Háskóli Íslands

Hugvísindasvið

Heimspeki

Hin vitiborna vél

Er viðeigandi að gervigreind sinni öllum mannanna verkum eða er hún takmörkunum háð?

Ritgerð til BA-prófs í heimspeki

Jakob Sindri Þórsson

Kt.: 130291-2879

Leiðbeinandi: Salvör Nordal

Janúar 2017

Ágrip

„Hugsandi“ vélar hafa fremur verið viðfangsefni skáldskapar en vísinda. Á því hefur orðið breyting. Nú keppast tölvunarfræðingar við að hanna öflug gervigreindarforrit sem sýna fram á afar mikla færni í aðskiljanlegustu verkefnum. Tilhneiging er til að álíta að hin vitræna hegðun þessarra forrita sé merki um vitsmuni og að vél geti búið yfir hugarástandi. Meðal annars þess vegna telja ýmsir fræðimenn að henni séu ekki takmörk sett umfram menn. Andstæð viðhorf gera því skóna að enginn eðlismunur sé á færustu gervigreindarforritum og einföldum tölvuforritum. Gervigreind skilji ekki aðgerðir sínar og sé ófær um að bera ábyrgð. Aðgerðir og aðgerðarleysi hennar séu einungis röklegar lausnir tölvu á tilteknu vandamáli. Með hliðsjón af þeim rökum skýrist hvers vegna varhugavert geti verið að fela gervigreind verkefni sem hún kann fræðilega og raunverulega að geta leyst af hendi en krefjast mannlegs skilnings og ábyrgðar svo að viðunandi sé.

Að lokinni umfjöllun um skilgreiningar og hugmyndir ýmissa fræðimanna um gervigreind, möguleika hennar og takmarkanir, er skoðað hvort og hvernig hún getur nýst á þremur mikilvægum sviðum samfélags manna, þ.e. í réttarkerfinu, heilbrigðisþjónustunni og hernaði. Er leitast við að vega og meta andstæð sjónarmið til að komast til botns í hvenær gervigreind geti og eigi að sinna verkefnum án atbeina mannlegrar dómgreindar.

Efnisyfirlit

Inngangur

Síðustu ár og áratugi hefur tækniþróun verið mjög ör. Miklar tæknibyltingar hafa átt sér stað sem hafa haft mikil áhrif á líf og störf manna. Ein slík tæknibylting felst í skjótri framför gervigreindar, þ.e. véla sem með mikilli einföldun má segja að hegði sér eins og hugsandi vera.

Helstu markmiðin með þróun gervigreindarforrita er að láta þau ná æ meiri færni í afmörkuðum verkefnum eða leikjum. Til dæmis tókst skáktölvunni Deep Blue að sigra Gary Kasparov, þáverandi heimsmeista, í einvígi árið 1997.[1] Gervigreindarforrit sinna nú orðið ýmsum verkefnum í daglegu lífi margra. Notkun forrita eins og t.a.m. Siri verður sífellt algengari, en Siri er hjálparhella í iPhone snjallsímum Apple fyrirtækisins og aðstoðar við ýmis verk sé hún beðin um það, s.s. að skipuleggja fund eða stefnumót.[2] Sumir fræði- og vísindamenn telja að gervigreind geti og muni geta sinnt ýmsum verkefnum sem við höfum hingað til talið aðeins vera á færi manna. Í einhverjum tilvikum eru slíkar spár mjög líklegar sé tekið mið af þróuninni sem hefur átt sér stað. Á hinn bóginn er nauðsynlegt að greina á milli þeirra verkefna sem gervigreind getur unnið og þeirra sem gervigreind ætti að vinna.

Í þessari ritgerð mun ég færa rök fyrir því að engan eðlislægan greinarmun sé að finna á færustu gervigreindarforritum og öðrum tölvuforritum sem eru einfaldari að allri gerð. Með hliðsjón af þeim rökum skýrist hvers vegna varhugavert er að fela gervigreind verkefni sem krefjast mannlegs skilnings og ábyrgðar þótt hún geti fræðilega eða raunverulega sinnt þeim. Í þessu ljósi standast ýmsar fullyrðingar um getu og möguleika gervigreindar ekki. Dæmi um slíkar fullyrðingar má finna hjá Kenneth M. Colby, James B. Watt og John P. Gilbert sem héldu fram að gervigreind gæti leyst sálfræðinga af hólmi.[3] Annað dæmi er að finna hjá Thomas W. Simpson og Vincent C. Müller sem færa rök fyrir því að það sé siðferðilega réttlátt að beita sjálfstýrðum vígvélum sem ráðast á óvini og óvinveittar herbúðir í hernaði án beinna fyrirmæla frá mönnum.[4]

---

[1] Kristinn R. Þórisson, „Vélvitund, meðvitund og sjálfsvitund í kjötvélum og vélmennum," *Veit efnið af andanum? Sjö fyrirlestrar um meðvitundina*, ritstj. Steinar Örn Atlason og Þórdís Helgadóttir (Reykjavík: Heimspekistofnun Háskóla Íslands, 2009), 129.
[2] Natalie Harrison og Teresa Brewer, „Apple Launches iPhone 4S, iOS 5 & iCloud," heimasíða Apple Inc. https://www.apple.com/pr/library/2011/10/04Apple-Launches-iPhone-4S-iOS-5-iCloud.html.
[3] Kenneth Mark Colby, James B. Watt og John P. Gilbert, „A Computer Method of Psychotherapy: Preliminary Communication," *The Journal of Nervous and Mental Disease* 2/142 (1966): 148-152.
[4] Thomas W. Simpson og Vincent C. Müller, „Just War and Robot's Killings," *Philosophical Quarterly* 33/263 (2016): 302-322.

Gagnrýni John R. Searles á gervigreind afhjúpar hvaða takmörkunum hún er háð. Hann leitast við að svara spurningum eins og þeirri hvort gervigreind skilji aðgerðir sínar og mögulegar afleiðingar af þeim. Ef við föllumst á að gervigreind geti ekki skilið aðgerðir sínar og borið ábyrgð heldur grundvallist aðgerðir og aðgerðarleysi hennar einungis á röklegri lausn tölvu á tilteknu vandamáli hlýtur að leiða af því að gervigreind eigi ekki að sinna verkefnum og úrlausnarefnum sem krefjast skilnings og ábyrgðar.

Í I. hluta ritgerðarinnar er fjallað um skilgreiningu á gervigreind ásamt því að farið er stuttlega yfir hugmyndir ýmissa fræðimanna um möguleika gervigreindar bæði í dag og náinni framtíð. Í II. hluta verður útskýrt í hverju gagnrýni Searles felst. Í III. hluta verður farið yfir staðhæfingar um gervigreind á sviði hernaðar, dómskerfis og heilbrigðisþjónustu með hliðsjón af þeirri gagnrýni sem fram kemur í II. hluta og þær vegnar og metnar í því skyni að komast að niðurstöðu um hvenær gervigreind ætti ekki að sinna verkefnum án atbeina mannlegrar dómgreindar.

Niðurstaða mín er að gervigreind sé og verði slíkum takmörkunum háð að hún geti ekki orðið meira en öflugt hjálpartæki mannsins. Hugmyndir um að hún sinni verkefnum sem krefjast skilnings og ábyrgðar án mannlegrar umsjónar tel ég ekki vera á rökum reistar.

I. hluti

Innan hinna ýmsu fræðasviða sem fást við gervigreind er ekki alltaf augljóst hvað átt er við með því hugtaki. Gervigreind (e. artificial intelligence) er í grunninn forrit sem byggist á flóknu safni skipana um verkefni sem tölva vinnur.[5] Með mikilli einföldun er skilgreiningin á gervigreind sú að vél líki eftir vitrænu hátterni sem menn tengja við mannshugann.[6] Þá er látið liggja milli hluta hvort eiginleg hugsun eigi sér stað eða við hvað er miðað þegar fjallað er um vitræna eða vitsmunalega hegðun. Framfarir á sviði gervigreindar síðustu ár sýna að „eftir því sem reiknigeta eykst og hugbúnaðurinn verður margbrotnari þá eykst geta vélanna til að hegða sér á þann hátt sem við myndum kalla „greindan"".[7] En jafnvel þó að fullkominni vélbrúðu tækist að leika eftir öll mannsins verk, og framkvæmdi þau jafnvel betur en sjálfur maðurinn, þýðir það ekki að slík brúða sé meðvituð eða hafi skilning á því sem hún gerir. Kristinn R. Þórisson segir að með tíð og tíma verði „að öllum líkindum" hægt að smíða „vélmenni sem haga sér algjörlega eins og manneskjur að öllu leyti – til dæmis úr nægilega fullkomnu líkani af heilanum byggðu á gervitauganetum – án þess að meðvituð upplifun þurfi nokkuð að koma þar við sögu".[8] Enn þá er langt að bíða þess að slík vélbrúða verði sköpuð. Í það minnsta þurfa framfarir að vera gífurlega miklar og örar næstu ár og áratugi eigi þessi sýn Kristins að verða að veruleika.

Algengt er í almennri umræðu að tala um gervigreind í samhengi við einföld og sértæk verkefni sem tölvu tekst að leysa, svo sem að tefla skák eða bera kennsl á andlit á myndum. Margir hafa glímt við skilgreininguna á gervigreind og hvað hún feli í sér en meðal þeirra fyrstu var Alan M. Turing. Hann spurði í grein sinni „Reikniverk og vitsmunir": „Geta vélar hugsað?"[9] Turing hélt því fram að skilgreiningarvandi fælist í sjálfri spurningunni því að erfitt væri að segja nákvæmlega hvað átt væri við með „hugsun" og „vél".[10] Því verður að nálgast verkefnið á annan hátt.

---

[5] Alan M. Turing, „Reikniverk og vitsmunir," *Hugur* 1/7, þýð. Atli Harðarson, (1995): 38. Turing notar að vísu ekki hugtakið gervigreind heldur hugsandi vélar um sama fyrirbærið.
[6] Stuart Russell og Peter Norvig, *Artificial Intelligence: A Modern Approach* (New Jersey: Prentice Hall, 2009), 2.
[7] Kristinn R. Þórisson, „Vélvitund, meðvitund og sjálfsvitund í kjötvélum og vélmennum," *Veit efnið af andanum? Sjö fyrirlestrar um meðvitundina*, 129.
[8] Sama rit, 134.
[9] Turing, „Reikniverk og vitsmunir," *Hugur* 1/7, 32.
[10] Sama rit, 32.

Turing ímyndaði sér að vél tæki þátt í svokölluðum hermileik. Hermileikur fer venjulega fram með þremur þátttakendum. Í dæmi Turings taka karlmaður, A, kvenmaður B og spyrill C þátt í leiknum. A og B eru í öðru herbergi en C sem sendir þeim spurningar og á að greina af svörunum hvort er hvað. Turing taldi best að leika leikinn með fjarrita. Hlutverk A í leiknum er að villa um fyrir C svo að spyrillinn giski rangt en B á að reyna að hjálpa C með svörum sínum. Hugmynd Turings var að láta vélina leika hlutverk A og að markmið leiksins væri að spyrillinn C gæti sér til um hvort væri maður og hvort vél. Kosturinn við leikinn að mati Turings er hann dregur skarpa línu á milli líkamlegra og vitsmunalegra hæfileika mannsins. Vélin sé ekki dæmd úr leik einungis vegna þess að hún hefur ekki líkama eins og maðurinn. Hið eina sem væri prófað væru vitsmunalegir hæfileikar en á því sviði taldi Turing að tölva gæti keppt.[11] Í grein sinni spáði Turing því að innan fimmtíu ára (þ.e. fyrir árið 2000) yrði hægt að forrita tölvur svo vel að „ekki [væru] meira en 70% líkur á að meðalspyrill [þekkti] viðmælendur rétt eftir að hafa rætt við þá í 5 mínútur".[12] Einnig taldi hann að málnotkun almennings myndi breytast þannig að honum yrði ekki framandi að líta svo á að vélar gætu hugsað.[13]

Eins og Turing benti á er ekki alveg skýrt hvað átt er við með „vél". Takmarkaði hann dæmi sitt um hermileikinn við að stafræn tölva gegndi hlutverki C.[14] Turing útskýrði stafrænar tölvur þannig að þeim sé ætlað að ráða við allar aðgerðir sem reiknandi maður getur framkvæmt og að hlutar hennar séu yfirleitt þrír: Geymsla, reikniverk og stýriverk. Sá pappír sem maðurinn notar við reikniverk sín sem og það sem hann leggur á minnið við sama verk samsvarar geymslu tölvunnar. Reikniverkið eru reikniaðgerðir tölvunnar og mannsins. Skipanatafla er svo þær reiknireglur sem eru í geymslunni eða á pappír mannsins og minni hans en stýriverkið sér um að þær séu rétt framkvæmdar og í réttri röð.[15] Vél sem hermir eftir manni reikna er forrituð til þess. Að forrita vél er að þýða yfir á form skipanatöflu hvernig maður fer t.d. að því að reikna. Það að forrita vél til að leika hlutverk A í hermileiknum er því að setja í hana skipanatöflu sem lætur hana leika A.[16]

[11] Sama rit, 32-33.
[12] Sama rit, 42.
[13] Sama rit, 43.
[14] Sama rit, 35.
[15] Sama rit, 36.
[16] Sama rit, 38.

Turing fjallaði stuttlega um Charles Babbage sem hugðist smíða stafræna tölvu en hann var prófessor við stærðfræði í háskólanum í Cambridge frá 1828 til 1839. Ástæðan fyrir því að Turing nefnir Babbage til sögunnar er til að benda á að stafræn tölva krefjist hvorki rafmagns né heldur taugakerfis. Babbage hugðist búa til vél sína úr hjólum og spjöldum. Turing hélt því fram að allar stafrænar tölvur væru jafngildar að vissu leyti og því hefði notkun rafmagns og taugakerfis fræðilega séð ekkert að segja.[17] Turing færði rök fyrir því með hvaða hætti stafrænar tölvur væru altækar eða jafngildar. Hann byrjaði á því að fjalla um stakrænar vélar, sem stafrænar tölvur tilheyra. Það eru vélar sem smella úr einni stöðu í aðra. Til skýringar er gott að hugsa sér rofa sem færist til um 120° gráður á einnar sekúndu fresti. Hægt er að stöðva rofann með því að grípa í stöng á vélinni sem hann er á. Í einni stöðu rofans kviknar ljós. Turing sagði stöður rofans þrjár sem hann nefndi q_1, q_2 og q_3. Staða stangarinnar á vélinni væri inntak hennar og annað hvort i_o eða i_1. Á hverju augnabliki ákvarðaðist staða rofans í vélinni af augnablikinu á undan og inntakinu. Eina sýnilega vísbendingin um stöðu rofans væri úttak vélarinnar en það er ljósið sem kviknar. Með því að vita upphafsstöðu vélar og öll inntaksmerki er mögulegt að spá fyrir um hvernig slíkar stakrænar vélar hegði sér um alla framtíð.[18] Þetta er þó í reynd flókið. Turing tók sem dæmi að vél sem var í notkun í Manchester á hans tíma hafi getað haft $2^{165.000}$ mismunandi stöður en rofinn í dæminu hér að framan hefði aðeins þrjár. Með töflu yfir hegðun stakrænu vélarinnar væri mögulegt að spá fyrir um hegðun hennar. Turing sagði að ekkert mælti gegn því að slíkir útreikningar væru framkvæmdir í stafrænni tölvu. Forsendurnar sem skipta máli eru þær að stafræna tölvan geri það nógu hratt og að fyrir hverja nýja stakræna vél þurfi að forrita hana upp á nýtt. Að því gefnu geti stafræn tölva hermt eftir hvaða stakrænu vél sem er.[19] Þessum eiginleika stafrænna tölva er lýst með því að segja að þær séu altækar. Altækar vélar hafi þann kost, sagði Turing, að geta unnið alla útreikninga ef þær eru nógu hraðar og sérstaklega forritaðar í hverju tilviki. Þannig séu þær jafngildar í þessum skilningi.[20]

Með hliðsjón af þessari hugmynd um stafrænar tölvur endurskoðaði Turing hermileikinn. Í stað þess að reyna að komast að því hvort stakræn vél gæti leikið leikinn vel spurði hann hvort hægt væri að velja eina ákveðna stafræna tölvu, laga hana til og láta hana

[17] Sama rit, 38.
[18] Sama rit, 39-40.
[19] Sama rit, 41.
[20] Sama rit, 42.

fá forrit við hæfi, þannig hún léki hlutverk A með fullnægjandi hætti gegn manni í hlutverki B. Turing gerði sér fulla grein fyrir að ekki samþykktu allir hina breyttu nálgun á því hvort vélar gætu hugsað en hann taldi að mætti sannreyna á grundvelli hennar hvort vélar hugsuðu. Upphaflegu spurninguna „geta vélar hugsað?" væri þýðingarlaust að ræða.[21] Þrátt fyrir það þá skoðaði hann nánar gagnrýni sem tefla mætti fram í tengslum við hermileikinn, þ. á m. rök prófessors Jefferson gegn því að vél gæti verið jafnoki heilans nema hún gæti ort sonnettu eða samið tónverk út frá hugsunum sínum og tilfinningum en ekki bara með tilviljanakenndri uppröðun á táknum. Sagði hann að ekki dygði til að vélin semdi þessi verk, hún yrði líka að vita að hún hefði samið þau og fundið til ánægju yfir vel unnu verki.[22] Þessi sjónarmið eru mjög áhugaverð í ljósi gagnrýni Searles á gervigreind sem farið verður nánar yfir í næsta hluta ritgerðarinnar. Gagnrýnin snýr að því að vélar geti ekki verið færar um skilning eða íbyggið hugarástand. Turing hélt því fram að eina leiðin til að komast að því hvort einhver hugsi eða finni til sé að láta hinn sama þreyta munnlegt próf. Hann segir enga aðra raunhæfa leið mögulega. Ætli maður að komast að því hvort einhver hugsi svo öruggt heiti þá sé aðeins ein leið fær en það er að vera sá maður og finna sig hugsa. Það sé aftur á móti sjálfsveruhyggja að dómi Turings. Sama megi segja um andmælin gegn hermileiknum. Að halda því fram að vél hugsi ekki þrátt fyrir að hún geti lýst „tilfinningum" og „hugsunum" ekkert síður en maður sé því sjálfsveruhyggja. Slík afstaða veldur vandkvæðum því að um leið og því er hafnað að vél hugsi sem þó stenst hið munnlega próf um að tjá „hugsanir" og „tilfinningar" eins og menn verður að hafna því að slíkt próf sanni að menn hugsi og finni til. Þess háttar viðhorf kann vel að vera rökrétt í strangasta skilningi en það er til siðs að fallast á þá viðteknu skoðun að aðrir menn hugsi og treysta þeim vísbendingum sem eru fyrir því og fást t.d. með samræðum.[23] Til þess að komast að því hvort vél beiti einungis brellum eða hugsi í alvörunni áréttaði Turing að hermileikurinn væri viðunandi próf. Væru svör vélarinnar jafngóð mannsins sem hún etur kappi við leyfði hann sér að efast um að því mætti með réttu lýsa sem einföldum brögðum.[24] Turing bætti því við að hann teldi ekki þörf á að blanda meðvitund inn í umræðuna um

[21] Sama rit, 42.
[22] Sama rit, 46.
[23] Sama rit, 47.
[24] Sama rit, 47-48.

hugsandi vélar. Honum virtist sem ekki væri nauðsynlegt að komast til botns í leyndardómnum sem meðvitundin er til að svara því hvort vélar geti hugsað.[25]

Kosturinn við prófið hjá Turing er hve framkvæmd þess er einföld. Gallar þess gera það hins vegar að verkum að prófið verður ekki áreiðanlegur mælikvarði á færni gervigreindar almennt. Einu forritin sem geta tekið Turing prófið eru þau sem eru fær um að tjá sig og eiga í samskiptum. Slíka færni er ekki að finna í fjölmörgum gervigreindarforritum sem sum hver sinna mjög flóknum verkefnum. Einnig er hægt að draga sterklega í efa að niðurstöður prófsins leiði það í ljós sem Turing ætlaði að það gerði. Er tvennt ólíkt að fallast á að tölva sýni vitsmunalega hegðun ef hún stenst prófið eða að segja fullum fetum að hún hugsi. En hvað veldur þessari tilhneigingu til að eigna forritum hugarástand? Ein ástæða gæti verið sterk þörf til að persónugera eitthvað sem sýnir eða virðist sýna vitsmunalega hegðun. Þetta kemur berlega í ljós hjá Joseph Weizenbaum í bók hans *Computer Power and Human Reason* en það sem vakti hann til umhugsunar voru einmitt sterk viðbrögð fólks við gervigreindarforritinu ELIZU, sem hann þróaði og var eitt fyrsta forrit sinnar tegundar, og hve fljótt þeir sem höfðu kynni af því litu á það sem viti borinn einstakling.[26] Í stað þess að líta á forritið sem áhugavert tækniafrek leit fólk framhjá því og leit svo á að Weizenbaum hefði skapað eitthvað annað og meira. ELIZA var sérhæft í tungumálagreiningu. Forritið var fært um að halda uppi samræðum við fólk eftir handriti. Frægasta handrit ELIZU, sem vakti mikla athygli, var DOCTOR. Handritið var einfalt í hönnun og líkti eftir viðtalstækni sálfræðinga sem kennd er við Carl Rogers. Viðtalstæknin felst m.a. í að endurorða svör sjúklingsins í framhaldsspurningum.[27] Þrennt undraði Weizenbaum og hafði djúpstæð áhrif á hann. Í fyrsta lagi var það fjöldi þeirra starfandi sálfræðinga og fræðimanna, s.s. Colby, Watt og Gilbert, sem töldu að ELIZA, með fyrirvara um að forritið yrði þróað áfram, gæti orðið að virku meðferðartæki og komið í stað sálfræðinga.[28] Í öðru lagi var það sterk tilhneiging til að persónugera forritið meðal þeirra sem höfðu átt í samskiptum við það.[29] Í þriðja lagi var það sú skoðun að fundist hefði almenn lausn á vandamálinu sem laut að svokölluðum tölvuskilningi á raunverulegu tungumáli.[30] Sumt

[25] Sama rit, 48.
[26] Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (San Francisco: W. H. Freeman and Company, 1976), 6-7.
[27] Sama rit, 3.
[28] Sama rit, 5; Colby, Watt og Gilbert, „A Computer Method of Psychotherapy: Preliminary Communication," *The Journal of Nervous and Mental Disease* 2/142, 148-152.
[29] Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*, 6.
[30] Sama rit, 7.

virðist ekki hafa breyst mikið síðan Weizenbaum skrifaði bókina sína um miðjan áttunda áratug síðustu aldar. Líkt og þá verður sömu þátta vart í almennri og fræðilegri umræðu um gervigreind. Heimspekingarnir Nick Boström og Daniel C. Dennet hafa t.d. gerst talsmenn hugmyndarinnar um formhyggju hugans og vélhyggju mannsins.[31] Sennilegt má telja að Weizenbaum, væri hann enn á lífi, þætti ýmislegt athugavert við slíkar hugmyndir og teldi að þær gæfu til kynna að gervigreind væri ekki háð þeim takmörkunum sem honum fannst eðlilegt að miða við. Weizenbaum áleit stóran hluta vandans felast í hugmyndum manna um greind. Hann sagði að almennt væri ríkjandi einfölduð mynd hennar. Öll þau veraldlegu vandamál sem maðurinn tækist á við krefðist margháttaðrar nálgunar. Mætti vel vera að vélar réðu við ýmislegt en það kynni að vera óviðeigandi eigi að síður að láta þær sjá um það. Þrátt fyrir að segja mætti um bæði menn og vélar að upplýsingavinnsla ætti sér stað væri grundvallarmunur á því ferli.[32] Vinnsla upplýsinga gæti ekki verið eini rökstuðningurinn fyrir því að vélar ættu að gera allt sem maðurinn gerir.

Weizenbaum hélt því fram að vera væri skilgreind að stórum hluta með hliðsjón af þeim vandamálum sem hún stæði andspænis. Mannvera stæði andspænis úrlausnarefnum sem vél gæti aldrei nokkurn tímann verið látin standa andspænis. „Maður er ekki vél," sagði hann einfaldlega.[33] Það segi ekkert um skilning að baki verknaðar þótt ómögulegt sé að greina á milli hvort maður eða tölva liggi að baki honum. Weizenbaum sagði að gera yrði meiri kröfur en þær einvörðungu að einhver eða eitthvað gæti framkvæmt eitthvað. Tölva kæmi ekki í stað manns þegar um væri að ræða virðingu fyrir öðrum, skilning og kærleika.[34] Í inngangi að bók sinni velti hann fyrir sér af hverju menn væru eins gjarnir og raun ber vitni á að skilja mannlega hugsun og skilning út frá vél- og formhyggju, eins og maðurinn væri að öllu leyti vélrænt kjötflykki sem mætti fullkomlega endurskapa í tölvu.[35] Þetta álítur hann vera grundvallarspurningu í nálgun sinni á því af hverju fólk hafi ríka tilhneigingu til að trúa að til geti orðið hugsandi vélar, vélar sem hefðu til að bera dómgreind og skilning á heiminum og umhverfi sínu.

[31] Nick Boström „How Long Before Superintelligence?" *Linguistic and Philosophical Investigations* 5/1 (2006), 11-30; Þorsteinn Gylfason. „Teikn og tákn." *Stúdentablaðið* 4/61 (1985): 17-19.
[32] Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation,* 203.
[33] Sama rit, 203.
[34] Sama rit, 269.
[35] Sama rit, 12.

Hluta vandans segir Weizenbaum felast í viðhorfi til greindar sem er nú á dögum mæld eftir einföldum megindlegum skölum.[36] Greindarvísitölupróf eiga að sýna fram á hversu greindur maður er og stendur miðað við aðra. Vandinn við þessi próf og þetta viðhorf gagnvart greind er ekki að prófin sjálf séu slæm heldur eru þau ófullkomin og þar af leiðandi ófullnægjandi. Þau segja ekki alla söguna. Samkvæmt Weizenbaum er helst tvennu ábótavant. Í fyrsta lagi er sköpun ekki eingöngu afsprengi greindar heldur samspil ýmissa þátta, svo sem innsæis og visku. Í öðru lagi hafa greindarvísitölupróf haft áhrif á hina almennu hugmynd um greind á þann hátt að við teljum hana algjörlega mælanlegt fyrirbæri sem er óháð allri viðmiðun. Án viðmiðunar eða samhengis er greind merkingarlaust fyrirbæri. Það þarf alltaf viðmið til að fá merkingu. Almennt er viðmiðið svo samofið okkur, í gegnum menningu, tungumál og félagsaðstæður, að við gerum okkur ekki grein fyrir því. En hvorki menning okkar, tungumál eða félagsaðstæður eru algild og eilíf sannindi og þess vegna verður alltaf að meta greind í samhengi við viðeigandi bakgrunn.[37] Weizenbaum fjallaði um ómenntaða móður sem gat ekki myndað málfræðilega réttar setningar á móðurmáli sínu en væri þrátt fyrir það fær um fágaðar og gáfulegar greiningar um eigin fjölskyldu.[38] Annað dæmi sem hann tók hefur orðið að ákveðinni staðalímynd hins fluggreinda manns og snýr að því hversu illa slíkum manni tekst til í einkalífinu. „Prófessorinn“ þykir ljóngáfaður en sérvitur og ófær um tilfinninga- eða félagsgreind. Hér má draga tölvu inn í myndina sem getur verið fær um að leggja hæfustu skáksnillinga mannkyns að velli þótt hún sé með öllu ófær um að skipta á barni.[39] Weizenbaum taldi þessa nálgun á greind hluta þeirrar villu sem ætti sér stað í hinni „ófrjóu umræðu“ um þróun gervigreindar sem stæði manninum framar. Weizenbaum taldi það vera merkingarleysu að halda því fram að tölvur gætu leyst af hendi viðkvæm verkefni manna eins og t.a.m. sálgreinis.[40]

Einnig er vert að nefna sjónarmið Weizenbaums um hversu færar tölvur geti orðið í að öðlast þekkingu. Hann segir að flestir tölvunarfræðingar hafi hafnað þeirri hugmynd að þekking sé ekkert annað en skipuleg niðurröðun mismunandi staðreynda.[41] Þetta skiptir máli því að halda mætti fram að gervigreind takmarkist af því sem hægt sé að kenna henni með

[36] Sama rit, 203.
[37] Sama rit, 204.
[38] Sama rit, 205.
[39] Sbr. sama rit, 205.
[40] Sama rit, 205-206.
[41] Sama rit, 207.

orðuðum eða skrifuðum skipunum. Sú er ekki raunin. Til eru forrit sem eru fær um að læra af umhverfi sínu, afla sér þekkingar. Dæmið sem Weizenbaum tók er af forriti sem lærði af hönnuði sínum að halda kústskafti í jafnvægi á öðrum enda. Í því tilfelli var augljóst að vélin fékk engin orðuð eða skrifuð fyrirmæli. Aftur á móti nefndi Weizenbaum að þrátt fyrir það væru ýmsar takmarkanir augljósar ef við aðeins veittum þeim athygli. Fyrst benti hann á að öll gervigreind sem lærði þyrfti ávallt ákveðna upphafsþekkingu. Hún gæti ekki byrjað á núllpunkti heldur þyrfti að mata hana á grunnupplýsingum og þar á ofan þyrfti hún lærdómsforrit.[42] Næst tók hann dæmi af handabandi tveggja manna. Þekkingin sem fælist í því að taka í höndina á öðrum sagði hann vera hreyfigreind og lágmarksforsenda hennar væri að hafa hendur. Mannleg þekking fælist í því að sumu leyti að hafa líkama. Við værum líkamsverur og að ákveðna þekkingu hefðum við aðeins vegna þeirra samskipta sem við höfum við aðrar líkamsverur. Að lokum benti hann á að tungumálið væri alls ekki jafn gegnsætt og margir halda. Þannig væru skilaboð oft háð því hver sendi þau og hver tæki við þeim. Skilaboð sem segi: „Kem með sjövélinni, elska þig, kveðja Bill," hafi mjög ólíka merkingu fyrir eiginkonu Bills sem átti von á honum heim með flugvél en vissi ekki klukkan hvað og fyrir konu sem beið ekki Bills og kom á óvart óvænt ástarjátning hans.[43] Því dregur Weizenbaum þá ályktun að menn fáist við alls kyns úrlausnarefni sem engin tölva gæti mögulega leyst úr. Jafnframt hélt hann því fram að lífverur skilgreindust að miklu leyti af þeim vandamálum sem þær þyrftu að glíma við.[44] Þrátt fyrir að þær ynnu úr upplýsingum líkt og vélar gerðu þær það ekki endilega á sama hátt.

Áður en gervigreind sleit barnsskónum leyfðu ýmsir fræðimenn eins og Weizenbaum sér að draga í efa að hún gæti nokkurn tímann staðið undir þeim háleitu hugmyndum sem margir höfðu um að hún gæti orðið jafnoki mannsins. Bjartsýnin var vissulega mikil. Herbert Alexander Simon hélt því t.d. fram um miðjan sjöunda áratug síðustu aldar að innan tuttugu ára mætti búast við því að gervigreind gæti sinnt öllum verkefnum mannsins.[45] Marvin Minsky taldi um svipað leyti að markmiðinu yrði náð að mestu leyti á lífaldri næstu kynslóðar.[46] Hins vegar varð ljóst með árunum að mun erfiðara yrði að ná slíkum markmiðum en bjartsýnustu

[42] Sama rit, 208.
[43] Sama rit, 208-209.
[44] Sama rit, 203.
[45] Herbert Alexander Simon, *The Shape of Automation for Men and Management* (New York: Harper & Row, 1965), 96.
[46] Marvin Minsky, *Computation: Finite and Infinite Machines* (New Jersey: Prentice-Hall, 1967), 2.

menn spáðu og ámóta framtíðarhugmyndir urðu fátíðari. Nú á síðustu árum hafa þær á hinn bóginn aftur gert vart við sig. Spár af svipuðu tagi má finna hjá heimspekingum og tölvunarfræðingum sem telja að vélar geti og muni geta sinnt öllum verkefnum sem maðurinn vinnur. Ray Kurzweil tölvunarfræðingur heldur því til dæmis fram að ekki aðeins verði gervigreind fær um alla mannlega hæfileika og getu á næstu áratugum heldur muni hún skara fram úr á öllum sviðum.[47]

Kurzweil telur að um leið og gervigreind nær sambærilegu stigi og maðurinn taki hún miklum framförum á örskotstíma. Þetta byggir hann helst á þeirri hugmynd að þekkingarflutningur sé svo auðveldur á milli véla að um leið og þær verði meðvitaðar geti þær sameinað þekkingu sína á ógnvænlegum hraða. Maðurinn hafi í grundvallaratriðum þróast á sama hátt og tekið framförum. Hann segir tungumálið hafa leikið lykilhlutverk og við getað miðlað upplýsingum á milli okkar sem hafi gert okkur kleift að afreka ýmislegt í hópi sem við hefðum aldrei getað ein og sér. Vélar séu einnig færar í að miðla þekkingu sín á milli ásamt því að hafa fullkomið minni og því geti þróun þeirra á þessum forsendum orðið margfalt hraðari en hún getur orðið hjá manninum.[48] Forsendurnar fyrir slíku er að gervigreind verði jafnoki mannsins á sínu sviði. Kurzweil telur að líklega verði hægt að búa til fullkomna skönnun af heila mannsins og virkni hans og þannig greina hann og endurskapa sem tölvuforrit. Hann áætlar að ekki verði lengra að bíða en til ársins 2029.[49] Hann gefur lítið fyrir að gervigreind hafi ekki staðið undir væntingum upphafsmanna hennar og telur að byrjunarörðugleikarnir hafi verið eðlilegar enda sé það venjan þegar ný og spennandi tækni er annars vegar. Hann bendir á að helstu styrkleika gervigreindar nú um stundir sé að finna í afmörkuðum verkefnum og að mörg forrit sem við flokkum ekki undir gervigreind eigi þar heima í raun og veru.[50] Vandinn sé ekki sá að gervigreind hafi ekki náð mikilli færni heldur viðhorfið til hennar. Lausnir hennar á ýmsum vandamálum séu smættaðar og taldar einfaldar. Kurzweil vitnar í Rodney Brooks, forstöðumann gervigreindarseturs Massachusetts Institute of Technology (MIT), sem sagði að um leið og vandamál væru leyst með gervigreind hyrfu ákveðnir töfrar yfir vandamálinu. Þar sem lausnin fælist að endingu bara í tölvuútreikningi

[47] Ray Kurzweil, *The Singularity Is Near: When Humans Transcend Biology* (New York: Penguin Group, 2005), 204-205.
[48] Sama rit, 204.
[49] Sama rit, 204-205.
[50] Sama rit, 206.

væri það ekki merkilegra en hver önnur tölvuvinnsla.[51] Kurzweil telur því gert óþarflega lítið úr því hve framarlega gervigreind raunverulega er.

Hugmyndir Kurzweils byggjast á þeirri forsendu að heilann megi endurskapa og endurvirkja í tölvu sem með hraða sínum og fullkomnu minni geti þróast á gífurlegum hraða og þannig tekið fram úr manninum á öllum sviðum. Þessi nálgun er ekki óalgeng hjá þeim sem telja gervigreind geta orðið manninum fremri í öllum verkum og allri hæfni. Svipaða hugmynd má finna hjá heimspekingnum Nick Boström þótt hann sé hógværari í spádómum sínum en Kurzweil. Boström telur líkt og Kurzweil að framtíð gervigreindar felist að mestu leyti í því að skilja heilann og endurskapa virkni hans í tölvuforriti.[52]

Fræðilegur grunnur þessara hugmynda Turings, Kurzweils og Boströms felst í kenningu um hugann sem Þorsteinn Gylfason þýðir sem *hákirkjukenningu hugsunarfræðinnar* en nafnið fær hann frá heimspekingnum Daniel C. Dennett. Í einföldu máli snýst hún um að mannshugurinn felist í heilaferlum sem séu ekkert annað en reiknanlegar aðgerðir sem megi endurskapa í tölvum. Forsenda þess að vélar geti orðið jafnstæðar manninum hvað varðar hugsun, umfram það að apa einungis eftir honum, felst í því að mannshugurinn „sé ekki annað en *hugbúnaður* eða *forrit* sem svo heitir á tölvumáli".[53] Í greininni „Teikn og tákn" sem er unnin eftir fyrirlestur Þorsteins um þessa kenningu gervigreindarmanna má sjá hana þrískipta: 1) Að mannleg hugsun sé úrvinnsla úr boðum. 2) Að úrvinnsla boða sé einber reikningur sem svo má sundurgreina í mismunandi reikningslistir. 3) Að teikn reikningslistanna hafa merkingu sem tengir hugsun fólks við heiminn. Merkingin sé svo regla sem tengir teikn við eitthvert merkingarmið. Þannig tengi finnsku orðin Pohjolan talo við Norræna húsið sem er mannvirki en ekki heiti.[54]

Þorsteinn færir rök gegn *hákirkjukenningu hugans* í grein sinni ásamt því að fjalla um ýmis andmæli gegn kenningunni. Helstu rök Þorsteins felast í því sem hann nefnir brigðhyggju sem hann setur fram til höfuðs löghyggju um mál og merkingu. Löghyggja felur í sér að það sem gerir teikn að skiljanlegum táknum sé merkingin sem þeim er gefin. Annars væru þau merkingarlaus. Merkingin er sú regla sem tengir teikn við eitthvert merkingarmið.[55] Þorsteinn

[51] Sama rit, 207.
[52] Boström „How Long Before Superintelligence?" *Linguistic and Philosophical Investigations* 5/1, 11-30.
[53] Þorsteinn Gylfason. „Teikn og tákn." *Stúdentablaðið* 4/61, 17.
[54] Sama rit, 17-18.
[55] Sama rit, 18.

bendir á að mörg orð tungumála gegni ýmsum mismunandi hlutverkum en það kallar hann *fjölkynngi* tungumálsins. Löghyggja geri aftur á móti ráð fyrir mismunandi reglum um hverja merkingu orðs. Nýstárleg notkun Þorsteins á orðinu *fjölkynngi* hér að framan verður honum svo efniviður í dæmi um að merking þess fari ekki fyrir ofan garð og neðan þrátt fyrir að flestir Íslendingar þekki orðið úr allt öðru samhengi. Þannig ályktar hann að þar sem auðveldlega megi brjóta reglur um merkingu orðs og þar sem reglur sem ávallt má brjóta séu engar reglur hljóti löghyggja að missa marks. Einu mótrök löghyggjunnar sem ekki lenda í vítahring segir Þorsteinn vera að hún greini „skýrt og skipulega milli eiginlegrar og óeiginlegrar eða afbrigðilegrar merkingar" en það segir hann ekki vera hægt.[56] Í því sambandi vísar hann í Searle og athugunar hans á ensku sögninni „cut": „Bill cut the cake", „Sally cut the grass" og „John cut my hair". Hér á við sama eiginlega merking sagnarinnar í öll þrjú skiptin. Á íslensku þyrfti þrjár sagnir til að lýsa athöfnunum þremur en ekki eina: Skera, slá og klippa. Ólíkt fyrra dæminu hvarflar ekki að neinum að þessar þrjár setningar hafi sömu eiginlegu merkingu, segir Þorsteinn, og Searle ætti ekki að þurfa neinnar íslensku með, „hvers vegna skyldum við ekki segja skilning eða merkingu sagnarinnar *cut* breytast eftir því hvert verkfærið er sem notað er til viðeigandi athafnar hverju sinni?"[57] Þorsteinn brýtur vandamálið ekki til mergjar í grein sinni og segir að margt sé ósagt um brigðhyggju. Hvað sem því líði sé þó „hákirkjukenningin [...] fallin, og það fyrir næstum barnslegum rökum".[58]

[56] Sama rit, 19.
[57] Sama rit, 19.
[58] Sama rit, 19.

II. hluti

Löngu áður en gervigreind varð að sjálfstæðri fræðigrein var hún umfjöllunarefni heimspekinga. Bæði René Descartes og Gottfried Wilhelm Leibniz fjölluðu um hana. Descartes sagði tvær leiðir ávallt færar til að greina á milli manns og vélar. Í fyrsta lagi að vél gæti aldrei líkt eftir, með fullnægjandi hætti, hæfileikanum til að tengja saman orð og tákn til að tjá hugsanir en það gætu alsljóustu menn. Í öðru lagi að þrátt fyrir að vél gæti gert ýmislegt vel og ef til vill betur en maður myndi henni fatast í ýmsu öðru. Þá yrði ljóst að athafnir hennar væru háðar líffæraskipan hennar. Descartes taldi óhugsandi að vél yrði búin til sem hefði líffæri sem leyfðu henni að aðhafast með þeim hætti sem skynsemi leyfi mönnum að hafast að.[59] Enda var hann uppi fyrir tíma stafrænu tölvunnar.

Leibniz dró líka í efa möguleika gervigreindar á því að verða hugsandi vera. Hann taldi ekki hægt að lýsa skynjuninni og því sem ylti á henni með vélhyggjurökum. Dæmið sem hann tók er áhugavert en því svipar til dæmisins sem Searle fjallaði um. Leibniz sagði að ef um væri að ræða vél „sem væri þannig saman sett að hún framleiddi hugsun, tilfinningu og skynjun, þá mætti hugsa sér hana stækkaða með því að halda sömu hlutföllum, þannig að unnt væri að komast inn í hana eins og í myllu. Og að því gefnu, þá fyndi maður ekki annað, með því að skoða hana að innan, en hluta sem ýttu hver á annan, en aldrei neitt sem gæti gert grein fyrir skynjuninni sjálfri".[60]

Searle stígur sömu glímu en á mun ítarlegri hátt í grein sinni „Hugur, heili og forrit". Hann gerir greinarmun á tvenns konar gervigreind, „róttækri" (e. strong) og „hófsamri" (e. weak). Megingildi hófsamrar gervigreindar við rannsóknir á mannshuganum segir Searle felast í því að hún sé gífurlega öflugt verkfæri sem geri okkur kleift að sannreyna ýmsar tilgátur á mjög nákvæman hátt. Róttæk gervigreind felur í sér að gervigreind forrituð á réttan hátt sé ekki einungis tól til rannsókna á mannshuganum heldur sjálf hugur sem býr yfir raunverulegum skilningi og vitsmunum.[61]

Searle beinir spjótum sínum að róttækri gervigreind. Hann tekur sérstaklega fyrir gervigreindarforrit Rogers Schank og samstarfsmanna hans þar sem það er skýrt dæmi um

[59] René Descartes, *Orðræða um aðferð*, þýð. Magnús G. Jónsson (Reykjavík: Hið íslenzka bókmenntafélag, 1998), 126-127.
[60] Gottfried Wilhelm Leibniz, *Orðræða um frumspeki*, ritstj. Ólafur Páll Jónsson, þýð. Gunnar Ágúst Harðarson (Reykjavík: Hið íslenzka bókmenntafélag, 2004), 150-151.
[61] John R. Searle, „Hugur, heili og forrit," *Hugur* 1/7, þýð. Ólafur Páll Jónsson (1995): 64.

gervigreindarforrit af því tagi sem hann vill skoða, auk þess sem hann þekkir það vel.[62] Hann tekur fram að sömu rök megi færa um forritið ELIZU eftir Weizenbaum og í raun allar Turingvélar sem líkja eftir mannlegu hugarstarfi. Searle lýsir því í stuttu máli hvað felist í forriti Schanks. Tilgangur þess er að líkja eftir hæfileika manna til að skilja sögur. Forritið er fært um að svara spurningum um sögur sem koma ekki berlega fram. Þannig er hægt að mata það af sögum og síðan spyrja það spurninga og fá svör við þeim sem líkjast því að maður hefði lesið söguna og svarað. Searle segir fylgismenn róttækrar gervigreindar halda tvennu fram um þessa hæfni forritsins sem hann telur hvoruga geta staðist. Í fyrsta lagi að segja megi að forritið skilji í raun söguna og í öðru lagi að það sem vélin og forritið geri endurspegli hinn mannlega hæfileika sem felst í því að skilja sögu og svara spurningum um hana.[63]

Ein leið til þess að prófa kenningar um hugann segir Searle felast í því að spyrja sjálfan sig hvað það þýddi ef manns eigin hugur ynni í raun eftir lögmálum kenningarinnar.[64] Searle prófar kenninguna hér að ofan með hugartilrauninni um „kínverska herbergið“. Hann ímyndar sér að hann sé lokaður inni í herbergi þar sem honum berist ógrynni kínverskra tákna. Fyrir honum er kínverskt ritmál merkingarlaust með öllu. Þannig getur hann ekki einu sinni verið viss um að hann sé með kínversk tákn fyrir framan sig frekar en japönsk eða merkingarlaust pár. Í kjölfar þessa fyrsta skammts af táknum fær hann annan skammt af sama letri nema ásamt honum eru reglur á ensku sem greina hvernig beri að tengja seinni skammtinn við þann fyrri. Reglurnar skilur hann eins vel og hver annar enskumælandi maður. Þær gera honum kleift að tengja saman fyrri og seinni skammtinn af formlegu táknunum. Með formlegum táknum á hann við að lögunin er það eina sem gerir honum kleift að þekkja þau. Þriðja skammtinn af kínverskum táknum fær Searle svo ásamt leiðbeiningum á ensku sem gera honum kleift að tengja tákn úr þriðja skammtinum við hina fyrri tvo. Í leiðbeiningunum kemur fram hvernig hann eigi að skila til baka ákveðnum kínverskum táknum sem viðbragði við tiltekinni lögun í þriðja skammtinum. Án þess að hann viti það kallar fólkið sem lætur hann hafa táknin fyrsta skammt „handrit”, annan „sögu” og þriðja „spurningar”. Táknin sem hann skilar til baka sem viðbragð við tiltekinni lögun í þriðja skammtinum kallar fólkið „svör við spurningum” og reglurnar „forrit”. Searle víkkar út hugartilraunina með því að láta afhenda sér auk kínversku táknanna sögur á ensku sem hann skilur vel og spurningar um þessar sögur sem hann svarar á

[62] Sama rit, 64.
[63] Sama rit, 65.
[64] Sama rit, 66.

ensku. Með tímanum verður hann svo fær að fylgja reglunum um meðferð kínversku táknanna og forritararnir svo góðir í að skrifa forritin að utan herbergisins væri engin leið að greina á milli svara hans og innfædds Kínverja. Svör hans við ensku spurningunum yrðu einnig óaðgreinanleg frá enskumælandi mönnum. Utan frá séð væru svör hans annars vegar á ensku og hins vegar á kínversku jafngóð.[65] Samt eru kínversku svörin fundin með því með því að möndla með ótúlkuð formleg tákn. Kínversku svörin eru því fengin með því að framkvæma reiknanlegar aðgerðir á formlega skilgreindum táknum. Searle er þannig að haga sér eins og tölva og er orðinn staðgengill hennar.

Í ljósi framangreindrar hugartilraunar er hægt að skoða áðurnefndar fullyrðingar fylgismanna róttækrar gervigreindar. Í fyrsta lagi var því haldið fram að forrituð tölva skildi sögur og í öðru lagi að forritið endurspeglaði mannlegan skilning á sögunni. Í hugartilrauninni hjá Searle er ljóst að hann skilur ekki stakt orð í kínversku sögunni. Þrátt fyrir að skila af sér svörum sem eru óaðgreinanleg frá innfæddum Kínverja skilur hann ekkert. Það greinir hann frá Kínverja sem hefði skilað sömu svörum. Þannig býr tölva, sem leysir verkin af hendi með sama hætti og Searle, ekki yfir neinum skilningi. Í tilviki kínverskunnar er Searle „tölva“ og tölva, sem væri ekki Searle, hefði ekkert fram yfir hann. Af þeirri ástæðu skilur forrit Schanks því ekkert í sögunni, það möndlar einfaldlega með formleg tákn. Seinni fullyrðingin um að forritið endurspegli mannlegan skilning gengur heldur ekki upp ef forritið skilur ekki söguna. Searle segir að tölvan og forritið séu ekki nægjanleg skilyrði fyrir skilningi þar sem þau virki án þess að skilja. Í kjölfarið veltir Searle fyrir sér hvort þau séu nauðsynleg skilyrði yfirhöfuð eða hvort þau skipti máli fyrir skilning. Samkvæmt Searle halda róttækir gervigreindarsinnar því fram að það sem felist í skilningi hans á sögu á ensku sé ekkert annað en möndl með tákn þar sem viðameiri táknavinnsla fer þá fram en þegar hann möndlaði með kínversk tákn. Magn eða umfang táknavinnslunnar greini þannig á milli tilviksins „þar sem ég skil“ frá tilvikinu „þar sem ég skil ekki“. Searle viðurkennir að hugsunartilraun hans hrekji ekki þessa fullyrðingu en hún sé hins vegar ósennileg miðað við hvað tilraunin felur í sér. Eigi fullyrðingin að vera sönn þarf að vera hægt að skrifa forrit sem hefur sama inntak og úttak og venjulegt fólk. Þar að auki verði að vera mögulegt lýsingarstig þar sem segja má að fólk vinni eins og forrit.[66] Af þessu má leiða að forrit Schanks sé ekki fært um skilning en geti verið hluti

[65] Sama rit, 66-67.
[66] Sama rit, 67.

af því hvað skilningur er. Þetta viðurkennir Searle en segir jafnframt að við höfum enga ástæðu til að ætla að slíkt sé satt. Af hugartilrauninni má leiða, þótt það sannist ekki með henni, að forritið skipti engu máli fyrir skilning á sögu. Allt sem gervigreind hefur væri hægt að láta Searle hafa í dæminu og hann, líkt og forritið, gæti unnið með kínversku án þess að skilja nokkuð í henni. Í tilvikinu með enskuna skilur Searle allt. Ekkert bendir til þess að skilningur hans á ensku hafi átt neitt skylt með tölvuforriti. Það vinnur eftir reiknanlegum aðgerðum á formlega skilgreindum táknum og dæmið sýnir að slíkar aðgerðir skipta engu verulegu máli fyrir skilning. Erfitt er að trúa því að skilningur mannsins felist í slíku. Searle segir kjarna málsins vera þann að það skipti ekki máli eftir hvaða algerlega formlegu reglum tölvan vinnur, þau verði aldrei nægjanleg skilyrði skilnings því að lítið mál er að fylgja þeim eftir án þess að skilja neitt. Engin rök segir hann hafa verið færð sem styðji það að slíkar formlegar reglur séu nauðsynlegar skilningi eða leggi honum eitthvað til þar sem engin rök geta sýnt fram á að við vinnum yfirhöfuð eftir formlegum forritum þegar við skiljum, eins og þegar hann skildi ensku í tilrauninni.[67]

Hvað skilur á milli? Hvað felst í skilningi Searles á ensku sem skortir í möndlinu með kínversku táknin? Searle segir svarið vera að hann viti hvað ensku setningarnar merkja meðan hann hafi ekki hugmynd um hvað þær kínversku eigi að þýða.[68] Spurningin verður þá hvað felist í því og hvers vegna við getum ekki komið því fyrir í vél? Áður en Searle leggur fram tillögu að svari reynir hann að koma í veg fyrir algengan misskilning á því hvað „skilningur" sé ásamt því að taka fyrir ýmsa gagnrýni á hugartilraun sína.[69] Gagnrýnendur Searles hafa bent honum á að skilningur getur verið mismikill. Searle samþykkir það en segir jafnframt að það hafi ekkert með kjarna málsins að gera. Augljós dæmi sé hægt að finna þar sem „skilningur" á við og þar sem hann á ekki við, sem er það eina sem röksemdafærsla hans krefst. Þannig segir Searle að hann skilji sögur á ensku, síður sögur á frönsku, enn síður sögur á þýsku og sögur á kínversku alls ekki. Því næst tekur hann dæmi um ýmis tæki sem skilja ekki neitt, svo sem bíla og samlagningarvélar. Vissulega segjum við að hlutir svo sem bílar og samlagningarvélar hafi skilning og aðra vitsmunalega eiginleika en það er einföld persónugerving sem ekkert sannar. Það er tvennt ólíkt að segja að samlagningarvél skilji samlagningu en ekki deilingu og að halda því fram að hún hafi skilning á samlagningu en ekki

[67] Sama rit, 68.
[68] Sama rit, 68.
[69] Sama rit, 68-69.

deilingu. Searle segir það áhugavert hvernig við eignum hlutum eigin íbyggni (e. intentionality). Samlagningarvél sem skilur samlagningu á ekkert skylt við skilning manna á samlagningu eða það hvernig maður getur skilið ensku. Ef forrit Schanks skilur sögur eins og samlagningarvél skilur samlagningu er tilgangslaust að ræða málið. En það er ekki viðhorf róttækra gervigreindarsinna. Þeir staðhæfa að tölvur geti búið yfir samskonar vitsmunum og mennirnir. Searle segir að sér líki hve afdráttarlaus sú fullyrðing er. Hann telur hægt að færa rök fyrir því að forrituð tölva skilji jafn mikið og samlagningarvélar, sem sagt ekki neitt.[70] Ekki sé um að ræða stig lítils skilnings eins og til dæmis skilnings Searles á þýsku, heldur sé skilningurinn alls enginn.

Næst tekur Searle fyrir sex helstu rök gagnrýnenda sinna og gerir skipulega grein fyrir þeim ásamt því að færa rök gegn þeim.

Fyrstu rökin sem hann fjallar um nefnir hann kerfisrökin. Í stuttu máli snúast þau um að þótt Searle í herberginu skilji ekki sjálfur söguna á kínversku sé hann einfaldlega hluti af stærra kerfi sem skilji söguna. Hlutar kerfisins eru þá allt það sem hann fær til sín í herbergið: Reglurnar, kínversku táknin og þau blöð og skriffæri sem hann svarar með. Viðbrögð Searle við þessari gagnrýni eru að breyta hugartilrauninni um kínverska herbergið í mann. Kerfið sem skilur söguna og Searle sjálfur eiga samkvæmt því að verða hluti af einum manni. Sá maður man reglurnar og kínversku táknin og gerir alla útreikninga í huganum. Kínverska herbergið, kerfið sem skilur samkvæmt rökfærslunni, er þannig orðið að einum manni. Þrátt fyrir þetta er allt við það sama segir Searle. Searle skilur enga kínversku inni í manninum og sama máli gegnir um kerfið því að það hefur ekkert umfram manninn. Searle fjallar sérstaklega um aðra útgáfu af sömu rökum þar sem nálgunin er eilítið önnur. Þar er samþykkt að maðurinn sem hefur innbyrt allt kerfið hafi ekki til að bera skilning á kínversku á sama hátt og innfæddur en maðurinn sem formlegt táknvinnslukerfi vinni þannig úr sögunni að halda megi því fram að hann raunverulega skilji kínversku.[71] Í dæminu eru þá tvenn aðskilin undirkerfi í manninum. Annað skilur ensku og hitt kínversku. Searle svarar því þannig að undirkerfið sem skilur ensku viti hvað sagan er um og svarar spurningunum eins vel og það getur með hliðsjón af innihaldi sögunnar á meðan kínverska kerfið veit ekkert.[72] Enska undirkerfið þekkir merkingu þess sem vísað er til í sögunni. Kínverska undirkerfið veit aðeins að á eftir einu

[70] Sama rit, 69-70.
[71] Sama rit, 70-71.
[72] Sama rit, 71.

tákni kemur annað tákn. Slíku undirkerfi tekst því ekki að skilja kínversku eins og Searle skilur ensku. Í þessu samhengi fjallar Searle um Turingprófið. Ljóst er af þessu dæmi að hér er um tvö kerfi að ræða sem bæði standast Turingprófið en aðeins annað býr yfir skilningi. Það að kerfi standist Turingprófið er því ekki rök fyrir því að það skilji. Kerfið sem skilur enskuna hefur margfalt meira til að bera en það sem möndlar með kínverskuna segir Searle. Forsendur kínverska undirkerfisins eru ekkert annað en inntak og úttak með forriti þar á milli. Ef það eru nægjanleg skilyrði skilnings virðast alls konar óvitsmunaleg undirkerfi hafa vitsmuni. Þannig vinni maginn t.d. úr gögnum og fylgi ótal forritum en fáir fengjust til þess að segja hann skildi nokkurn skapaðan hlut. Nú væri hægt að benda á að upplýsingarnar sem koma inn í herbergið og maturinn sem berst maganum hafi áhrif á það hvort skilningur geti átt sér stað. Searle bendir á að enginn munur sé á því sem berst kerfunum tveimur. Frá sjónarhóli þess sem er í herberginu eru ekki neinar upplýsingar í kínversku táknunum fremur en í matnum.[73] Áður en Searle segir skilið við kerfisrökin fyrir róttækri gervigreind veltir hann fyrir sér tengslum hennar við sálfræðina. Róttæk gervigreind reyni að vera grein sálfræðinnar. Ef henni á að takast það verði hún að geta greint þau lögmál sem hugurinn vinnur eftir frá lögmálum óhugrænna kerfa. Annars sé ekki að vænta neinna svara hjá róttækri gervigreind um hið hugræna. Staðhæfingar gervigreindarsinna grafa oft undan eigin markmiðum. Þannig vitnar Searle í McCarthy sem segir: „Það má segja um jafn einfalt tæki og hitastilli að hann hafi skoðun, og að hafa skoðun virðist vera einkenni flestra tækja sem geta leyst vandamál.”[74] Án þess að færa nein sérstök rök gegn þessari fullyrðingu höfðar Searle til innsæisins. Hann segir rannsókn á huganum hefjast á staðreyndum eins og þeirri að manneskjur hafi skoðanir en ekki hitastillar eða samlagningarvélar. Sé því hafnað er nauðsynlegt að kanna það nánar.[75] Searle spyr því hvað þurfi til svo að hitastillir geti talist hafa raunverulegar skoðanir. Gæti hitastillir haft sterka skoðun eða látið sér standa á sama? Gæti hann verið taugaveiklaður eða öruggur? Searle segir hitastillinn rétt eins og magann eða samlagningarvélina ekki eiga möguleika á því.[76]

Önnur rökin af sex sem Searle skoðar nefnir hann vélmennisrökin. Þau eru á þá leið að í staðinn fyrir forrit Schanks yrði skrifað annað forrit sem væri komið fyrir inni í vélmenni.

[73] Sama rit, 72.
[74] Sama rit, 73.
[75] Sama rit, 73.
[76] Sama rit, 73-74.

Vélmennið væri ekki einungis fært um að taka við og skila út formlegum táknum heldur hegðaði það sér eins og það skynjaði, sæi og hreyfði sig ásamt ýmsu öðru, svo sem að borða og drekka. Vélmenninu væri þannig kleift að athafna sig og því væri alfarið stjórnað af tölvuheila. Þannig byggi það yfir raunverulegum skilningi. Fyrst vekur Searle athygli á því að vélmennisrökin telja að vitsmunir felist ekki aðeins í formlegri táknvinnslu eins og forrit Schanks gerir heldur þurfi að vera til að dreifa orsakatengslum við umheiminn.[77] Aftur breytir Searle upphaflega kínverska herberginu en í þetta sinn verður umbreytingin þannig að Searle í herberginu stýrir vélmenninu sem hann veit þó ekki af. Með sama hætti og áður er lýst berast honum kínversk tákn en nú gegnum myndavélar vélmennisins sem eru sjón þess. Hann heldur áfram að möndla með kínversku táknin, tekur við þeim og svarar eftir reglunum með því að skila til baka viðeigandi táknum. Síðan ímyndar Searle sér að sum þessara tákna hafi áhrif á aflgjafa vélmennisins þannig það hreyfi sig, bregðist við. Aftur er engu aukið við sjálfan skilninginn hjá forritinu. Upplýsingar berast frá skynjurum vélmennisins og með formlegri táknavinnslu sendir Searle áfram skipanir til aflgjafa vélmennisins án þess að verða nokkurs vísari um eitt eða annað.[78] Enn er engan skilning að finna.

Þriðju gagnrýnisrökin, hin svokölluðu heilalíkansrök, ganga út á að hannað yrði forrit sem ynni ekki með upplýsingar um heiminn, eins og upplýsingarnar um sögurnar hjá Schank, heldur væri líkt eftir taugaboðum í heila Kínverja sem skilur sögur á kínversku og svarar spurningum um þær. Forritið fengi síðan sögur á kínversku og spurningar um þær og myndi líkja eftir hinum formlegu taugaboðum í heila Kínverja og skila þannig frá sér svörum á kínversku. Kæmi jafnvel til greina að ekki eitt heldur mörg forrit, sem ynnu á svipaðan hátt, líktu eftir virkni þessara taugaboða, eins og ætla má að mannsheili geri þegar hann vinnur úr táknum tungumáls. Þá væri ekki hægt að neita því að vél sem ynni eftir þessum hópi forrita skildi sögur á kínversku.[79] Þessa afstöðu um að endurskapa taugaboð heilans í tölvu er bæði að finna hjá Kurzweil og Boström. Searle hefur leika á því að segja að slík nálgun sé undarleg í gervigreindarfræðum. Hans skilningur á grundvallarhugmynd gervigreindarfræða væri sú að við þyrftum ekki að vita hvernig heilinn starfar til að vita hvernig hugurinn starfar. Tilgátan væri sú að til væri „millistig þar sem heilaferli væru reiknanlegar aðgerðir á formlegar einingar sem mynduðu kjarnann í hinu hugræna, og að þetta gæti átt sér stað í allskonar heilastarfsemi

[77] Sama rit, 74.
[78] Sama rit, 74-75.
[79] Sama rit, 75.

á sama hátt og hvaða tölvuforrit sem er má keyra með ólíkum vélbúnaði".[80] Samkvæmt róttækri gervigreind sé einmitt ekki meira samband á milli hugar og heila en á milli forrits og tölvu. Því ætti að vera hægt að skilja hugann án taugalífeðlisfræði. En þrátt fyrir að hann telji þessa nálgun undarlega svarar hann rökunum efnislega. Hann telur að þrátt fyrir að heilastarfsemin væri leikin eftir með þessum hætti væri það ekki fullnægjandi til að hægt væri að fullyrða að um skilning væri að ræða. Hann breytir nú kínverska herberginu í flókna vatnslögn sem hann vinnur með eftir reglum. Á öllum samskeytum eru kranar sem hann getur skrúfað fyrir og frá. Sérhver samskeyti svara til taugamóta og ef opnað er fyrir allar réttu lagnirnar þá bunar kínverska svarið út úr úttakinu á lögninni. Enn segist Searle eiga bágt með að sjá hvar skilningurinn ætti að vera. Vatnslögnin tekur við kínversku og líkir eftir taugamótum sem eftir réttum formlegum leiðum skilar kínversku en enn og aftur skilur hann ekki kínversku og vatnslögnin ekki heldur. Searle segir gallann við heilalíkanið felast í því að það líki eftir röngum þáttum heilans. Formleg uppbygging taugamóta nægi ekki ein og sér án orsakabundinna eiginleika heilans.[81]

Séu ofangreind andmæli skoðuð hvert í sínu lagi kunna þau að virka sem ósannfærandi gagnrýni á kínverska herbergið en samantekið verða þau öflugri. Þessa afstöðu nefnir Searle samsetningarrökin. Samkvæmt þeim er skapað vélmenni með heilalíkanstölvu sem komið er fyrir í eftirlíkingu af hauskúpu og hún svo forrituð með öllum taugamótum mannsheilans. Öll hegðun tölvunnar verður þannig óaðgreinanleg frá mannlegri hegðun og allt er þetta eitt órofa kerfi sem vinnur ekki út frá einföldu inntaki og úttaki. Searle samþykkir að hér sé komið tilvik þar sem erfitt væri að gera greinarmun á vélmenni og manni og því væri skynsamlegt, að því gefnu að mann skorti mótrök, að ætla að vélin hefði íbyggni. Raunar þarf ekki að vita neitt um vélmennið til að ætla það. Searle segir hins vegar að lítil hjálp sé í þessum samsettu rökum fyrir róttæka gervigreindarsinna en þeir gera ráð fyrir því að nægjanlegt skilyrði fyrir íbyggni felist í formlegu forriti með réttu inntaki og úttaki. En það að vélmenninu sé eignuð íbyggni hafi ekkert með formleg forrit að gera. Ef hegðun þess og atgervi er nógu líkt manni ætlum við því hugarástand eins og manna. Aftur á móti segir Searle að ef við hefðum vitneskju um hvernig gera ætti grein fyrir hegðun vélmennisins án þess að ætla því hugarástand myndum við ekki eigna því íbyggni, sér í lagi ekki ef við vissum til þess að það væri formlegt forrit.[82]

[80] Sama rit, 76.
[81] Sama rit, 76.
[82] Sama rit, 77.

Searle leggur til enn eina hugartilraun. Í tilfelli vélmennisins sem hegðar sér eins og maður ímyndum við okkur að inni í því starfi maður sem fær hrá formleg tákn frá skyntækjum vélmennisins og lætur frá sér hrá formleg tákn til gangverks þess, allt eftir gríðarlegu reglusafni. Auk þess gerir hann sér ekki grein fyrir því að hann stjórnar vélmenni. Þannig vinnur hann einungis með tákn án þess að skilja merkingu þess en hann veit hvaða aðgerðir eru viðeigandi í sérhverju tilviki. Vélmennið verður þannig að hugvitssamlegri vélbrúðu. Nú væri ástæðulaust að eigna vélbrúðunni íbyggni. Formlega táknvinnslan heldur áfram en eini staðurinn þar sem íbyggni er að finna er hjá manninum sem stýrir brúðunni. Hann þekkir hins vegar ekkert til þess ástands sem brúðan er í. Ætlun hans er ekki að stjórna brúðunni heldur er hún einungis fólgin í því að möndla með tákn. Til þess að skýra þetta nánar fjallar Searle um frumstæð dýr eins og apa. Við getum ekki skilið hegðun þeirra, að hans dómi, án þess að eigna þeim íbyggni enda sjáum við að þeir eru að mörgu leyti líkt sköpunarverk og við sjálf. Þeir hafa t.d. augu og eyru, nef og munn. Við ályktum að samkvæmni í hegðun þeirra orsakist af því að þeir búi yfir hugarástandi sem liggur hegðun þeirra til grundvallar. Svipaða niðurstöðu fengjum við með því að skoða vélmennið í samsetningarrökunum. Verður þó ljóst að um leið og við vitum að formlegt forrit ræður hegðun vélmennisins og að eiginleikar efnislegra hluta þess, útlima og skynfæra, skipta engu máli, þá myndum við hafna niðurstöðunni um að vélmennið hefði íbyggni.[83]

Searle fjallar stuttlega um tvenn önnur gagnrýnisrök. Fyrri rökin nefnast aðrir hugar sem fjallað var stuttlega um í I. hluta og Turing reyndi að sneiða hjá með prófinu sínu. Rökin grundvallast á því að eina leiðin til að vita yfirhöfuð að annað fólk hafi vit sé að ráða það út frá hegðun þess. Takist tölvu að standast slíkt hegðunarpróf hljóti að verða að fallast á að hún hafi vitsmuni. Searle svarar þessu snaggaralega og segir að vandamálið snúist ekki um það hvernig við vitum að annað fólk hafi vitsmuni. Vandinn felist í því hvað það sé sem maður eignar fólki þegar því eru eignaðir vitsmunir.[84] Kínverska herbergið sýnir fram á að eingöngu útreikningar og útkoman úr þeim, þ.e. einföld, formleg táknvinnsla, séu ekki nægjanleg skilyrði vitsmuna vegna þess að það geti vel verið til án þeirra.

Seinni rökin, sem Searle nefnir völundarhúsarökin, snúast um vélarnar sjálfar. Rökin ganga út á það að gagnrýni Searles eigi einungis við um hliðstæðutölvur og stafrænar tölvur.

[83] Sama rit, 78.
[84] Sama rit, 79.

Ef það sem Searle heldur fram er satt og orsakaferli, hver sem þau eru, séu nauðsynleg forsenda íbyggni, þá mun verða hægt í óljósri framtíð að búa til tölvur eða tæki sem búa yfir þeim - og það mun vera gervigreind. Því beinast rök Searles ekki að þeim möguleika gervigreindar að geta búið til og skýrt vitsmuni. Searle hefur lítið við þetta að athuga fyrir utan að hann telur þessa nálgun afbaka markmið gervigreindar. Grundvöllur kenningarinnar um gervigreind felist einmitt í þeirri hugmynd að hugarferli eru reiknanlegar aðgerðir á formlega skilgreindar einingar. Þetta gagnrýnir Searle. Ef grundvöllurinn er einhver annar, þá er kenningin ekki lengur hin sama. Searle segir ekki hægt að gagnrýna það sem hefur enga prófanlega tilgátu.[85]

Eins og kom fram áður setti Searle fram spurningu um hvað fælist í því að skilja ensku annars vegar og að skilja ekki kínversku hins vegar. Hvað það væri sem hann hefði í fyrra tilvikinu en skorti í seinna. Einnig spurði hann af hverju ætti ekki að vera hægt að gefa vél „þetta", hvað sem „þetta" er. Í fljótu bragði sér Searle ekkert sem fræðilega útilokar að vél geti skilið ensku eða kínversku þar sem líkami manns og heili eru í ákveðnum skilningi einmitt slík vél. Á hinn bóginn dregur hann sterklega í efa hvað snertir hugartilraunina um kínverska herbergið að hún geti skilið ef hún vinnur eftir reiknanlegum aðgerðum á formlega skilgreindum einingum. Searle telur það ekki felast í því að hann sjálfur sé „forrit" að hann skilji ensku. Hann telur t.d. að hann sé þá ótölulegur fjöldi forrita en það sé ekki forsenda skilnings heldur það að hann sé lífvera með ákveðna líffræðilega gerð sem getur alið af sér íbyggni eins og skynjun, skilning, athafnir og nám. Hann segir þetta hluta þess sem hann vildi benda á, að einungis það sem hafi þessa orsakabundnu eiginleika geti búið yfir íbyggni. Það þýði samt ekki að önnur efnisleg og efnafræðileg ferli geti ekki leitt af sér sömu áhrif.[86] Meginpunkturinn hjá Searle er sá að hrein formleg kerfi búi ekki yfir og muni aldrei búa yfir íbyggni ein og sér. Searle segir að öll þau rök fyrir róttækri gervigreind sem hann hafi heyrt einblíni á skuggamyndir vitsmunanna eins og það sé raunveruleikinn.[87]

Í lok greinar sinnar svarar hann ýmsum spurningum til þess að skýra algjörlega afstöðu sína. Vekur athygli að hann telur að hægt sé að búa til vél með manna höndum sem geti hugsað en forsendur þess séu að slík vél sé byggð upp af efni sem er mun líkara okkar eigin líkömum en tölvur eru. Vélin byggi þá yfir taugakerfi og taugafrumum og öllu tilheyrandi sem væri líkt

[85] Sama rit, 79.
[86] Sama rit, 80.
[87] Sama rit, 81.

mannslíkamanum. Hann telur einnig mögulegt að framkalla meðvitund, í raun hvað sem vera skal, með öðrum efnasamböndum en þeim sem mannslíkaminn samanstendur af.[88]

Searle áréttar í lok greinar sinnar að líking þess efnis að samband hugar og heila sé eins og forrits og vélbúnaðs bregðist í ýmsum atriðum. Í fyrsta lagi sé það sú staðreynd að setja megi forrit í mismunandi stafrænar tölvur eða vélar þannig að stundum virðist afkáralegt og útilokað að líta svo á að þær öðlist með því íbyggni. Dæmi um það er að koma forriti sem virðist skilja sögur á kínversku fyrir í vatnslögnum. Í öðru lagi nefnir hann að forrit séu eingöngu formleg sem greini þau frá íbyggnum hugarferlum. Sú skoðun að það rignir skilgreinist ekki sem ákveðin formgerð heldur sem hugmynd sem ræðst af veraldlegum staðreyndum. Hugmyndina má setja fram á ótal setningafræðilega vegu í mismunandi málkerfum og því hefur hún enga formgerð í þeim skilningi. Í þriðja lagi segir Searle að hugarástand sé afurð heilastarfseminnar en forritið ekki afurð tölvunnar.[89]

Searle reynir enn fremur að svara því hvaðan þessi algengi misskilningur komi að formleg forrit séu uppistaða hugarferla. Hann er ekki viss hvað veldur en telur að hugmyndin um að tölvur séu eftirlíkingar af heilastarfsemi, sem var ekki upphaflegt markmið með þeim, hefði átt að vekja tortryggni. Engum detti í hug að tölvulíkan af báli kveiki eld út frá sér. Því sé erfitt að átta sig á hvers vegna nokkrum detti í hug að tölvulíkan af skilningi geti skilið. Hann veltir upp þeim möguleika að ákveðinn misskilningur sé fólginn í margræðum skilningi á hugtakinu upplýsingavinnsla. Þannig sé því haldið fram að hugurinn stundi upplýsingavinnslu líkt og tölvur vinni með upplýsingar. Searle gerir greinarmun á þessu tvennu með því að benda á að tölva býr yfir setningafræði en ekki merkingarfræði. Sé tölva látin leggja saman 2 og 2, þá gefi hún okkur 4. Hún hafi hins vegar enga hugmynd um að 4 merki 4. Þegar við leggjum saman 2 og 2 og fáum 4 þá leggjum við merkingu í hvað felst í hverjum þætti. Sést vel á þessu dæmi hve ólík upplýsingavinnslan er. Searle segir að annað hvort sé upplýsingavinnsla skilgreind með eða án íbyggni.[90] Ef íbyggni er nauðsynleg upplýsingavinnslu þá möndli tölvan aðeins með tákn en ef ekki, þá vinni tölvan úr upplýsingunum en aðeins í sama skilningi og samlagningarvél, hitaskynjarar og magar. Þannig velti það á manni en ekki vél að túlka inntakið og úttakið sem upplýsingar. Einnig bendir hann á að atferlis- og verkhyggja sé enn ríkjandi innan gervigreindarfræða og fyrir þá

[88] Sama rit, 81.
[89] Sama rit, 82-83.
[90] Sama rit, 83-84.

hefð sé Turingprófið dæmigert. Tilhneiging sé til þess að eigna því sem svipi til vitsmuna manna hugarástand. Searle telur að yrði atferlis- og verkhyggjunni kastað fyrir róða myndi verða auðveldara að sneiða hjá ruglingi milli eftirlíkinga og eftirmynda. Verkhyggjuna sé hins vegar áhugavert að skoða nánar. Ein óorðuð forsenda hennar samkvæmt Searle er að heilinn skipti hugann engu máli. Í róttækri gervigreind er forritið það eina sem skiptir máli. Markmiðið er að skapa huga með forriti. Ef hugurinn er ekki bæði röklega og raunverulega aðgreinanlegur frá heilanum þá er engin von um að skapa hið hugræna með forriti. Grunnforsenda róttækrar gervigreindar er að hugarferli séu einfaldlega reiknanlegar aðgerðir á formlegum táknum, þ.e. tölvuforrit, og að heili sé ekkert annað en ein þeirra mörgu véla sem geta keyrt forritið.[91]

Í ljósi þessarar róttæku tvíhyggju verður niðurstaða Searles skýr. Einungis ákveðnar vélar, svo sem heilar, geta hugsað. Róttækri gervigreind takist því ekki, með því að hunsa að mestu mannvélina, að segja mikið um hugann. Viðfangsefni hennar samkvæmt eigin skilgreiningu séu forrit sem eru ekki vélar. Tvíhyggjan er oft varin, segir Searle, með því að halda því fram að heilinn sé stafræn tölva. Það segir hann gefa augaleið, allt sé stafræn tölva. Hins vegar felist íbyggni ekki í því að keyra tölvuforrit eins og dæmið um kínverska herbergið sýnir fram á og geti því ekki byggst á forriti um það hvernig heilinn framleiði íbyggni, því að ekkert forrit sé nægjanleg forsenda þess.[92]

[91] Sama rit, 84-85.
[92] Sama rit, 86.

III. hluti

Í I. hluta hér að framan var bent á að hugtakið gervigreind vísaði til vélar sem líkti eftir vitrænu hátterni sem menn tengja við mannshugann. Vegna framfara á sviðinu aukist sífellt geta vélanna til að hegða sér á hátt sem við myndum venjulega telja vísbendingu um hugsun. Spurningin er hvort megi eigna vél, sem maðurinn býr til og forritar, eiginleika sem greinir manninn frá öðrum dýrum. Hvort megi kalla hina fullkomnustu gervigreind, nú þegar eða í óljósri framtíð, hina vitibornu vél, vél með meðvitund, skoðanir og hugsun. Og hvort sé réttlætanlegt að fela slíkri vél að framkvæma öll mannanna verk, hver sem þau eru og við hvaða aðstæður sem er.

Turing setti m.a. fram tillögur að því með hvaða skilyrðum megi halda fram að gervigreind teljist hugsa. Weizenbaum áleit að slík afstaða smættaði manninn. Hann væri ekki afsprengi einbers „upplýsingavinnslukerfis“ heldur annað og meira en það. Lífveran maðurinn væri skilgreind út frá þeim úrlausnarefnum sem hún stæði andspænis. Vél gæti aldrei nokkurn tímann verið látin standa andspænis þeim vandamálum sem maðurinn stæði frammi fyrir.[93]

Aðrir fræðimenn, Kurzweil og Boström, telja líkur á því að gervigreind geti orðið jafnoki mannsins eða tekið honum fram þegar tæknin leyfir að líkt verði mjög nákvæmlega eftir taugaboðum heilastarfsemi. Slíkri þróun fleygi fram. Í hákirkjukenningunni, sem Þorsteinn Gylfason eignar í höfuðatriðum Dennet, felst auk þess að mannleg hugsun sé úrvinnsla úr boðum, úrvinnsla boðanna sé reikningslist og að reikningsteiknin hafi merkingu sem tengir hugsun fólks við heiminn.[94]

Í samræmi við viðhorf Weizenbaums hefur greind vélarinnar oftlega verið ofmetin eða greind mannsins vanmetin – nema hvort tveggja sé. Vél á ekki fjölskyldu og vini, tilheyrir ekki samfélagi, vex ekki úr grasi með öðrum af sömu tegund, eignast ekki maka, er sneydd allri ástríðu og skortir þegar af þessum ástæðum margt það sem skilgreinir manninn sem vitiborna eða hugsandi veru.

Í II. hluta var ítarlega fjallað um hugartilraun Searles um kínverska herbergið. Niðurstaða hans er sú að vél geti ekki, þegar öllu er á botninn hvolft, búið yfir íbyggni. Munurinn á samlagningarvél og gervigreind sé stigsmunur en ekki eðlismunur.

[93] Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*, 203.
[94] Þorsteinn Gylfason. „Teikn og tákn.“ *Stúdentablaðið* 4/61, 18.

Í stuttu máli má segja að Searle og Weizenbaum séu þeirrar skoðunar að maðurinn geti ekki skapað veru í sinni mynd. Gervigreind sé ýmsum takmörkunum háð þótt hún gæti fræðilega séð leikið margt eftir sem maðurinn getur gert vegna vitsmuna sinna. Vél sýnir ekki samúð eða skilning, jafnvel þó að hún sé forrituð til að líkja eftir viðbrögðum á grundvelli samúðar eða skilnings við ákveðnar aðstæður. Hér er einnig freistandi að velta vöngum yfir ábyrgð á háttsemi gervigreindar sem hún getur eðli máls samkvæmt ekki borið sjálf. Í skáldsögu Marys Shelley um Frankenstein er siðfræði gervigreindar lykilatriði. Ef unnt er að smíða vél sem hefur til að bera greind, getur hún þá einnig búið yfir tilfinningum? Ef hún finnur til, ber hún þá réttindi og skyldur eins og maður? Slíkar hugrenningar virðast vera jafn fjarstæðukenndar og þær að ætla skepnum að svara til saka fyrir glæpi sem þær hafa framið. Annað eins og þvíumlíkt var þó ekki framandi íbúum meginlands Evrópu fyrr á öldum þar sem finna má mörg dæmi þess að höfðað væri mál gegn dýrum og þau pyntuð „til sagna“ á strekkingarbekknum, ekki vegna þess að nokkur rannsóknardómari léti sér raunverulega til hugar koma að þau myndu játa heldur til þess að uppfylla skilyrði málsmeðferðar þess tíma réttarfars áður en þau væru látin þola dóm og refsingu.[95] Til eru heimildir fyrir því að asni hafi verið dæmdur í Tyrklandi fyrir að valda barni líkamstjóni á fyrri hluta 20. aldar.[96]

Skilningur á muninum á réttu og röngu er forsenda sakhæfis. Íslensk hegningarlög geyma ákvæði um að mönnum undir 15 ára aldri skuli ekki refsað og ekki þeim sem sökum geðveiki, andlegs vanþroska eða hrörnunar, rænuskerðingar eða annars samsvarandi ástands voru alls ófærir á þeim tíma, sem þeir unnu verkið, til að stjórna gerðum sínum.[97]

Gefur augaleið að vél, hversu fullkomin sem hún er, getur ekki axlað skyldur á borð við þessar, ekki frekar en tyrkneskur asni. Ávallt yrði maður ábyrgur fyrir verkum sem vél ynni, því að vélin er skynlaus.

Ýmsir virtir fræðimenn hafa eigi að síður mikla trú á gervigreind. John McCarthy, tölvunarfræðingur og einn upphafsmanna gervigreindar og höfundur hugtaksins (e. artificial intelligence), hélt því fram í rökræðu við Weizenbaum að allt sem dómarar í réttarkerfinu vissu og ynnu eftir væri mögulegt að þýða á tölvuforrit. Því væri fullkomlega eðlilegt að reyna

[95] Eward Payson Evans, *The Criminal Prosecution and Capital Punishment of Animals* (London: William Heinemann, 1906), 139.
[96] Páll Sigurðsson, *Svipmyndir úr réttarsögu: Þættir um land og sögu í ljósi laga og réttarframkvæmdar* (Reykjavík: Skjaldborg, 1992), 128.
[97] Sbr. almenn hegningarlög, nr. 19/1940, 14. og 15. gr.

að skapa vélar sem gætu tekið ákvarðanir í dómsmálum.[98] Colby, Watt og Gilbert þóttust hafa skapað meðferðartæki sem gæti með tímanum leyst sálfræðinga af hólmi, eins og áður greinir.[99] Thomas W. Simpson og Vincent C. Müller færðu rök fyrir því að það væri siðferðilega réttlætanlegt að beita sjálfstýrðum vígvélum sem tækju eigin ákvarðanir um að ráðast gegn óvini í hernaði.[100] Sameiginlegt með réttarkerfinu, heilbrigðiskerfinu og m.a.s. stríðsrekstri er að samfélag manna hefur komið sér saman um siðferðilegar grundvallarreglur sem gilda á þessum sviðum. Efast má um að gervigreind verði fær um að vega slík álitaefni inn í aðgerðir sínar. Líkast til mun hún fara stystu og hagkvæmustu leiðina að hverri lausn og láta sér hverja fyrirstöðu í léttu rúmi liggja.

Að mínu mati hafa verið færð sannfærandi rök fyrir því að gervigreind sé alls ófær um íbyggið hugarástand, skilning og samúð. Þessi augljósi grundvallarmunur á mönnum og vélum, að þeir eru líkamsverur og félagsverur en þær ekki, tel ég að skýri hvers vegna gervigreind geti ekki tekið sæti dómara, læknis eða herforingja. Upplýsingavinnsla ein og sér er takmörkuð þó að hún sé öflug á tilteknum sviðum. Án samkenndar og skynjunar verður hún fljótt samfélagsmein.

Í október árið 2016 kom út grein rannsakenda við háskólana í Lundúnum, Sheffield og Pennsylvaníu þar sem þeir greindu frá því að hafa búið til gervigreindarforrit sem tókst með 79% nákvæmni að segja til um hvernig dómar Mannréttindadómstóls Evrópu myndu falla. Aðferð tölvuforritsins byggist á nýjustu tækni í tungumálagreiningu og sjálfstæðu námi véla. Niðurstaða rannsóknarinnar benti til þess að staðreyndir í dómsmálum, sem auðvelt er fyrir tölvur að vinna með, séu mikilvægasti þátturinn við úrlausn þeirra.[101] Því virðist sem hugmynd McCarthys sé smám saman að verða að veruleika. Vissulega er það vísindalegt afrek að vél takist að spá fyrir um úrslit dóma með þessum hætti en er það vísbending um að vélin eigi að dæma í málum – jafnvel þótt árangurinn væri enn betri, e.t.v. 99kant%? Hverjir væru kostirnir við það? Í fljótu bragði má geta sér þess til að tölvan komist ávallt að hlutlægri niðurstöðu og dragi ekki taum annars málsaðila á kostnað hins. En tölva í dómarasæti yrði tæplega fær um

[98] Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*, 207.
[99] Colby, Watt og Gilbert, „A Computer Method of Psychotherapy: Preliminary Communication," *The Journal of Nervous and Mental Disease* 2/142, 148-152.
[100] Simpson og Müller, „Just War and Robot's Killings," *Philosophical Quarterly* 33/263, 302-322.
[101] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daneil Preoţiuc-Pietro og Vasileios Lampos, „Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective." PeerJ Computer Science https://peerj.com/articles/cs-93/.

að setja sig í spor þeirra sem hún dæmir. Skilningur gervigreindarinnar á aðstæðum þeirra sem leita til dómstóla er enginn. Ekki er hægt að höfða til tilfinninga hennar.

Sigurður Líndal segir að lög séu „sjaldan svo afdráttarlaust orðuð að mat og virðing þess sem túlkar þau og framfylgir þeim, svo sem dómara eða handhafa stjórnvalds, hafi ekki einhver áhrif. Þannig getur réttlætisvitund og siðgæðisviðhorf sem ríkjandi eru í þjóðfélagi og lífsskoðun þess sem lög túlkar, til dæmis dómara, ráðið því hvernig regla er endanlega mótuð“.[102] Hafsteinn Dan Kristjánsson vitnar m.a. í Ronald Dworkin í umfjöllun sinni um siðferði og lögfræði og segir að löggjöf og önnur framkvæmd sé túlkuð í ljósi þeirra siðferðilegu meginreglna sem réttlæta hana til að komast að niðurstöðu um lög.[103] Slíkum rökum er teflt fram gegn strangri vildarréttarkenningu (lagalegum pósitífisma) sem hafnar því að siðferði sé hluti af lögum. Skúli Magnússon og Hafsteinn Þór Hauksson orða það þannig að gagnrýnendur valdboðskenningar um lögin hafi hafnað raunvísindalegri nálgun: „Í stað þess að „smætta“ samfélagslegt fyrirbæri eins og lögin í staðreyndir væri fræðimanninum nauðsynlegt að setja sig í spor þeirra sem lifa og hrærast í lögunum og líta á reglur sem réttlætingu eða ástæðu athafna, en ekki aðeins sem fyrirboða um viðurlög. [...] Samfélagslegur veruleiki, þ.á m. lögin, er byggður upp af afstöðu, athöfnum og orðræðu lifandi og hugsandi fólks sem tekur, eða getur a.m.k. tekið, vitræna afstöðu til afstöðu sinnar, athafna, o.s.frv.“[104]

Augljóslega er þetta ekki tæmandi umræða um þessa réttarheimspekilegu togstreitu en hún þjónar þeim tilgangi að sýna fram á að lagaleg úrlausnarefni eru oft flóknari en svo að þær megi leiða fram með reikningslistum. Mál eru misjafnlega augljós í eðli sínu. Þótt líta megi á spádómsgáfu gervigreindar um dóma Mannréttindadómstólsins sem vísindalegt afrek og vísbendingu um að dómarar víki sjaldan af braut strangrar lagalegrar aðferðar við að komast að niðurstöðu, þá brást gervigreindin hvað sem öðru líður í 21% tilvika. Voru það þau tilvik sem voru matskennd umfram önnur?

[102] Sigurður Líndal, *Um lög og lögfræði: Grundvöllur laga – réttarheimildir*, (Reykjavík: Hið íslenzka bókmenntafélag, 2010), bls. 34-35.
[103] Hafsteinn Dan Kristjánsson, *Að iðka lögfræði: Inngangur að hinni lagalegu aðferð*, (Reykjavík: Bókaútgáfan CODEX, 2015), bls. 49.
[104] Skúli Magnússon og Hafsteinn Þór Hauksson, „Klassískur vildarréttur – valdboðskenningin um lögin,” *Úlfljótur* 2/64 (2012): 208.

Dómstólar hafa í ljósi tíðaranda og nýrra sjónarmiða um réttlæti gerbreytt túlkun sinni á lögum.[105] Slíkur sveigjanleiki er nauðsynlegur í ljósi grundvallnarreglna um að dómstólar eigi að skila réttlátri niðurstöðu. Tölva gerir ekki ráð fyrir því.

Taka má dæmi af heilbrigðissviði sömuleiðis. Tölvufyrirtækið IBM er langt komið með þróun gervigreindar sem nefnist Watson. Watson eru margir vegir færir og einskorðast ekki við afmörkuð verkefni. Hann er gæddur þeim eiginleika að geta lesið mikið magn upplýsinga og túlkað þær.[106] Nú þegar er hann nýttur í heilbrigðisþjónustu. Hann er heilbrigðisstarfsfólki mikilvæg hjálparhella. Hann geymir gríðarlegt magn rannsóknarniðurstaðna á heilbrigðissviði og er sífellt mataður á nýjum gögnum. Yfirsýn yfir slíkt gagnamagn og úrvinnsla þess er klárlega einn af helstu kostum tölva almennt og gervigreindar sérstaklega. Þessi tækniþróun gerir heilbrigðisstarfsfólki kleift að sinna starfi sínu betur en áður og kalla fram upplýsingar sem einum manni yrði ofviða á jafnskömmum tíma. Læknisþjónusta batnar þar með. Watson er þó enginn læknir þótt hann rati rétt á sjúkdómsgreingar og meðferðarrúrræði í einhverjum tilvikum.

Engum blandast hugur um að gervigreindarforrit eru dýrmæt hjálpartæki sem geta aukið afköst. Þau hljóta samt sem áður að vera þjónn mannsins en ekki öfugt. Colby, Watt og Gilbert álitu að gervigreind gæti komið í stað sálfræðings. Sálfræðimeðferð væri afurð upplýsingavinnslu sem leiddi til ákvarðana eftir ákveðnum reglum.[107] Hér virðist sálfræðingurinn með sama hætti og dómarinn áður smættaður til að koma til móts við hugmyndir fræðimannanna um möguleika gervigreindar. Oftrú á tækni er líkleg til að leiða okkur inn á hættulegar brautir. Hin hliðin á þeim peningi er vantrú á manninn.

Engin vél mun þurfa að standa andspænis úrlausnarefnum sem eru tilkomin vegna líffræðilegra og tilfinningalegra þarfa hennar. Weizenbaum segir að umfang mannlegrar greindar ákvarðist að mestu af mannúð mannsins og að sérhver önnur greind, hversu fullkomin sem hún kann að vera, verði ávallt utan marka hins mannlega sviðs.[108] Meðferðarsamband er

[105] Sjá t.d. dóm Hæstaréttar Íslands 21. október 2001 í máli nr. 129/2001, þar sem ekki þóttu efni til þess lengur að halda við þeirri dómstólamynduðu reglu íslensks skaðabótaréttar að áhættutaka farþega sem tekur sér far með ölvuðum ökumanni leiði til niðurfellingar bótaréttar. Þess í stað var niðurstaða málsins látin ráðast af reglum um eigin sök í 2. mgr. 88. gr. umferðarlaga nr. 50/1987.
[106] David Ferrucci, Anthony Levas, Sugato Bagchi, David Gondek og Erik T. Mueller, „Watson: Beyond Jeopardy!“ *Artificial Intelligence* 199 (2013): 104.
[107] Colby, Watt og Gilbert, „A Computer Method of Psychotherapy: Preliminary Communication,” *The Journal of Nervous and Mental Disease* 2/142, 150.
[108] Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*, 223.

mikilvægt í hjúkrun og læknisþjónustu.[109] Heilbrigðisstarfsfólk þarf að geta mætt skjólstæðingum sínum á félagslegum, siðferðilegum og sammannlegum nótum. Þegar vel tekst til um persónuleg tengsl stuðlar það að gagnkvæmu trausti sem sjúklingurinn hagnast á, ekki síst í því ljósi að heilbrigðisstarfsmaðurinn lætur sér frekar annt um hagi hans en ella.[110] Þessa færi sjúklingur á mis í vélrænu umhverfi.

Að síðustu er rétt að athuga gervigreind í hernaði. Þróun hennar á því sviði hefur vakið alþjóðlega umræðu um siðferðileg álitamál og takmörk sem kann að þykja nauðsynlegt að setja henni. Nú er ekki aðeins spurt hvað gervigreindin „geti gert“ heldur einnig hvort hún „eigi að gera“.

Með nýrri tækni er mögulegt að búa til vélar sem geta valið skotmörk án mannlegs atbeina og eytt þeim.[111] Raunhæft er orðið að hanna sjálfvirkar og sjálfstýrðar vígvélar sem ráðast gegn þeim sem þær skilgreina sem óvini.[112] Þær eru þeim augljósu kostum búnar í stríðsrekstri að enginn ferst ef þeim er grandað. Herlið getur beðið í öruggri fjarlægð á meðan gervigreindarvopnin ráðast á fjandmennina enda eigi þau að geta greint hverjir þeir eru.[113] Það er m.ö.o. ætlast til að gervigreindin virði upp á eigin spýtur alþjóðareglur um beitingu vopnavalds, *jus in bello*, þ.m.t. mannúðarlög og Genfar-sáttmála, sbr. t.d. 48. gr. viðauka hans frá 1977 sem kveður á um að greina skuli á milli almennra borgara og bardagamanna í vopnuðum átökum. Ókosturinn er að gervigreindin lækkar þröskuldinn fyrir árásir.[114]

Viðbrögðin hafa verið sterk á alþjóðavettvangi. Flestir tala gegn þessari þróun. Mannúðarsamtök hafa dregið í efa að vél geti raunverulega gert greinarmun á bardagamönnum annars vegar og almennum borgurum hins vegar eða jafnvel bardagamönnum sem hafa gefist upp. Árásir slíkra véla séu því handahófskenndar og

[109] Sbr. t.d. Elizabeth J. Pask, „Trust: An essential component of nursing practice —implications for nurse education.“ *Nurse Education Today* 3/15 (1995): 190-195 og Hildegard E. Peplau, „Interpersonal relations: A theoretical framework for application in nursing practice.“ *Nursing Science Quarterly* 1/5 (1992): 13-18.
[110] Gary S. Carr, „Negotiating trust: A grounded theory study of interpersonal relationships between persons living with HIV/AIDS and their primary health care providers.“ *Journal of the Association of Nurses in AIDS Care* 2/12, 41.
[111] Ronald C. Arkin, *Governing Lethal Behavior in Autonomous Robots* (New York: Chapman&Hall, 2009), 9.
[112] Future of Life Institute. „Autonomous Weapons: an Open Letter from AI & Robotics Researchers.” Future of Life http://futureoflife.org/open-letter-autonomous-weapons#signatories.
[113] Arkin, *Governing Lethal Behavior in Autonomous Robots*, 9.
[114] Future of Life Institute. „Autonomous Weapons: an Open Letter from AI & Robotics Researchers.” Future of Life http://futureoflife.org/open-letter-autonomous-weapons#signatories.

ólögmætar.[115] Auk þess birtu margir helstu gervigreindarsérfræðingar heims opið bréf 28. júlí 2015 á alþjóðlegri ráðstefnu um gervigreind (International Joint Conference on Artificial Intelligence (IJCAI)) þar sem lagt var til að framþróun og framleiðsla gervigreindar til notkunar í vígvélum yrði bönnuð.[116] Alþingi Íslendinga brást við þessu bréfi með þingsályktun. Í henni segir: „Alþingi ályktar að lýsa stuðningi við áform og viðræður um alþjóðlegt bann við framleiðslu og beitingu sjálfvirkra og sjálfstýrðra vígvéla og felur ríkisstjórninni að fylgjast með þróun þessara mála á vettvangi Sameinuðu þjóðanna og annars staðar þar sem það á við."[117]

Bann við slíkum vélum er m.a. rökstutt með því að þær gætu lent í höndum hryðjuverkahópa.[118] Önnur rök lúta að alþjóðalögum en eins og fyrr er vikið að verður gervigreind ekki dregin til ábyrgðar fyrir verk sín. Á bak við vélina verði að vera maður. Ein lykilforsenda *jus in bello* um réttlæti í stríði snýst einmitt um að hægt sé að láta þann sem brýtur af sér í stríði svara til saka. Robert Sparrow heimspekingur bendir á að lögmálið um mismunun í hernaði, sem gerir kröfu um að andstæðingar geri greinarmun á lögmætum og ólögmætum skotmörkum, sé skýrt dæmi um hversu mikilvægt það er að hægt sé að draga hina brotlegu til ábyrgðar. Hervald án ábyrgðar eða aðhalds af refsireglum breyti eðli styrjalda til hins verra. Það sé kjarni kantísks siðferðis að borin sé virðing fyrir manninum.[119] Skilningur er forsenda þess að manngildi sé haft í hávegum. Jafnvel þótt ábyrgðarreglur gerðu ráð fyrir að einstaklingar eða ríki bæru ábyrgð á lögbrotum gervigreindar yrði sífellt erfiðara að koma almennilega lögum yfir þá vegna aukins sjálfstæðis vélanna.[120]

Í skyldusiðfræði Immanuels Kant er því haldið fram að maður eigi alltaf að breyta eftir þeirri lífsreglu sem maður vildi að aðrir breyttu eftir sömuleiðis.[121] Þetta er skilyrðislausa skylduboð Kants, einnig orðað þannig að maður eigi alltaf að koma fram við aðra menn sem markmið í sjálfu sér en ekki einvörðungu tæki í einhverjum öðrum tilgangi.[122] Ekki verður

---

[115] Sbr. Bonnie Docherty, *Losing humanity: The case against killer robots* (New York: Human Rights Watch, 2012), 4 og 24.
[116] Future of Life Institute. „Autonomous Weapons: an Open Letter from AI & Robotics Researchers." Future of Life http://futureoflife.org/open-letter-autonomous-weapons#signatories.
[117] Alþt. 2015-2016, A-deild, þskj. 1486 – 68. mál.
[118] Future of Life Institute. „Autonomous Weapons: an Open Letter from AI & Robotics Researchers." Future of Life http://futureoflife.org/open-letter-autonomous-weapons#signatories.
[119] Robert Sparrow, „Killer robots," *Journal of Applied Philosophy* 1/24 (2007): 67.
[120] Sama rit, 73.
[121] Immanuel Kant, *Grundvöllur að frumspeki siðlegrar breytni*, ritstj. Vilhjálmur Árnason og Ólafur Páll Jónsson, þýð. Guðmundur Heiðar Frímannsson (Reykjavík: Hið íslenzka bókmenntafélag, 2003), 140.
[122] Sama rit, 153.

séð að hægt sé að gera slíkar kröfur til vélar. Vél hikar ekki við að ná markmiði sínu en mannslíf getur verið háð slíku hiki, einkum og sér í lagi í vopnuðum átökum.

Þegar sjálfvirkri og sjálfstýrðri vígvél er beitt er í raun handahófskennt hver er felldur og hver ekki. Gildir einu þótt um almenna borgara sé að ræða en ekki hermenn. Það væri ekki nema einskær óheppni að þessi en ekki hinn yrði drepinn. Notkun þess háttar vopna gefur í skyn, að mati Sparrows, að þetta hafi litla þýðingu og í því birtist djúpstætt virðingarleysi við líf einstakra manna.[123] Vél hefur ekkert aðhald af vanlíðan eða samviskubiti, hvort sem hún nefnist vígvél eða samlagningarvél. Það eru forréttindi mannsins.

[123] Sparrow, „Killer robots," *Journal of Applied Philosophy* 1/24, 68.

Lokaorð

Gervigreind er heillandi og ögrandi í senn vegna þess hversu mannleg hún virðist vera. Mörg þekkjum við hina hljómþýðu rödd hjálparhellunnar í símanum sem ráðleggur okkur um veitingastaði í nágrenninu og hvernig við komumst þangað. Í fræðasamfélaginu verður vart tveggja andstæðra póla í afstöðunni til gervigreindar. Sumir vilja eigna henni mannlega hegðun, jafnvel ofurmannlega hegðun, aðrir flokka hana einfaldlega með öðrum vélum eins og einfalda samlagningarvél. Vissulega fer ákveðinn ljómi af gervigreindinni þar með. Spurningin er hvort það sé ekki rétt og eðlilegt þegar litið er til þeirra margháttuðu nota sem megi hafa hana til, hvort gervigreind sé ekki ofmetin á kostnað mannlegrar greindar og skilnings, þannig að henni verði ekki gert hærra undir höfði en hún á skilið.

Þótt færni gervigreindar virðast vera fá takmörk sett mun hana ávallt skorta mannlega eiginleika á borð við skilning, samúð og íbyggið hugarástand. Í því felast takmörk hennar. Jafnvel þótt hún „geti gert“ er engan veginn sjálfsagt að hún „eigi að gera“. Slíkir mannlegir eiginleikar eru nauðsynlegir þegar mikið ber undir eins og dæmin sanna hér að framan um dómskerfið, heilbrigðisþjónustuna og hernaðinn. Gervigreindina skortir þá eiginleika og getur því ekki talist jafnoki mannsins. Hún á ekki til réttlætiskennd dómarans, samúð læknisins og ályktunarhæfni hermannsins. Gervigreind er ekki fær um mannleg tengsl, ábyrgðartilfinningu og skilning á afleiðingum aðgerða sinna.

Ekki á að fela fyrirbæri, sem ekki er fært um að skilja og bera ábyrgð, ábyrgðarfull verkefni sem krefjast skilnings. Það fer í bága við viðteknar siðferðiskröfur í samfélagi manna. Gervigreind á að vera þjónn mannsins og þannig getur hún auðveldað honum ýmis verkefni. Það býður hættunni heim að fela henni alfarið viðkvæm verkefni eins og læknis- og hjúkrunarþjónustu, dómstörf eða stríðsrekstur. Gervigreind er þannig ýmsum takmörkunum háð, hvað sem líður hugsanlegri færni hennar, og því er óviðeigandi að hún sinni öllum mannanna verkum.

Með því að gaumgæfa rökin fyrir því hvað gervigreind skortir verður ljóst hver munurinn á henni og öðrum vélum er. Hann er stigsmunur en ekki eðlismunur.

Heimildaskrá

Aletras, Nikolaos, Dimitrios Tsarapatsanis, Daneil Preoţiuc-Pietro og Vasileios Lampos. „Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective." PeerJ Computer Science, birt 24. október 2016, https://peerj.com/articles/cs-93/ (skoðað 17. desember 2016).

Arkin, Ronald C. *Governing Lethal Behavior in Autonomous Robots*. New York: Chapman&Hall, 2009.

Boström, Nick. „How Long Before Superintelligence?" *Linguistic and Philosophical Investigations* 1/5 (2006): 11-30.

Carr, Gary S. „Negotiating trust: A grounded theory study of interpersonal relationships between persons living with HIV/AIDS and their primary health care providers." *Journal of the Association of Nurses in AIDS Care* 2/12 (2001): 35-43.

Colby, Kenneth Mark, James B. Watt og John P. Gilbert. „A Computer Method of Psychotherapy: Preliminary Communication." *The Journal of Nervous and Mental Disease* 2/142 (1966): 148-152.

Descartes, René. *Orðræða um aðferð.* Þýðandi Magnús G. Jónsson. Reykjavík: Hið íslenzka bókmenntafélag, 1998.

Docherty, Bonnie. *Losing humanity: The case against killer robots*. New York: Human Rights Watch, 2012.

Evans, Eward Payson. *The Criminal Prosecution and Capital Punishment of Animals*. London: William Heinemann, 1906.

Ferrucci, David, Anthony Levas, Sugato Bagchi, David Gondek og Erik T. Mueller. „Watson: Beyond Jeopardy!" *Artificial Intelligence* 199 (2013): 93-105.

Future of Life Institute. „Autonomous Weapons: an Open Letter from AI & Robotics Researchers." Future of Life, birt 28. júlí 2015, http://futureoflife.org/open-letter-autonomous-weapons#signatories (skoðað 3. desember 2016).

Hafsteinn Dan Kristjánsson. *Að iðka lögfræði: Inngangur að hinni lagalegu aðferð*. Reykjavík: Bókaútgáfan CODEX, 2015.

Harrison, Natalie, og Teresa Brewer. „Apple Launches iPhone 4S, iOS 5 & iCloud." Heimasíða Apple Inc., birt 4. október 2011,

https://www.apple.com/pr/library/2011/10/04Apple-Launches-iPhone-4S-iOS-5-iCloud.html (skoðað 8. janúar 2017).

Kant, Immanuel. *Grundvöllur að frumspeki siðlegrar breytni*. Ritstjórar Vilhjálmur Árnason og Ólafur Páll Jónsson. Þýðandi Guðmundur Heiðar Frímannsson. Reykjavík: Hið íslenzka bókmenntafélag, 2003.

Kristinn R. Þórisson. „Vélvitund, meðvitund og sjálfsvitund í kjötvélum og vélmennum." *Veit efnið af andanum? Sjö fyrirlestrar um meðvitundina.* Ritstjórar Steinar Örn Atlason og Þórdís Helgadóttir. Reykjavík: Heimspekistofnun Háskóla Íslands, 2009.

Kurzweil, Ray. *The Singularity Is Near: When Humans Transcend Biology.* New York: Penguin Group, 2005.

Leibniz, Gottfried Wilhelm. *Orðræða um frumspeki.* Ritstjóri Ólafur Páll Jónsson. Þýðandi Gunnar Ágúst Harðarson. Reykjavík: Hið íslenzka bókmenntafélag, 2004.

Minsky, Marvin. *Computation: Finite and Infinite Machines* New Jersey: Prentice-Hall, 1967.

Pask, Elizabeth J. „Trust: An essential component of nursing practice —implications for nurse education." *Nurse Education Today* 3/15 (1995): 190-195.

Páll Sigurðsson. *Svipmyndir úr réttarsögu: Þættir um land og sögu í ljósi laga og réttarframkvæmdar*. Reykjavík: Skjaldborg, 1992.

Peplau, Hildegard E. „Interpersonal relations: A theoretical framework for application in nursing practice." *Nursing Science Quarterly* 1/5 (1992): 13-18.

Russell, Stuart, og Peter Norvig. *Artificial Intelligence: A Modern Approach.* New Jersey: Prentice-Hall, 2009.

Searle, John R. „Hugur, heili og forrit." *Hugur* 1/7. Þýðandi Ólafur Páll Jónsson (1995): 64-86.

Sigurður Líndal. *Um lög og lögfræði: Grundvöllur laga – réttarheimildir*. Reykjavík: Hið íslenzka bókmenntafélag, 2010.

Simon, Herbert Alexander. *The Shape of Automation for Men and Management.* New York: Harper & Row, 1965.

Simpson, Thomas W., og Vincent C. Müller. „Just War and Robot's Killings." *Philosophical Quarterly* 33/263 (2016): 302-322.

Skúli Magnússon og Hafsteinn Þór Hauksson, „Klassískur vildarréttur – valdboðskenningin um lögin," *Úlfljótur* 2/64 (2012): 191-209.

Sparrow, Robert. „Killer robots." *Journal of Applied Philosophy* 1/24 (2007): 62–77.

Turing, Alan M. „Reikniverk og vitsmunir." *Hugur* 1/7. Þýðandi Atli Harðarson. (1995): 32-63.

Weizenbaum, Joseph. *Computer Power and Human Reason: From Judgment to Calculation.* San Francisco: W. H. Freeman and Company, 1976.

Þorsteinn Gylfason. „Teikn og tákn." *Stúdentablaðið* 4/61 (1985): 17-19.

Lög og alþjóðasamningar o.fl.

Almenn hegningarlög, nr. 19/1940.

Alþt. 2015-2016, A-deild, þskj. 1486 – 68. mál.

Genfar-sáttmálinn 12. ágúst 1949, viðauki um vernd almennra borgara í alþjóðlegum vopnuðum átökum, (viðauki I), Genf, 8. júní 1977.

Hæstiréttur Íslands, dómur 21. október 2001 í máli nr. 129/2001.