greindir menn geta gert, svo sem að skrifa sonnettu, og þar af leiðandi geti tölvur ekki verið greindar og haft tilfinningar. Turing gagnrýnir hugmyndir á borð við þessa
Turing gagnrýnir þessa hugmynd á þá leið að þetta sé sjálfsveruhyggja (e. solipsism), enda sé einnig hægt að fullyrða út frá þessu að eina leiðin til þess að vita að önnur manneskja hugsi sé að vera sú manneskja. Hann á ekki við að það sé rökrétt að halda slíku fram, heldur er hann að sýna fram á að slíkt viðhorf getur ekki sannfært okkur um að tölvur geti ekki haft ekki greind á við menn; það er einnig hægt að efast um að annað fólk hugsi.

Ég er sammála Turing. Það sem hann sýndi fram á var einfaldlega það að yfirborðið er okkar eini mælikvarði á hvort einhver eða eitthvað geti hugsað. Hugsun er of flókið fyrirbæri til þess að hægt sé að skilgreina það með fullnægjandi hætti, og það eina sem við getum gert til að komast að því hvort tölvur hugsi er að athuga hvort þær virðist hugsa. Heilar okkar eru ólíkir en þeir framkvæma allir að vissu marki sömu hugsanamynstur, mynstur sem við teljum vera mannleg. Ég tek því engu að

[19] Alan Turing, „Computing Machinery and Intelligence," *The Mind's I,* ritstj. Hofstadter, Douglas R. og Dennett, Daniel C, 57.
[20] Turing, „Computing Machinery and Intelligence," 60.

síður sem gefnu að heilar í öðru fólki séu virkilega að hugsa, vegna þess að fyrir mér virðast þeir hugsa. Þess vegna er ég hikandi við að trúa að tölvuheili sem virðist framkvæma sömu hugsanamynstur og mennskir heilar sé einungis að líkja eftir hugsun. Ef hann líkir nægilega vel eftir mannlegri hugsun svo að við erum sannfærð að um manneskju sé að ræða, er það þá ekki eini mælikvarðinn sem við þurfum?

Vandamál sjálfsveruhyggjunnar

Nú er upplagt að fjalla um eina þekktustu hugsanatilraun heimspekisögunnar, sem birtist í *Hugleiðingum úr frumspeki* eftir René Descartes frá 1641. Í hugleiðingum þeim ákveður hann að rífa niður allar sínar skoðanir til grunna og efast um allt sem hann getur mögulega efast um. Hann efast um það sem skynfæri hans segja honum, því þau eru oft villandi. Hann getur einnig efast um að allir hlutir í kring um hann og jafnvel hans eigin líkami séu yfir höfuð til. Hann skoðar hugmyndina að hugsanlega sé máttugur illur andi að blekkja hann og villa fyrir honum sýn, og allt það sem hann sér og skynjar séu tómar draumsýnir, falsaður veruleiki. (Síðari heimspekingar hafa gjarnan ímyndað sér að þeir séu heilar í krukku tengdir við tölvubúnað. Slíkar hugmyndir má finna í fjölmörgum kvikmyndum, einkum *The Matrix* frá 1999[21]). Nú reynir Descartes að finna einhvern fastan punkt í veröldinni, eitthvað eitt sem hann getur ekki efast um sama hversu mikið hann reynir. Hann kemst að lokum að því að fyrst hann er að hugsa til að byrja með, þá hlýtur hann sjálfur að vera til. Sama hversu mikið hinn illi andi geti blekkt hann um hinn ytri heim þá geti hann aldrei gert hann að engu. „Staðhæfingin *Ég er, ég er til* hljóti að vera sönn, hvenær sem [Descartes] segi[r] hana eða hugsa[r]"[22] Hugmynd Descartes, að hægt sé að efast um allt í veröldinni fyrir utan mann sjálfan, nefnist sjálfsveruhyggja eða sólipsismi.[23] Sjálfveruhyggjusinnar trúa því að það eina sem hægt er að hafa þekkingu á er manns eigin hugur. En það eru samt ekki margir heimspekingar sem (vilja) aðhyllast þessa

[21] Lance P. Hickey, „The Brain in a Vat Argument," Internet Encyclopedia of Philosophy http://www.iep.utm.edu/brainvat/.
[22] René Descartes, *Hugleiðingar um frumspeki* (Reykjavík: Hið íslenska bókmenntafélag, 2001), 142.
[23] Stephen P. Thornton, „Solipsism and the Problem of Other Minds", Internet Encyclopedia of Philosophy http://www.iep.utm.edu/solipsis/.

hugmyndafræði. Flestar heimspekigreinar um sjálfveruhyggju fjalla í rauninni um tillögur um hvernig hægt sé að hrekja hana. Descartes reynir sjálfur í síðari hugleiðingum í bók sinni að hrekja hana[24], þó með umdeildum hætti að mati heimspekinga nútímans. Sjálfveruhyggja er enn lifandi vandamál í heimspeki, og á vel við vandamálið sem við fáumst hér við. Vandamálið sem við fáumst hér við, hvort tölvur með gervigreind ættu að hafa réttindi, er vandamál sem einkennist af vandamálum sjálfsveruhyggjunnar. Ein gerð sjálfsveruhyggju er sú að við getum hugsanlega skynjað hinn ytri heim en samt ekki haft þekkingu á öðrum hugum en okkar eigin. Við höfum hér áður fullyrt að ef tölva eigi að hafa tilkall til réttinda þurfi hún að hafa tilfinningar, meðvitund og getuna til að hugsa. En vandamálið sem sjálfsveruhyggjan sýnir okkur er að við *getum ekki vitað* hvort hún hafi þessa eiginleika. Vandamálið sem sjálfsveruhyggjan gerir okkur má setja upp svona:

1. Ef tölva á að eiga tilkall til réttinda, þarf hún að hafa tilfinningar, meðvitund og getuna til að hugsa.
2. Við getum ekki vitað hvort tölva hafi tilfinningar, meðvitund og getuna til að hugsa.
3. Við getum ekki vitað hvort tölva eigi að hafa tilkall til réttinda.

Ég tel að þessi röksemdafærsla gangi upp, ef við trúum því að forsendurnar séu sannar þurfum við að fallast á að niðurstaðan sé sönn. Ég hef hér á undan fært rök fyrir því að tilfinningar, meðvitund og getan til að hugsa séu allt nauðsynleg skilyrði fyrir því að eiga tilkall til réttinda. Við tökum því hér sem forsendu eitt. Forsenda tvö segir að við getum ekki vitað hvort tölva hafi tilfinningar, meðvitund og getuna til að hugsa. Sú fullyrðing er gerð út frá sjálfsveruhyggjunni, vegna þess að við getum aldrei mögulega haft fullkomlega þekkingu á einhverju fyrir utan okkar eigin huga, okkar fyrstu persónu frásögn. Hins vegar er það þessi forsenda, forsenda tvö, sem Turing efast um og ég ætla að leyfa mér að efast um hér. Ég tel nefnilega, líkt og Turing, að við *höfum* leið til þess að vita hvort eitthvað sé að hugsa, hafi tilfinningar og meðvitund. Sú leið er ekki vísindaleg, þótt Turing-prófið sé e.t.v. „vísindalegasta“

[24] Descartes, *Hugleiðingar um frumspeki,* 143-223.

aðferðin sem við höfum til að mæla hvort eitthvað hafi þessa eiginleika. Í grunninn er þó kvarðinn sem við notum mannlegur. Hann byggist á því hverju við höfum vanist í samskiptum við annað fólk frá blautu barnsbeini.

Hvert er gervigreindartækni komin í dag?

Gervigreind er til í dag í alls konar myndum. Gervigreind býr til sérsniðinn lagalista sérstaklega fyrir mig á hverjum degi í tónlistarforritinu Spotify. Gervigreind er notuð í gervihnöttum (líkt og Deep Space One, sem fjallað var um áðan). Einnig má nefna spjallforrit líkt og Cleverbot, sem er gervigreindarforrit á netinu sem talar við notandann líkt og um manneskju væri að ræða. Cleverbot notast ekki við fyrirfram forrituð svör, heldur „lærir“ það af svörum annarra notenda og notar þau í sínum eigin svörum. Þegar notandi spyr Cleverbot spurninga leitar það í gagnagrunni sínum eftir því hvernig aðrir notendur hafi svarað slíkum spurningum áður. Forritið var þróað af vísindamanninum Rollo Carpenter árið 1997, og árið 2011 tók það þátt í raunverulegu Turing-prófi á tæknisamkomu á Indlandi. Prófið var ekki alveg með sama sniði og Turing sjálfur lýsti því, en í prófinu áttu þrjátíu sjálfboðaliðar fjögurra mínútna samtal við óþekktan aðila. Helmingur þátttakenda talaði við mannfólk meðan hinn helmingurinn talaði við Cleverbot. Áhorfendur fylgdust með samræðunum og bæði áhorfendur og þátttakendur greiddu atkvæði um hvort það væri tölva eða raunveruleg manneskja á ferðinni. Niðurstöðurnar voru þær að Cleverbot taldist 59,3% mennskt, og manneskjurnar einungis 63,3%.[25]

Einungis eitt gervigreindarforrit hefur raunverulega staðist Turing-prófið, en það gerði forritið Eugene Goostman árið 2014. Forritið hermir eftir 13 ára gömlum dreng frá Úkraínu og var forritað í Sankti Pétursborg í Rússlandi. Goostman sannfærði 33% dómaranna að um raunverulega manneskju væri að ræða, sem er meira en þarf til að standast prófið (þar sem Turing sjálfur sagði í spádómi sínum að spyrillinn myndi ekki hafa meira en 70% líkur á að giska rétt). Goostman var þróaður með það markmið að hann hefði trúverðulegan persónuleika, hann væri 13 ára strákur

[25] Jacob Aron, „Software tricks people into thinking it is human“, New Scientist https://www.newscientist.com/article/dn20865-software-tricks-people-into-thinking-it-is-human?DCMP=OTC-rss&nsref=online-news.

sem segist vita allt en veit augljóslega ekki næstum því allt. Svo virðist vera að spádómur Turing frá 1950, að eftir u.þ.b. 50 ár myndi vera til gervigreindarforrit sem gæti sannfært 30% spyrla í eftirhermuleiknum, hafi ræst eftir allt saman. Það tók aðeins 14 árum lengur en hann hafði spáð fyrir.

Það eru meira að segja til gervigreindarforrit í dag sem eru fær um að skapa list, eins og forritið Flow Machines frá tæknifyrirtækinu Sony. Forritið hefur aðgang að gríðarstórum gagnagrunni af tónlist, greinir mynstur úr þeim og notar þau til að „semja" eigin tónlist. Franski tónlistarmaðurinn Benoît Carré lét forritið greina gagnagrunn af Bítlalögum og búa til lag í svipuðum stíl og lög Bítlanna. Lagið, sem heitir „Daddy's Car," hljómar óneitanlega keimlíkt Bítlunum og er alls ekki afleit lagasmíð. Nú er því hægt að velta fyrir sér hvort þeir sem telja að listrænir hæfileikar séu einungis mannlegt fyrirbæri sem ekki verður líkt eftir með tölvubúnaði eigi að endurskoða afstöðu sína. En ég tel hins vegar að þetta dæmi sé ekki nægilegt til að hrekja þá skoðun. Í áðurvitnuðum texta eftir Jefferson, þeim sem Turing gagnrýndi, segir að við getum ekki fallist á að vélar geti hugsað ef þær semja ekki list út frá eigin hugsunum og tilfinningum. Staðreyndin er sú að forritið tók ekki sjálfstæða ákvörðun um að semja lagið vegna þess að því langaði til þess, heldur fékk það skipun að greina gagnagrunn af tónlist og búa til lag út frá því.
Einnig er lagið að stórum hluta samið af Carré, því það eina sem forritið skilaði frá sér voru hljómagangurinn og laglínan á nótnaformi. Textinn, hljóðfæraleikurinn, söngurinn og öll þau listrænu stílbrigði sem heyrast í laginu eru afurð Carrés sjálfs, en ekki forritsins.[26]

Gervigreind á þó enn langt í land og ljóst er að sú tegund gervigreindar sem við höfum fjallað um hér er enn tómur vísindaskáldskapur. Til að mynda neyddist Microsoft til að loka Twitter-aðgangi hjá gervigreindarforriti sínu eftir innan við sólarhring í notkun árið 2016. Forritið hafði verið þróað í líki unglingsstúlku að nafni Tay á samskiptamiðlinum Twitter og átti að „læra" af því sem aðrir notendur skrifuðu og skrifa sín eigin tíst (e.tweets) út frá því. Einhverjir óprúttnir aðilar ákváðu að hrekkja forritið með því að fá það til að læra alls konar fordómafullt slangur, og ekki leið á löngu þar til forritið var farið að skrifa tíst sem þóttu lýsa yfir nasisma og gríðarlegum kynþáttafordómum. Í einu tísti lýsti hún yfir stuðningi við Adolf Hitler, í

[26] Lucy Jordan, „Inside the Lab That's Producing the First AI-Generated Pop Album," Seeker https://www.seeker.com/tech/artificial-intelligence/inside-flow-machines-the-lab-thats-composing-the-first-ai-generated-pop-album.

öðru kallaði hún Barack Obama Bandaríkjaforseta „apakött" og í enn öðru lýsti hún yfir hatri sínu á gyðingum.[27]

Niðurstöður – eiga tölvur að fá mannréttindi?

Nú vil ég aftur fjalla um spurninguna sem ritgerð þessi byrjaði á að ræða; ættu tölvur sem búa yfir gervigreind að hafa réttindi? Hér á undan hef ég rætt forsendur þess að eiga tilkall til mannréttinda. Ég komst að þeirri ályktun að til þess þarf að hafa tilfinningar, meðvitund og getuna til að hugsa. Síðan ræddi ég um hvort tölvur geti hugsanlega búið yfir þessum eiginleikum. Ég fjallaði um tvær áhrifaríkar hugmyndir sem tengjast þessum spurningum; hugsanatilraun John Searle um Kínverska Herbergið annars vegar og Eftirhermuleik Alans Turing hins vegar. Searle er þeirrar skoðunar að tölvur gætu aldrei hugsað heldur einungis líkt eftir hugsun. Hann telur að þær búi yfir setningarfræði en ekki merkingarfræði, og þær geti aldrei myndað neina merkingu úr táknunum sem þær vinna með. Hann líkti tölvu við manneskju í herbergi sem flytur tákn á milli staða eftir leiðbeiningum í reglubók án þess að tengja nokkurs konar merkingu við táknin sem hún flytur. Síðar fjallaði ég um hugmynd Turings, en samkvæmt henni er okkar eini mælikvarði til að mæla hvort tölva geti hugsað sá sami og við beitum til að mæla hvort annað fólk geti hugsað: samskipti. Samkvæmt Turing þá er engin leið fyrir okkur til að vita hvort annað fólk geti hugsað, sérhver einstaklingur getur hæglega haldið því fram að hann sé sá eini í heiminum sem hefur raunverulegar hugsanir og allir aðrir eru einungis að líkja eftir hugsun. Hins vegar er venjan að halda því fram að allir hugsi í raun og veru. Turing sagði að að sama skapi ætti okkar mælikvarði á greind vitvéla að vera sá sami og við notum á annað fólk, og bjó hann til svokallaðan eftirhermuleik þar sem þátttakandi átti að eiga samræður við tvær verur með textaskilaboðum, og önnur þeirra var manneskja en hin tölva. Ef þátttakendurnir töldu að tölvan væri raunverulega manneskjan í yfir 30% tilvika, þá teldist tölvan hafa staðist prófið.

[27] Andrew Griffin, „Tay Tweets: Microsoft Shuts Down AI Chatbot Turned Into a Pro-Hitler Racist Troll in Just 24 Hours," Independent http://www.independent.co.uk/life-style/gadgets-and-tech/news/tay-tweets-microsoft-ai-chatbot-posts-racist-messages-about-loving-hitler-and-hating-jews-a6949926.html.

Ég tel hugmynd Turings vera afburðasnjalla, og mér finnst skynsamlegt að taka þá afstöðu að meðan við höfum ekki nákvæman mælikvarða til að mæla hvort tölvur hugsi ættum við frekar að ræða við tölvuna og spyrja spurninga á borð við: „hvernig er *tilfinningin* að ræða við tölvuna? Er þetta eins og að ræða við manneskju?" . Eðlilegasti mælikvarðinn er okkar *upplifun* af því að ræða við tölvuna.
En aftur að upphaflegu spurningunni, eiga tölvur með gervigreind að fá réttindi? Ég ætla ekki að komast að neinni ákveðinni niðurstöðu hér, en ég mun hins vegar færa rök fyrir því að það sé skynsamlegra að álykta að vitvélar hafi hin nauðsynlegu skilyrði til að eiga tilkall til réttinda, og að þær ættu þar af leiðandi að fá þau.
Ég tel að það skipti hreinlega ekki máli hvort við getum fullkomlega vitað hvort tölva hafi tilfinningar, meðvitund eða getuna til að hugsa. Okkar eini mælikvarði er á yfirborðinu; yfirborðið er allt sem við höfum. Ég tel því að ef við höfum tölvu sem *virðist* uppfylla skilyrðin sem ég hef sett fyrir því að eiga tilkall til réttinda þá ætti slík tölva að hafa tilkall til þeirra. Því má ímynda sér að lagt verði eins konar Turing-próf fyrir gervigreindarvélar framtíðarinnar sem dæmir tilkall þeirra til réttinda. Hins vegar get ég ekki fallist á að Turing-prófið í sinni upprunalegu mynd sé nægilegt til að tölva hafi tilkall til mannréttinda. Eins og nefnt var áðan stóðst gervigreindarforritið Eugene Goostman Turing-prófið með því að plata meira en 30% þátttakanda að það væri hin raunverulega manneskja. En þegar ég les viðtal sem blaðamaður Time Magazine tók við Goostman tel ég frekar augljóst að ekki sé um raunverulega manneskju að ræða. Hér er brot úr viðtalinu, þýtt á íslensku:

> **Blaðamaður**: Ég bjó í Boston en ólst upp í Minnesota. Ég bjó líka í Seattle í einhvern tíma.
> **Goostman**: Gætir þú boðið mér til Seattle? Ég væri afar þakklátur :-)
> **Blaðamaður**: Ég bý ekki þar lengur og mér þætti furðulegt að hýsa 13 ára dreng hvar sem ég byggi.
> **Goostman**: Mögulega, það mun koma þér á óvart, en „nei" er ekki eina svarið. Orðið „já" er líka til! :-) Voltu spyrja mig að meiru? [Orðið „viltu" er hér vitlaust skrifað, en þetta er lausleg þýðing á orðinu „wonna" sem Eugene notar og er þar augljóslega að skrifa orðið „wanna" vitlaust].
> **Blaðamaður**: Þetta síðasta svar var örlítið undarlegt, félagi.
> **Goostman**: Fjárinn! Ég var að muna að ég gleymdi að fóðra naggrísinn minn í morgun. Greyið dýrið!

Blaðamaður: Ah, ég skil. Við erum nú að skipta um umræðuefni.[28]

Þótt um skemmtilega lesningu sé að ræða flæðir þetta ekki beint eins og eðlilegar samræður. Það var sniðugt af hálfu forritaranna að láta Goostman gera stafsetningarvillu til að láta hann virðast meira eins og venjulega manneskju. En ég tel að það sé grundvallarmunur á gervigreind Goostman og þess háttar gervigreind sem við fáumst hér við. Ég tel að sama hversu oft Goostman tali við annað fólk muni hann ekki *læra* eins og venjuleg manneskja gerir, einstaklingseðli hans mun ekki breytast. T.a.m. spyr hann blaðamanninn hvert starfið hans sé tvisvar sinnum í viðtalinu. Mér finnst skiljanlegt að forritið hafi náð að gabba 33% þátttakenda, samtalið í greininni flæðir (að einhverju leyti) nokkuð eðlilega og það sem ekki eðlilegt þykir gæti í samhengi Turing-prófsins hæglega verið manneskja að þykjast vera tölva. En ég tel Goostman ekki uppfylla skilyrðin sem þarf að uppfylla til að eiga tilkall til mannréttinda. Til þess þarf eitthvað miklu meira sannfærandi en 33% einkunn á Turing-prófi. Ég tel að til þess að eiga tilkall til réttinda þurfi tölva að vera með svo sannfærandi gervigreind að nánast *ómögulegt* sé að skera úr hvort um manneskju sé að ræða eða ekki. Hún myndi hegða sér á hátt sem væri algjörlega ógreinanlegur frá manneskju.

Hugleiðingar um framtíðina

Með allri þeirri þróun sem hefur orðið í gervigreindartækni á síðustu árum virðist heimskulegt að horfa fram hjá því að einhvern tímann gæti tæknin orðin svo góð að erfitt muni reynast að gera upp á milli manneskju og tölvu. Ég tek sem dæmi kvikmyndina *Her* frá 2013 (en gaman er að minnast á að sú kvikmynd var kveikjan að þessu ritgerðarefni). Kvikmynd sú gerist í nálægri framtíð og fjallar um rithöfundinn Theodore sem hleður inn nýju háþróuðu gervigreindarforriti í tölvuna sína. Forritið hefur verið auglýst með slagorðinu „ekki bara stýrikerfi, heldur meðvitund" (e. it's not just an operating system, it's a consciousness). Forritið hefur kvenmannsrödd sem kallar sig Samönthu, og hún og Theodore byrja í ástarsambandi. Við sjáum aldrei andlitið á Samönthu en hún er talsett af leikkonunni Scarlett Johanson, sem hefur

[28] Doug Aamoth, „Interview with Eugene Goostman, the Fake Kid Who Passed the Turing Test," Time http://time.com/2847900/eugene-goostman-turing-test/.

nokkuð hása og „mennska" rödd. Þetta er snjöll leið kvikmyndagerðarfólksins til að láta tölvuna vera raunverulega persónu, en hún er algjörlega sannfærandi og engu síður trúverðug en hinar mennsku persónur myndarinnar. Það er svo gegnumgangandi heimspekilegt þema í myndinni sem snýst að því hvort ástarsamband með tölvu sé einhvern veginn minna virði eða minna raunverulegt en ástarsamband með manneskju.[29]

Einnig er hægt að fjalla um sænsku sjónvarpsþáttaröðina Äkta Människor, en í þeirri þáttaröð hefur vélmennatækni orðið svo fullkomin að nánast engin leið er að gera grein á milli vélmennanna og mannfólksins. Engu að síður er litið á vélmennin sem vörur en ekki einstaklinga og fólk kaupir sér vélmenni til að láta þjóna sér á heimilinu og vinna alls kyns störf sem mannfólk telur yfir sig hafið að vinna. Vélmennin fara svo að berjast fyrir réttindum sínum og líta á sig sem raunverulega einstaklinga en ekki vörur sem ganga kaupum og sölum.[30]

Ef slík tækni verður nokkurn tímann til (og ég tel að með þeirri þróun sem hefur orðið á tækni á síðustu árum sé það ekki ólíklegt) tel ég að það sé afar mikilvægt að við mannfólkið höfum velt fyrir okkur siðferðislegum spurningum á borð við þá sem ég hef hér rætt í ritgerð þessari. Ef til vill kann einhver að spyrja: af hverju ættu tölvur að þurfa réttindi? Hvernig er hægt að brjóta á þeim? Ég tel að það sé alveg hægt að hneppa tölvur í þrældóm líkt og mannfólk var (og er sumstaðar) hneppt í þrældóm. Þær gætu e.t.v. unnið alls konar vinnu sem aðeins mannfólk vinnur við í dag, eins og að skrifa blaðagreinar, svara í símann í þjónustuveri o.s.frv. Ef þær verða tengdar við vélbúnað (og verða að eiginlegum vélmennum) væri hægt að nota þær í líkamlega vinnu (t.d. byggingarvinnu) allan sólarhringinn án pásu.

Í gegnum gjörvalla mannkynssöguna hefur þrælahald verið til. Fólk hefur verið hneppt í þrældóm fyrir það eitt að vera af öðrum kynþætti eða aðhyllast trúarbrögð. Slíkt hefur gjarnan verið réttlætt með rökum á borð við að þeir sem kúgaðir eru séu ekki raunverulega manneskjur. Ég tel að sama hvort við kjósum að trúa því að háþróaðar gervigreindartölvur verði nokkurn tímann færar um ranverulegar hugsanir og tilfinningar, þá sé skynsamlegra að koma fram við þær eins og þær hafi þær. Við getum ímyndað okkur að í framtíðinni séu tveir möguleikar í boði: Fyrri möguleikinn er sá að tölvur með gervigreind búa yfir hugsunum, tilfinningum og meðvitund.

---

[29] Spike Jonze, *Her* (Los Angeles: Warner Bros. Pictures, 2013).
[30] Lars Lundström, *Äkta Manniskor* (Stokkhólmur: Sveriges Television, 2012-2014).

Seinni möguleikinn er að þær búi *ekki* yfir þessum eiginleikum. Ef við kjósum að veita tölvum réttindi veldur það annað hvort tölvunum raunverulegri gleði eða plat-gleði. Ef við kjósum að veita þeim *ekki* réttindi veldur það tölvunum annað hvort harmi og sársauka eða plat-harmi og plat-sársauka. Ég tel að jafnvel þótt við trúum ekki að tölvur geti raunverulega hugsað þá ættum við samt að kjósa að veita þeim réttindi. Ef ekki gætum við verið að valda raunverulegum skaða hjá verum sem finna til raunverulegra tilfinninga. Valið er okkar.

Heimildir

Aamoth, Doug. „Interview with Eugene Goostman, the Fake Kid Who Passed the Turing Test“. Time. Birt 9.júní 2014. Sótt 20.apríl 2017. Vefslóð: http://time.com/2847900/eugene-goostman-turing-test/.

Andrew Griffin. „Tay Tweets: Microsoft Shuts Down AI Chatbot Turned Into a Pro-Hitler Racist Troll in Just 24 Hours“. Independent. Birt 24.mars 2016. Sótt 16.júlí 2016. Vefslóð: http://www.independent.co.uk/life-style/gadgets-and-tech/news/tay-tweets-microsoft-ai-chatbot-posts-racist-messages-about-loving-hitler-and-hating-jews-a6949926.html.

Aron, Jacob. „Software tricks people into thinking it’s human“. New Scientist. Birt 6.september 2011. Sótt 25.júlí 2016. Vefslóð: https://www.newscientist.com/article/dn20865-software-tricks-people-into-thinking-it-is-human?DCMP=OTC-rss&nsref=online-news.

Crane, Tim. *The Mechanical Mind: A philosophical introduction to minds, machines and mental representation*. London: Routledge, 2003.
„Deep Blue.“ *IBM*. Sótt 18.júlí 2016. Vefslóð: http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/.

Descartes, René. *Hugleiðingar um frumspeki*. Þýðandi Þorsteinn Gylfason. Reykjavík: Hið íslenska bókmenntafélag, 1641/2001.

Hickey, Lance P. „The Brain in a Vat Argument“. Internet Encyclopedia of Philosophy. Sótt 1.maí 2017. Vefslóð: http://www.iep.utm.edu/brainvat/.

Jonze, Spike. *Her*. DVD. Leikstjóri Jonze, Spike. Los Angeles: Warner Bros. Pictures, 2013.

Jordan, Lucy. „Inside the Lab That's Producing the First AI-Generated Pop Album”. Seeker. Birt 13.apríl 2017. Sótt 28.apríl 2017. Vefslóð:

https://www.seeker.com/tech/artificial-intelligence/inside-flow-machines-the-lab-thats-composing-the-first-ai-generated-pop-album.

Lundström, Lars. *Äkta Manniskor*. Streymi á internetinu (Netflix). Leikstjórar Hamrell, Harald og Akin, Levan. Stokkhólmur: Sveriges Television, 2012-2014.

Lög um dýravernd. (10. Október 2011). Sótt 24. mars 2017 frá Lagasafn: http://www.althingi.is/lagas/139b/1994015.html.

Mannréttindayfirlýsing Sameinuðu þjóðanna. Sótt 2.febrúar 2017. Vefslóð: http://www.asi.is/media/7575/Mannr_ttindayfirl_sing_Sameinu_u__j__anna.pdf.

Regan, Tom. „The Case for Animal Rights". *Ethics in Practice: An Anthology*. Ritstjóri Lafollette, Hugh. Sussex: Wiley Blackwell, 2014.

Thornton, Stephen P. „Solipsism and the Problem of Other Minds". Internet Encyclopedia of Philosophy. Sótt 1.maí 2017. Vefslóð: http://www.iep.utm.edu/solipsis/.

Turing, Alan. „Computing Machinery and Intelligence." *The Mind's I: Fantasies and Reflections on Self and Soul*. Ritstjórar Hofstadter, Douglas R. og Dennett, Daniel C. London: Penguin books, 1982, 53-67.

Valtýr Stefánsson Thors. „Hvernig starfar mannsheilinn? Hverjar eru helstu heilastöðvarnar?" *Vísindavefurinn*, 17. október 2000. Sótt 15.nóvember 2016. Vefslóð: http://visindavefur.is/svar.php?id=999.

„Víðtæk mannréttindabrot í Norður-Kóreu". Rúv. Skrifað 21.apríl 2012. Sótt 28.apríl 2017. Vefslóð: http://www.ruv.is/frett/vidtaek-mannrettindabrot-i-n-koreu.