Hugur | 34. ár, 2024 | s. 12–22

Hugsað með tækninni

Aþena Ýr Ingimundardóttir og Gamithra Marga ræða við Þórhall Magnússon um sköpun og gervigreind

Þórhallur Magnússon er rannsóknaprófessor við Hugvísindasvið Háskóla Íslands og prófessor í tónlist við Sussex háskólann í Brighton. Hann hefur fengist við rannsóknir á sviði sköpunar og gervigreindar, sérstaklega á sviði tónlistar. Í því samhengi hefur hann skrifað bækurnar Sonic Writing: Technologies of Material, Symbolic and Signal Inscription *(New York: Bloomsbury Academic, 2019) og* Live Coding: A User's Manual *(Cambridge: MIT Press, 2022). Ásamt því að skoða hlutverk tækni í mannlegri hugsun og tjáningu hefur hann þróað forrit og tækni til tónlistarsköpunar. Hann stýrir rannsóknarverkefninu* Intelligent Instruments *sem er styrkt af Evrópska rannsóknarráðinu og er hýst við Háskóla Íslands.*

Aþena Ýr Ingimundardóttir og Gamithra Marga ræða hér við Þórhall og beina sjónunum sérstaklega að möguleikum nýrrar gervigreindar til að vera skapandi og þeirra siðferðislegu spurninga sem þá vakna.

Í dag vinnur þú við þróun og rannsókn greindra hljóðfæra innan HÍ. Gætirðu sagt okkur frá ferli þínum og hvernig hann leiddi þig að þessari vinnu með gervigreind og skapandi tónlistartækni?

Ég er með bakgrunn í heimspeki þar sem ég einbeitti mér sérstaklega að tækniheimspeki og hugarheimspeki. Það var sá áhugi sem leiddi mig út í að forrita tölvur seint á síðustu öld, af því að ég hafði áhuga á að skoða hugmyndir um huga, sköpun og greind út frá hugfræðum (e. *cognitive science*). Hér býður gervigreindin og þróun nýrrar skapandi tölvutækni upp á vissa speglun á menningu og listir, og reynir virkilega á skilgreiningar okkar á meginhugtökum innan þeirra. Ég hef alltaf verið með annan fótinn í tónlist og fundist hún vera birtingarmynd menningar og hugmyndafræði, bæði sögulega og landfræðilega. Forritun og tilraunir í tónlist og tækni urðu þannig fyrir mér einskonar hagnýt heimspeki. Ég hafði verið að vinna í London með baskneskum félaga, Enrike Hurtado, að þróun tónlistarforrita sem urðu nokkuð vinsæl kringum 2000, ixi audio, og í tengslum við þá vinnu var mér boðinn styrkur til doktorsnáms í tölvuvísindum og gervigreind við Sussex háskólann í Brighton. Sussex hefur verið miðpunktur rannsókna í skapandi gervigreind og mætti þar nefna frumkvöðlastarf vinkonu minnar, Margaret Boden, en hún skrifaði árið 1990 bókina *The Creative Mind:*

Myths and Mechanisms þar sem hún þróar kenningar um sköpun og skoðar hugmyndina um að tölvur geti verið skapandi. Eftir að hafa starfað við nokkra breska háskóla, þá endaði ég við tónlistardeild Sussex þar sem ég þróaði nýja námsbraut í tónlist og tækni. Sussex háskólinn var stofnaður með það að markmiði að vera þverfaglegur og því hef ég unnið mikið með fólki í tölvuvísindum og hugfræðum. Þannig unnum við að rannsóknum í nýjum forritunarmálum fyrir tónlist og þróuðum skapandi gervigreind af ýmsu tagi.

Rannsóknir mínar hafa verið birtar í bókinni *Sonic Writing: Technologies of Material, Symbolic and Signal Inscriptions* sem kom út hjá Bloomsbury Academic árið 2019. Bókin rekur sifjafræði tónlistartækni aftur í aldir og sýnir hvernig sú tækni sem við notum í dag á sér rætur í eldri tækni, hvort sem það varðar hljóðfærasmíði, táknkerfi tónlistar eða hljóðritun hennar (þetta er það sem ég vísa til með *material, symbolic* og *signal inscriptions* í titlinum). Í bókinni varpa ég ljósi á þekkingarfræðilegt eðli tækninnar; hvernig við hugsum í gegnum hana og hvernig hún rammar inn mögulegar hreyfingar og hugsanir. Þegar ég lagði lokahönd á bókina var sú bylting sem hófst fyrir nokkrum árum á sviði spunagreindar (e. *generative AI*) ekki komin fram. Þó að gervitauganet (e. *artificial neural network*) séu áratuga gömul, þá var það fyrst með hröðum örgjörvum, gífurlegu magni af aðgengilegum gögnum og svo nýjum arkitektúr í hönnun tauganeta á síðustu nokkrum árum sem við förum að sjá þá mögnuðu þróun sem hefur orðið á spunagreind í dag, t.d. með ChatGPT eða Claude í textavinnslu, Midjourney eða Dall-E í myndvinnslu, og Udio, Suno eða Stable Audio í tónlist.

En gervitauganetin voru á radarnum og það var með þennan bakgrunn í rannsóknum á tónlistartækni og tækniheimspeki sem ég skrifaði umsókn til Evrópska rannsóknaráðsins um að skoða áhrif gervigreindar á samfélagið í gegnum tónlist. Verkefnið *Intelligent Instruments: Understanding 21st-Century AI Through Creative Music Technologies* er þverfaglegt, með kenningalegan grunn í hugvísindum en aðferðafræði tónlistar og tölvuvísinda. Um er að ræða tilraunakennd hugvísindi, þar sem við notum nýja tækni til að svara spurningum um atbeina og virkni gervigreindar ásamt því að skoða hvaða orðræða og hugtakamynstur verða til þegar fólk fer að hugsa með nýrri tækni. Verkefnið er hýst við Hugvísindasvið Háskóla Íslands með aðsetur í Veröld, Húsi Vigdísar (sjá www.iil.is). Meðfram vinnu minni við HÍ stunda ég líka rannsóknir og kenni við Sussex háskólann í Brighton.

Hvað felur starf þitt við þróun greindra hljóðfæra í sér, hver eru markmið þess og hvernig hefur listaheimurinn tekið verkefninu?

Það skemmtilega við að vinna að rannsóknum á gervigreind er að það eru mjög ólíkar skilgreiningar á því hvað greind er. Út frá vissu sjónarhorni má færa rök fyrir því að stofuplanta sé greindari en þau mállíkön sem við erum að vinna með í dag. Gervigreind á sér langa sögu og gengur út á að nota algrími til þess að bregðast við heiminum eða gögnum úr honum. Á 20. öldinni var mesti fókusinn á táknræna gervigreind (e. *symbolic AI*) þar sem tölvurnar fylgdu reglukerfi skilgreindu af forriturum, oft með aðgang að stórum gagnasöfnum. Þessi tegund

gervigreindar varð aldrei „mennsk“: tölvur urðu góðar í skák, gátu stýrt flóknum kerfum og voru góðar í sérhæfðum verkefnum, en það var alltaf einhver bjánaleg heimska og takmarkanir í þessari „greind“. Vélin gat sigrað Kasparov í skák en ekki raðað kubbum sem hvert þriggja ára barn getur. Það er hins vegar með djúpum gervitauganetum og spunagreind sem stökkbreyting verður í upplifun okkar á gervigreind. Allt í einu eru tölvurnar farnar að tala, búa til myndir og tónlist, og nú er svo komið að við getum ekki vitað svo auðveldlega hvort við séum í samskiptum við manneskju eða vél. Fólk er auðvitað slegið yfir hraðanum í þessari þróun og að mínu mati hafa hugvísindin aldrei verið mikilvægari en akkúrat nú til að bregðast við, skilja áhrifin og vinna í stefnumótun.

Markmið rannsóknarverkefnisins við HÍ er að þróa hljóðfæri með greind, þ.e. hljóðfæri sem hafa atbeina (e. *agency*), minni, vilja, skynjun og viðbragð, og geta þannig brugðist við því sem tónlistarmaðurinn gerir. Þessi hljóðfæri geta þróast, lært á hljóðfæraleikarann, unnið með honum eða veitt viðnám. Þau eru þjálfuð á gögnum sem tónlistarmaðurinn hefur áhuga á að vinna með og eru því orðin viss framlenging á honum. Hér notum við fyrirbærafræði til að skoða hvernig fólk bregst við þessum nýju hljóðfærum og þá í gegnum persónuleg tengsl manneskju og gervigreindar. Við höfum ekki áhuga á að vinna með stór gagnasöfn og þróa forrit þar sem maður ýtir á takka til búa til tónlist, heldur erum við fremur að einbeita okkur að því hvernig einstakir tónlistarmenn geta þjálfað sín eigin hljóðfæri til að verða persónulegri og meira spennandi í notkun. Þetta er munurinn á „big data“ og „small data“, eða í þessu samhengi, iðnaðargervigreindar og persónulegrar gervigreindar.

Viðbrögðin við verkefni okkar hafa almennt verið góð. Sérstaklega þegar fólk skilur að við erum ekki að vinna með risastór gagnasöfn að smíða tækni sem á að koma í stað tónlistarmannsins, heldur fremur að skoða snjallhljóðfærið sem framlengingu á honum. Við vinnum mikið með tónlistarfólki, þverfaglegum hópi vísindamanna og almenningi í þróun þessara hljóðfæra og höfum í því samhengi fengið verðlaun Evrópusambandsins fyrir lýðvísindi (e. *citizen science*), AI in Art verðlaun Ars Electronica og einnig Vísinda og nýsköpunarverðlaun HÍ í flokknum „samfélag“. Það er mjög skemmtilegt að vinna með fjölbreyttum hópi fólks og við gætum í raun ekki unnið svona rannsókn í fílabeinsturni akademíunnar, en samtöl, vinnustofur, notendapróf, þróun, tilraunir og tónleikar eru allt liðir í rannsóknaraðferð verkefnisins. Sagan sýnir okkur að þegar ný tækni kemur fram þá tileinka listamenn sér hana og skoða hvað þeir geta gert við hana. Tónlist hefur ávallt verið í fararbroddi tækninnar (t.d. vatnsorgel Forn-Grikkja, pípuorgel miðalda, eða hljóðgervlar á síðustu öld) og nú er tónlistarfólk að rannsaka gervigreind. Sú tilraun hófst árið 1956 þegar Lejaren og Hiller notuðu tegund vélanáms (e. *machine learning*), Markov Models, til að semja tónlist og framfarirnar hafa verið ótrúlegar síðan þá, sérstaklega á síðustu fimm árum með komu Transformer gervitauganetanna. Fólk er að vissu leyti hrætt við gervigreind, skiljanlega, þar sem stórfyrirtæki eru að nota sköpunarverk og samskipti fólks sem fóður í risalíkön, en hvað okkur varðar, þá hverfur slíkur ótti yfirleitt þegar við útskýrum að við erum við að vinna með ótal ólík algrími og í tilfelli þróunar

líkana þá er um að ræða lítil gagnasöfn sem eru framlenging á listamönnunum sjálfum og undir þeirra stjórn.

Í dag er oft sagt að gervigreind sé í sjálfu sér ófær um það að vera virkilega skapandi, en sé spennandi sem nýtt verkfæri sem hægt er að beita við sköpun. Ert þú sammála þessu eða er þetta kannski einföldun? Hefur vinna þín breytt áliti þínu á sambandi sköpunar og gervigreindar að einhverju leyti?

Allt sem við segjum um gervigreind í dag er einföldun því að það er svo margt í gangi og þróunin er svo hröð. Sumt sem virðist óhugsandi í dag gæti vel gerst eftir nokkur ár eða áratugi. Það sem gerir hlutina enn flóknari er að orðin sem við notum eiga stundum ekki við eða brotna þegar við vörpum þeim á vélar, orð eins og skynjun, skilningur, sköpun og svo framvegis. En nútíma gervigreind er mjög góð í að greina mynstur og formúlur. Það þýðir að ef okkur langar til að búa til tónlist eftir formúlu, þá gæti gervigreindin komið sér vel. Hún gæti t.d. reitt fram svo sannfærandi barrokkfúgur, Deltablús eða Berlínarteknó að við getum ekki vitað fyrir víst hvort að manneskja eða vél standi þar að baki. En þetta er tónlist fortíðarinnar, tónlist sem gervigreindin hefur verið mötuð á. Allt byggir þetta á statistík; að framreikna næstu orð eða tóna út frá þeim sem fyrr hafa komið. Tónlist er hér frekar einfalt dæmi af því að hún er óhlutbundin í eðli sínu, ólíkt tungumálinu. Gervigreindin getur samið sonnettu í líki Shakespeare og oft er það ágætlega gert hvað stíl varðar, en auðvitað kemst innihaldið ekki nálægt list Shakespeares sjálfs.

Hvað er það þá að vera „virkilega skapandi“ eins og þið spyrjið? Maggie Boden sem ég nefndi hér að ofan talaði um samsetjandi (e. *combinatorial*), kannandi (e. *explorative*) og umbreytandi (e. *transformational*) sköpun. Ljóst er að spunagreindin stendur sig ágætlega í tveimur fyrstu tegundum sköpunar, en getur hún virkilega breytt reglum listarinnar, komið með ný sjónarhorn, nýjar aðferðir og lausnir? Getur hún tjáð ástand manneskjunnar í flóknum heimi? Þetta er það sem góð list gerir meðal annars og það er langt í að gervigreindin geti slíkt. Til þess þyrfti hún helst að vera í líkama og partur af samfélagi sem er í þróun. Hún þyrfti að skynja heiminn, finna til, vera með hormóna sem taka okkur í allar áttir, vera þreytt og pirruð, glöð og spennt, og vera háð tímanum og dagrytma eins og við. Ætli hana þyrfti ekki að dreyma líka? Allt annað er hermun, og þá *simulacrum* en ekki *mimesis*. Ég vil þó taka fram að hér er ég ekki að setja fram einhverja grundvallartvíhyggju manns og vélar: staðreyndin er sú að menn eru oft mjög vélrænir (hafið þið hringt í „customer service“ í Bretlandi?) og vélarnar gætu í framtíðinni orðið líkamlegar á þann hátt sem ég minnist á að ofan. Athugum líka að sum tónlist sem við heyrum í útvarpinu og er búin til af manneskjum virðist vera á sviði eftirlíkinga á annarri tónlist og tilfinninga eða í raun mjög vélræn. Það má líka benda á að list er ekkert endilega að fjalla um tilfinningar, tónlist Schoenbergs eða Autechre eru hér góð dæmi, eða myndlist Mondrian og Bridget Riley, þó að hún veki oft upp tilfinningar í okkur sem hlustum og horfum.

Gervigreind er auðvitað spennandi verkfæri, hún getur flýtt fyrir, einfaldað og stungið upp á möguleikum sem okkur óraði ekki fyrir. Hér virkar hún sem

skissutækni og samræðufélagi. Við tölum við hana eins og völvu, eða spákonu, hún kemur fram með eitthvað sem við þurfum að túlka, meta og setja í samhengi. Ef við skoðum spunagreind stóru líkananna, þá höfum við fremur litla stjórn á því sem kemur út í texta, mynd og hljóði. Það er þó auðvelt fyrir okkur að breyta textanum og gera hann að okkar, en ef verið er að vinna með mynd eða hljóð, þá er gervigreindin enn sem komið er ekki að gera okkur kleift að breyta neinu að ráði. Hljóðið kemur út í einni hljóðskrá (engar rásir) og myndin í einu skjali (engin lög, eins og í Photoshop). Það er því erfitt að breyta og frekari kveikjusamræður (e. *prompting*) ná oft ekki þeim árangri sem notandinn vill. Þetta mun þó fljótt breytast. En í þeim litlu líkönum sem við vinnum með í Intelligent Instruments verkefninu, þá er gervigreindin frekar víkkun á okkar stíl, hugmyndum og aðferðum. Hljóðfærið verður eins konar víkkun (e. *prosthesis*), sem hjálpar okkur að hugsa og framkvæma, en út frá okkar forsendum. Við erum ekki að gefa hljóðfærunum kveikjur með texta, heldur að spila á þau í rauntíma.

En ljóst er að hér er um að ræða eitthvað annað og meira en tól eða verkfæri. Gervigreindin er nú orðin partur af hugsunarferlum okkar og er því orðin þekkingarfræðileg. Ég hef áður skrifað um hefðbundin og stafræn hljóðfæri sem þekkingarleg verkfæri (e. *epistemic tools*), þ.e. að við hugsum með þeim, þau skilgreina það sem hægt er að gera og eru einskonar framlenging á líkamanum. Það sem er að gerast núna er að við erum að fá viðmælanda, hálfgerðan andskota í tækninni, sem verður partur af vistkerfi okkar eigin hugsana. Þetta býður upp á marga möguleika, spennandi ný hugsunarferli, en einnig hættur.

Í grein þinni sem birtist á Vísi á síðasta ári skrifaðir þú: „Það er ljóst að gervigreind mun ekki stunda neinar frumlegar rannsóknir í sagnfræði, tungumálum, mannfræði eða heimspeki." Ljóst er að þú átt við gervigreind eins og hún er í dag, en telur þú mögulegt að þetta eigi eftir að breytast? Þ.e.a.s telur þú það mögulegt að gervigreind muni geta stundað sjálfstæðar rannsóknir í hugvísindum og komist að nýjum áhugaverðum niðurstöðum á sama hátt og manneskja?[1]

Þetta eru góðar spurningar en það er erfitt að svara án þess að skrifa heila bók. Við þyrftum fyrst að skoða hvað við meinum með „frumlegar rannsóknir" og svo á „sama hátt og manneskja". En jú, eins og staðan er í dag, þá er gervigreindin greinandi á mynstri og hún framleiðir mynstur. Hún skilur ekki neitt og veit ekki neitt. Hún hefur enga ætlun og á erfitt með að tengja úr einu sviði yfir á annað, nokkuð sem er einkennandi fyrir hinn skapandi huga. Þetta tengist skilgreiningu á skilningi en hér geng ég út frá muninum á því hvernig við kennum barni að reikna og svo þeirri staðreynd að mállíkönin reikna ekki á sama hátt. Þau reikna í raun ekki neitt, heldur giska á svarið út frá líkindum, nema ef þau ná að skilja spurninguna og senda svo reiknisdæmið yfir á aðra þjónustu (t.d. á forritunarmálið Python) og eru þá farin að nota aðra aðferð en gervitauganetin. Barn lærir tungumál af reynslu en mállíkan af líkindum þess að eitt orð fylgi öðru. Það má vel vera að þetta muni breytast í framtíðinni, en ekki í formi þeirrar

1 Greinina má finna á Vísi.is á þessari slóð: https://www.visir.is/g/20232400636d/ad-greina-gervigreind

tækni djúptauganeta sem við höfum í dag. Nú er mikið talað um almenna gervigreind (e. AGI – *Artificial General Intelligence*) og þá er möguleiki á að mállíkön tengist merkjanlegri greind og alls kyns snjallkerfum af ólíkum gerðum. Hún gæti verið líkamleg og með skynjara, þannig að ef við spyrjum hana hvort það sé rigning úti þá rétti hún út hendina fremur en að sækja gögnin frá Veðurstofunni. Hún gæti skilið bragðið af epli með því að hafa bragðskynjara (eða raftungu) og þannig grundvallað textalegan skilning sinn á líkamlegri upplifun. Þannig að já, það er möguleiki á að gervigreindin muni jafnvel skilja heiminn á einhvern hátt í framtíðinni en núna eru stórfyrirtækin í ógnarham að smíða stærri örflöguflota, sækja fleiri gögn – og jafnvel búa þau til, því að það er ekki nægilegt magn af textum – og svo þjálfa tauganetin oftar og betur. Þessi áhersla á að skala upp stærð netanna og magn gagnanna mun leiða af sér bættari tækni en hér er ekki um að ræða greind með skilning á heiminum.

Þessi tækni getur nú þegar gert ótrúlegustu hluti sem verkfæri í vísindum, fræðum, leikjum og listum. Fyrir nokkrum árum varð stökkbreyting í rannsóknum á próteinum með tilkomu AlphaFold2 frá Deepmind, þar sem gervigreindin gat komið með lausnir í prótínbrotningu (e. *protein folding*) á skömmum tíma, nokkuð sem hefði tekið manneskju áratugi að finna út úr. AlphaGo vann heimsmeistara í Go með leik sem þótti svo frumlegur að alþjóðasamfélag Go spilara var í áfalli. Stafræn hugvísindi nota mállíkön á margvíslegan hátt í rannsóknum og vélalestri. Ímyndum okkur til dæmis að brot úr forngrísku handriti finnist í leirkrukku í einhverju klaustri. Venjulega yrði það sent til helstu sérfræðinga sem myndu reyna að túlka það og setja í samhengi. Með stóru líkani sem hefði verið þjálfað á öllum grískum textum, öllum þýðingum á þeim sem og túlkunum og útleggingum, þá er möguleiki á að við gætum fengið ansi víða og góða útskýringu á hvað stendur í textabrotinu, hvaðan það kemur, hvenær það var skrifað o.s.frv. Þá væri líkanið að byggja á þekkingu allra vísindamanna sem hafa skrifað um grísk fornrit, en ekki á þekkingu einhvers eins hóps af túlkendum. Ljóst er þó að við getum aldrei vitað hverju hægt sé að treysta og því þyrftu alltaf að vera sérfræðingar sem meta tilgátur og niðurstöður gervigreindarinnar. Hún er því hjálpartæki en ekki sjálfstætt afl í vísindum og fræðum.

Til að gervigreind yrði fær um þetta, þá virðist augljóst að hún þyrfti að hafa mun betra grip á sannleikanum en hún gerir í dag, þ.e. að búa yfir skilningi á því hvað sannleikur er á sama eða svipaðan hátt og við gerum. Telur þú að gervigreind af því tagi sem við erum að þróa í dag geti nokkurn tíma orðið fær um það? Ef ekki, hvað telur þú hana vanta til þess? Þetta er í raun mjög stór heimspekileg spurning nokkuð dulbúin sem spurning um gervigreind, þ.e. hvernig öðlumst við þekkingu? En í dag gæti verið að við séum í fyrsta sinn að fylgjast með þróunarferli fyrirbrigðis að öðlast þessa getu. Því þykir okkur sérstaklega áhugavert að heyra hvernig þú sérð þetta ferli mögulega fyrir þér, út frá þekkingu þinni bæði á fræðilegu heimspekinni og tæknilegum veruleika gervigreindar í dag.

Ef við skiljum hvernig gervitaugnetin virka þá er nánast mótsögn að tala um sannleika og spunagreind í sömu hendingu. Þar sem líkönin vita ekki hvaðan

þau hafa upplýsingarnar þá er ekki hægt að sannreyna þekkingu þeirra. Þau geta reitt fram svör við spurningum en ekki stundað gagnrýna hugsun til að meta það sem þau eru að segja. Að meta uppruna upplýsinga á gagnrýninn hátt er þjálfun sem krakkar hljóta í skóla. Það er enn sem komið er ekki hægt að rekja svörin almennilega. Þó skal taka fram að það er til rannsóknarsvið sem heitir Explainable AI og reynir að vinna að lausnum á þessu vandamáli. Mállíkönin eru þjálfuð á textum á netinu og ekki er allt rétt sem þau lesa þar. Á netinu eru alls kyns textar og geta því líkönin magnað upp sjónarmið eða viðmið sem okkur þykir ekki eiga við í dag. Nú er til dæmis verið að banna bækur í Bandaríkjunum og þá vakna spurningar um það hvernig mállíkönunum verði ritstýrt út frá pólitískum sjónarmiðum. En ljóst er að sjónarmið minnihlutahópa fá ekki rödd á sama hátt og ríkjandi menning þó að oft sé svar líkansins á þann hátt að bent er á hinar ýmsu hliðar máls. Þannig eru líkönin gott dæmi um fullkomna afstæðishyggju. Spurning okkar ætti því að vera hvernig við treystum þekkingu þess sem veit ekki neitt?

Sannleikur er auðvitað hugtak sem heimspekin hefur verið að vinna með í árþúsundir án endanlegrar niðurstöðu. Það væri of langt mál hér að fara út í ólíkar skilgreiningar á sannleika, en segjum að við höfum tvær staðhæfingar: „Það er kaffibolli á borðinu“ og „Trump sýnir af sér fasískar tilhneigingar“. Vélmenni gæti svarað með nokkurri vissu hvort að fyrri staðhæfingin sé rétt, til dæmis með því að ganga að borðinu og sannreyna hana með því að taka upp bollann (ef hann er þar). Síðari staðhæfingin virðist velkjast um fyrir milljónum Bandaríkjamanna og þó að vélmenni tengt mállíkani myndi að öllum líkindum staðfesta að hún væri rétt, þá er hugmyndin um sannleika ekki jafn einföld hér og í fyrri setningunni. En það er erfitt að tala um þetta í svona stuttu máli. Hvað ef þetta væri t.d. prjónaður kaffibolli? Hann liti alveg eins út og venjulegur kaffibolli, en hann myndi ekki halda vökva, og svo framvegis.

Að mínu mati er eiginlega meira spennandi að hugsa um skilning fremur en sannleika í þessu samhengi. Hvernig skilur gervigreindin? Þau sem hafa lært eitthvað flókið, t.d. stærðfræði eða forritun þekkja hvernig maður fylgir oft reglum eins og páfagaukur þar til að maður skilur allt í einu vandamálið. Þetta er öðruvísi skilningur en sá sem góður fótboltamaður hefur úti á vellinum en það er skilningur líka. Wittgenstein er spennandi í þessu samhengi en hann talar um að skilningur sé fólginn í að fylgja reglu. En þá vandast málið, því að þetta er akkúrat það sem tölvur eru góðar í; að fylgja reglum. Wittgenstein var auðvitað að skrifa fyrir tíma tölvunnar, þó að þeir Turing hafi verið samtíma að vappa um kampusinn í Cambridge, en það er mikilvægt að hafa í huga að fyrir Wittgenstein var skilningur – og reglufylgni – félagslegt fyrirbæri. Nútíma hugfræði hafa svo í lengri tíma bent á líkamlega tengingu merkingar, til dæmis með breiðum hópi fræðimanna sem eru oft kenndir við hin fjögur E (e. *embedded*, *embodied*, *enactive*, *extended*). Eins tala sumir um að það þurfi vitund til að skilja á sama hátt og við gerum, og að vitund geti aðeins verið til í lífrænum kerfum. Því held ég að gervigreindin sé ekki að fara skilja heiminn eins og við í nánustu framtíð.

Heldurðu að fólk eigi eftir að líkjast meira gervigreind eða gervigreindin meira fólki í framtíðinni, eftir 10–15 ár? Eigum við líka bara eftir að bulla meira og reyna að finna

einhver gögn til að styðja við mál okkar, eða heldurðu að gervigreindin muni frekar verða betri í að finna heimildir fyrir sínar staðhæfingar?

Já, alveg örugglega hið síðara. Það er verið að vinna í að finna leiðir til að hægt sé að leita að uppruna þekkingar, sannreyna þekkingu og svo að blanda saman symbólískri gervigreind og þessum máll íkönum. Það er nauðsynlegt að gervigreind fari að geta staðið við það sem hún er að segja og sum gervigreind gerir það á vissan hátt eins og Bing til dæmis, þar sem maður fær oft vísanir í heimildir. Hið nýja kínverska DeepSeek máll íkan sýnir okkur svo hvaða hugsunarferli liggur að baki svarinu og það getur verið mjög hjálplegt til skilnings. Svo spyrjið þið líka hvort að við munum fara að líkjast gervigreindinni og ég held að það sé að gerast nú þegar. Fólk fer að aðlaga sig tungumáli gervigreindarinnar og hugsa eins og hún. Það gerist ósjálfrátt þegar nemendur og fólk í atvinnulífinu fer að láta gervigreindina skrifa fyrir sig. Það eru komin stefnumótaforrit, þar sem gervigreind skrifar skilaboð á milli fólks. Mörgum finnst þetta fremur skelfilegt og hér er mikilvægt að menntakerfið komi sterkt inn: þegar við erum komin með þessa tækni til að hugsa, framlengingu á heilanum, þá þurfum við að bregðast við mjög sterklega og endurskoða hvernig við metum okkur sjálf og hvernig við kennum, vinnum, og þjálfum gagnrýna hugsun. Í besta mögulega heimi er gervigreindin bara frábært tæki til að hjálpa fólki að hugsa rétt og fá gífurlega mikið af gögnum og draga saman efni. Hún getur hjálpað að draga saman þekkingu, lesa texta og túlka þá. En við erum sjaldnast í besta mögulega heimi og nýleg tengsl stórfyrirtækjanna við öfgapólitík vegur upp ugg. Spurningin er því ávallt um hlutverk gagnrýninnar hugsunar og hvernig við sannreynum það efni sem er framleitt.

Í greininni á Vísi spyrðu einnig hvort við þurfum að aðlaga okkur að stöðlum gervigreindar eða hvort við getum þróað okkar eigin menningu í henni. Gætirðu sagt okkur meira frá því hvað þú ert að hugsa um þegar þú spyrð þessa spurningu?

Þessi stóru máll íkön eru byggð á því sem finnst á netinu, á stafrænum bókum og öðrum gögnum sem þar er að finna. Þau eru þjálfuð af örfáum stórfyrirtækjum sem hafa ólíkan aðgang að gögnum og það er ekki alltaf á hreinu hvaða gagnasöfn eru notuð. En það sem ég var að benda á í greininni er að það er ekki gerður greinarmunur á hvaða textar eru notaðir, það úir og grúir af allskyns vitleysu á netinu, en svo hefur máll íkanið verið þjálfað í því að vera „kurteist“. En hvaða kurteisi er það? Hér er um að ræða kalifornískt siðferði, þar sem myndir af morðvopnum þykja eðlilegri en af móður að gefa barni brjóst. Líkönin hafa verið þjálfuð gífurlega vel af heilum her af fólki – mest láglaunafólki í Afríku – sem hefur fengið vissan siðastaðal frá hönnuðum tækninnar. Með þeim samruna tækni og pólitíkur sem við sjáum í dag er þó líklegt að upplýsingum verði stýrt meira, t.d. veit DeepSeek ekkert um það sem gerðist á Torgi hins himneska friðar og líklega mun ChatGPT fljótlega hætta að tengja Trump við fasisma. Það er þó skemmtilegt að oft er hægt að gabba tauganetin til að fara framhjá þeirri þjálfun.

Og þá kemur spurningin um „eigin menningu“. Hvernig munu ríki eins og Kína eða Tyrkland reyna að breyta virkni líkananna, því að margt sem þau segja

stangast á við skilgreiningar þeirra á æskilegri menningu? Hvernig munu alræðisríki dragast aftur úr í tækni við það að loka á þessi líkön, eins og t.d. Íranir hafa gert? Hvernig fer með menningu minnihlutahópa eins og frumbyggja Ameríku eða Skandinavíu? Hvernig verða tungumál varðveitt með gervigreind? Hér er um að ræða nýja heimsvaldastefnu og alþjóðavæðingu þar sem ein menning mun verða yfirsterkari en aðrar ef ekki verður brugðist við. Slík „mjúk" heimsvaldastefna hefur auðvitað verið í gangi í áratugi með kvikmyndum frá Hollywood og tölvuleikjum eins og menningarfræðingar hafa rætt í áratugi, en hérna er þetta óljósara og flóknara. Það er því ljóst að við töpum menningarlegu sjálfræði tímabundið við það að nota þessi mállíkön, líkt og við höfum tapað afli blaðamennskunnar þar sem að erlend stórfyrirtæki hirða allar auglýsingatekjur. Spurningin er hvernig við fáum þessa hluti til baka, hvernig við getum fundið menningu okkar farveg í þessum líkönum? Mér finnst við Íslendingar hafa staðið okkur ágætlega hér, bæði stjórnmálamenn og fyrirtæki í einkageiranum.

Þegar kemur að öllum viðfangsefnunum sem hugvísindi geta tekist á við þegar kemur að gervigreind, þá finnst mér eitt þeirra standa upp úr – og þú minnist einmitt á það í greininni – og það er dauðinn. Hér er ekki átt við deepfake tækni, heldur þjálfun gervigreindar á gömlum skilaboðum og öðrum gögnum þannig að hún geti hermt eftir okkar nánustu svo hægt sé að „ræða við þau" eftir að þau eru dáin. Fyrir lesendur sem ekki þekkja til þá er hér ekki um að ræða möguleika í fjarlægri framtíð, heldur er þessi gervigreindarhæfni nú þegar aðgengileg og notuð af ýmsu fólki.

Telur þú að þessi notkun gervigreindar gæti á einhvern hátt verið af hinu góða? Berum við ábyrgð á því að fá leyfi manneskju, áður en hún fellur frá, til þess að þjálfa gervigreind á gögnum sem við höfum eftir hana? Finnst þér að þetta ætti mögulega að vera eitthvað sem varðar við lög, núna eða í framtíðinni? T.d. hvort einstaklingar ættu að eiga eitthvað eins og höfundarrétt á eigin persónuleika?

Þetta eru góðar spurningar sem ég á erfitt með að taka afstöðu til. Persónuverndarsjónarmið koma sterkt inn en einnig þarf að gera rannsóknir í sálfræði, félagsfræði og menningarfræði varðandi það hvernig hugmyndir um dauðann og látið fólk breytast við svona lagað. Það er ljóst að þetta er að gerast, eins og þið bendið á, en hverjar eru afleiðingarnar? Við höfum auðvitað unnið ævisögur og kvikmyndir úr persónulegum gögnum fólks – bréfaskiptum, dagbókarskrifum, myndum og annað – og þá vakna siðferðisspurningar varðandi gögn og notkun þeirra. Við biðjum ekki endilega um leyfi til að skrifa ævisögu látinnar manneskju. En erum við ekki að tala um eitthvað annað hér? Er það ekki eðlisólíkt þegar við gerum kvikmynd um einhvern eða tökum persónu þeirra og endurgerum hana í lifanda líki? Þarf hinn látni að gefa leyfi til að þetta geti verið gert? Og hvaða reglur getur viðkomandi sett um það hvernig persónan er notuð? Hér er ærið verkefni fyrir siðfræðinga og í raun gott dæmi um það hvernig gervigreindin hefur skapað heimspekingum ótal umhugsunarefni – og launuð störf!

En þið talið um persónulíkingar (e. *impersonation*) og það er ekki endilega eitthvað sem hefur eingöngu með látna að gera, heldur er hægt að stela höfundareinkennum, rödd, stíl í listum og slíku. Hvar liggja mörkin hérna? Við erum

hreinlega ekki komin nógu langt í þessari umræðu og hér hafa siðfræðingar og lögfræðingar æríð verk að vinna. En það er hér sem mér finnst þessi spunagreind sem stórfyrirtækin eru að þróa í dag vera hættulegust og fremur ógeðfelld.

Hvort að þetta geti verið gott eða vont fyrir okkur veit ég ekki. Menningin breytist með tækninni og við sjáum núna hvernig ungt fólk hegðar sér ólíkt því þegar við vorum að alast upp. Þau eru upplýstari, með meiri aðgang að gögnum, en á sama tíma eru þau að kljást við kvíða, einmanaleika og þunglyndi sem eru beinar afleiðingar af þessum algrímum stórfyrirtækjanna. Upplýsingar og samskipti í gegnum snjalltæki eru ekki slæm í sjálfu sér, en smáforrit sem gera fólk að fíklum eru það (Norðmenn eru að spá í að banna slík forrit fyrir ungt fólk). Því get ég ekki tekið afstöðu núna til þess hvort að samskipti við persónulíki séu af jákvæðum eða neikvæðum toga, það á eftir að koma í ljós, og auðvitað er það persónubundið þar að auki.

Hverjir eru helstu möguleikarnir í þróun gervigreindar sem þú hefur áhyggjur af – og telur að krefjist meiri athygli?

Hættan við gervigreind er meira pólitísk, efnahagsleg og siðfræðileg fremur en að hún sé hættuleg í sjálfri sér. Það er notkun hennar sem er vandamál samtímans. Við sjáum örfá stórfyrirtæki sem eru orðin alltof stór og öflug, með hagkerfi á við meðalstórt land, taka völd yfir miðlum og stjórnmálum út um allan heim. Þau eigna sér heilu markaðina og má nefna hér Spotify í tónlist, Storytel í bókmenntum og Amazon í vörusendingum. Eigendur þessara fyrirtækja geta haft afgerandi áhrif á úrslit kosninga eins og nýleg dæmi sanna. Paul Virilio sagði að uppfinning skipsins sé einnig uppfinning skipbrotsins og þetta á við hér, nema að brot og hættur algrímanna eru ekki bundin við eitt svið eins og sjómennsku.

Persónulíkingar eru dæmi um slíka tækni. Hún getur virkað vel í kvikmyndaiðnaðinum, Skaupinu, og alls kyns gríni í smáforritum, eins og krakkarnir okkar eru sífellt að leika sér að, en hún er einnig verkfæri glæpamanna, t.d. í sviðsettum mannránum þar sem svindlarar hringja í ættingja með rödd hins „rænda“ og krefjast lausnargjalds. Það er hægt að búa til gerviatburði sem áttu sér aldrei stað. Sumir stjórnmálamenn munu eflaust lenda í vandræðum út af þessu en aðrir geta hafnað sannleikanum og sagt að um sé að ræða falsfréttir. Sannleikurinn fer að mást út. Hér má einnig nefna sjálfvirkar fréttir þar sem gervigreind skrifar fréttir sem fylgja ekki stöðlum fréttamennsku.

Ef við lítum á þetta eins og Virilio út frá hugmyndum hans um innbyggt slys hverrar tækni (fr. *l'accident intégral*), þá er ljóst að hætturnar liggja víða. Verðbréfa- og rafmyntamarkaðir eru undirlagðir gervigreind og það þarf lítið út af að bera til að þar fari allt til fjandans. Ótal aðrir hlutir eru í þróun sem erfitt er að sætta sig við, t.d. sjálfvirk vopn, þ.e. þegar vélmennum og drónum er stýrt af gervigreind og engin manneskja kemur þar að málum en raunverulegt fólk er að láta lífið. Firringin getur varla verið meiri.

Einnig eru í þróun sambland líffræðilegra tauganeta og gervitauganeta, þ.e. að tengja heila okkar við gervigreind og netið sjálft. Þeir sem eru að þróa þá tækni trúa því að við munum geta endurspilað minningar eða skipst á þeim, lært tungu-

mál í hvelli, hlustað á tónlist án hljóðs og tengst öðrum í gegnum hugsanir. Hægt verður að selja þær upplifanir, líkt og í myndinni *Strange Days*. Það eru milljarðar settir í að þróa þetta þessa dagana.

Þá er einnig verið að skoða möguleikann á gervilífi (e. *artificial life*) í þróun nýrra lausna á vandamálum sem gervitauganetin geta ekki leyst. Tauganetin byggja á fortíðinni og því sem þau eru fóðruð á, en gervilíf með erfðafræðilegum algrímum, flokkum (e. *swarms*) og öðrum aðferðum geta þróað nýja möguleika sem núverandi líkön eru ekki fær um. Sambland gervilífs við spunagreind gæti orðið mjög gjöful aðferð við að þróa nýja þekkingu sem er ekki eingöngu byggð á fortíðinni.

Svo má vara sérstaklega við þeirri viðleitni sumra til að búa til gervigreind með vitund. Það hafa verið skrifaðar lærðar greinar um þann möguleika og er fræðafólk innan hugfræða með ólíkar skoðanir á því hvort að þetta verði yfirleitt hægt. Sumir segja að vitund verði aðeins til í líffræðilegum kerfum en aðrir sjá möguleikann á að gervitauganet – sem jafnvel hefur líkama sem sér, hlustar og finnur til líkt og mannslíkaminn – geti líka öðlast vitund. En hvert erum við þá komin? Hvaða siðferðislegu spurningar vakna þegar tæknin okkar er komin með vitund? Vísindaskáldskapur undanfarinna áratuga hefur að sjálfsögðu unnið með þessar spurningar gífurlega vel og af honum er margt hægt að læra.

Það er endalaust hægt að telja upp svona sviðsmyndir, en sú nærtækasta er ef til vill sú að þegar þau gögn sem eru á internetinu (textar, myndir, hljóð) eru búin til af gervigreind, þá fer vitleysan stigmagnandi, því að ný líkön verða þjálfuð á vitleysunni í eldri líkönum. Ég hef t.d. séð myndir af Reykjavík notaðar á túristasíðum, sem eru ekki af Reykjavík. Þær líkjast bara Reykjavík. Hvernig munum við getað aðskilið hið sanna frá hinu ósanna? Munum við jafnvel hætta að hugsa á þeim nótum? Að við sættum okkur við líkindi raunveruleikans í miðlum og samskiptum?

Enn annar vandi við gervigreindina er stéttaskipting. Hver hefur aðgang að henni og hver ekki? Til dæmis geta efnaðri nemendur keypt áskrift af bestu líkönunum, og hreinlega framreitt betri ritgerðir en verr stæðir nemendur. Viðbragðið við þessu ætti þó að vera að breyta námsmati því að námsritgerðir eru í auknum mæli skrifaðar af gervigreind. Öll þessi tækni mun auka stéttskiptingu, til dæmis þegar Neuralink heilaviðmótið (e. *brain-computer interface*) kemur á markað.

Að lokum vil ég bara minnast á að ég held að þegar við tölum um gervigreind að þá erum við ekki að vísa í eitthvað eitt. Þetta eru þúsundir algríma sem virka á ólíkan hátt. Það er ekkert hættulegt við þau í sjálfu sér, en hættan liggur í því hvernig túrbókapítalisminn, pólitíkin, glæpahringir og aðrir nota þessi verkfæri. Á margan hátt má líkja þessu við þegar eðlisfræði síðustu aldar var að uppgötva kjarnaorkuna. Þetta er því vandamál sem er á sviði siðfræði, lögfræði og að endingu stjórnmála, en þar vandast þó málið, því að stjórnmálin eru algjörlega undir þessum stórfyrirtækjum komin, eins og sjá má af stefnuskrám flokka út um allan heim. Málið er æsispennandi og ekki fyrir hjartveika, en ljóst er að fyrir hug- og félagsvísindin eru ærin verkefni framundan.