Gervigreind, sýklar, atómsprengjur og allt þetta fína: Hugleiðing um bók eftir Mustafa Suleyman

Atli Harðarson skrifar 30. október 2024 11:46

Mustafa Suleyman (f. 1984) er frá London en býr nú í Kaliforníu. Hann er með áhrifamestu frumkvöðlum heims á sviði gervigreindar og einn af stofnendum fyrirtækisins DeepMind sem Google eignaðist árið 2014. Eftir það gegndi hann stjórnunarstöðum hjá Google til 2022 og er nú háttsettur innan Microsoft. Bók hans *The Coming Wave: AI, Power and Our Future* kom út haustið 2023 og vakti strax mikla athygli.

Bókin er 332 blaðsíður að lengd, fremur fljótlesin og auðskiljanleg leikmönnum. Hún er ákall um að heimsbyggðin vakni og geri sér grein fyrir þeim miklu hættum sem fylgja þróun gervigreindar. Með ritun hennar tekur Suleyman undir varnaðarorð fjölda vísindamanna í fremstu röð allt frá því Joseph Weizenbaum (1976) sendi frá sér bókina *Computer Power and Human Reason* fyrir nær hálfri öld til dagsins í dag. Það má til dæmis nefna að á síðasta ári undirrituðu hundruð sérfræðinga á sviði gervigreindarfræða yfirlýsingu þar sem segir að ráðstafanir til að draga úr hættu á gereyðingu af völdum þessarar tækni ættu að vera forgangsverkefni á alþjóðavettvangi rétt eins og varnir gegn farsóttum, kjarnorkustyrjöldum og annarri vá sem ógnar tilvist heilla samfélaga.[i]

Mikill máttur til illra verka

Gervigreind þróast hratt um þessar mundir. Það er aðeins um eitt og hálft ár síðan almenningur fékk aðgang að spjallvél OpenAI sem flestir þekkja undir nafninu „GPT 4“ og nú er hún notuð af milljónum manna við alls konar verkefni. (Bókstafirnir þrír standa fyrir „Generative Pre-trained

Transformer“ sem spjallvélin sjálf sagði mér að þýða með orðunum „skapandi fyrirfram þjálfaður umbreytir.“)

GPT tæknin er ung, hefur þróast á um áratug. Hún gerir vélum kleift að „læra“ mál með greiningu á textum sem liggja frammi á netinu. Þessi lærdómur er ólíkur máltöku barna. Hann krefst gríðarlegrar orku (sem líklega telur í gígawattstundum)[ii] þar sem þúsundir tölva suða og puða. Útkoman er samt fullt vald á mannamáli og meira til því spjallvélar af þessu tagi svara alls konar spurningum með lýtalausum texta og skálda svör fremur en að játa vankunnáttu. Engin leið er að sjá svör þeirra fyrir né rekja hvernig þau eru fengin. Það er ekki í mannlegu valdi að greiða úr kóðanum sem verður til þegar þær „læra“ – útkoman er véfrétt sem svarar stundum af furðu miklu viti en romsar stundum upp úr sér bulli og vitleysu.

Gagnsemi svona spjallvéla er mikil. Þær vinna hratt úr meira og flóknara gagnamagni en mannshugur ræður við, skrifa tölvuforrit og fleira og fleira. Í stuttu máli færir þessi tækni mönnum mikinn mátt sem nýtist ekki síður til illra verka en góðra. Viðvaranir Suleymans og fleiri vísindamanna eru vegna þess að sum illvirkin sem hægt er að fremja með hennar hjálp eru afar óhugnanleg. Við bætist að það er mikil óvissa um hversu hratt tæknin þróast. Menn vita ekki á þessari stundu hvort vélar muni af sjálfsdáðum geta smíðað æ fullkomnari útgáfur af sjálfum sér og þannig aukið getu sína án atbeina manna. Ekki er heldur vitað hvort úr því verði vélar sem vinna að eigin markmiðum hvað sem mannlegum hagsmunum líður og einnig er óvissa um hvaða tök menn hafa á að slökkva á slíkum tækjum. Um dýpri heimspekilegar ráðgátur tengdar gervigreind er líka óvissa. Það eru engin svör við því á hvaða sviðum vitsmuna og hugarstarfs vélarnar fara fram úr mönnum og ekki heldur við því að hve miklu leyti þær taka að líkjast lifandi verum.

Ethan Mollick sem kennir frumkvöðlafræði við Pennsylvaníuháskóla ræðir hætturnar á að menn missi allt vald á vexti þessarar tækni í lokakafla bókar sem kom út á þessu ári og fjallar um notagildi gervigreindar fremur en hætturnar sem fylgja henni. Hann segir að við vitum einfaldlega ekki hvort til verði vélar sem taka mönnum fram að vitsmunum og ekki heldur hvort þær muni fremur leggja okkur lið eða snúast gegn okkur og bætir við að „þessa hættu verði að taka alvarlega“ (Mollick, 2024, bls. 208).

Líftækni og hernaður

Ótti vísindamanna við vitsmunaskrímsli sem þeir hafa sjálfir skapað er að nokkru leyti ótti við hið óþekkta og óskiljanlega. Hætturnar sem Suleyman ræðir eru samt fyrst og fremst hættur sem þegar eru orðnar til eða eru að verða til. Verulegur hluti bókar hans fjallar um samspil gervigreindar og líftækni. Tæki til að „skrifa“ erfðaefni fást nú þegar á verði sem margir ráða við. Margir hafa líka aðgang að gervigreind sem gerir þeim til dæmis mögulegt að hanna veirur með tiltekna eiginleika.

Sem betur fer eru veirur sem eru bæði mjög smitandi og mjög banvænar sjaldgæfar í náttúrunni en ekkert útilokar samt að þær verði búnar til. Við erum aðeins hársbreidd frá þeim möguleika að bílskúrsfyrirtæki geti búið til farsóttir sem granda stórum hluta mannkyns.

Ragnarök af mannavöldum hafa verið möguleg og vofað yfir okkur frá því kjarnorkuvopnum tók að fjölga í kalda stríðinu. Sem betur fer hafa fáir slík vopn undir höndum en það sama gildir ekki um gervigreind og líftækni.

Önnur ógn sem Suleyman ræðir er smíði alsjálfvirkra og nær alsjáandi vígvéla. Ríki eða hryðjuverkasamtök sem geta herjað með vélmennum eru líklegri til að stjórna með skefjalausu ofbeldi en yfirvöld sem aðeins hafa dauðlegum og mennskum hermönnum á að skipa. Með þessu telja margir að hernaður komist á nýtt stig með óheyrilegum afleiðingum.

Suleyman ræðir í þessu sambandi um hvernig sjálfvirk eftirlitskerfi geti „metið“ þörf fyrir hernaðarlega íhlutun og sett drápstæki af stað án þess neinn maður viti hvers vegna. Hann ræðir líka hratt lækkandi verð á tæknibúnaði sem gerir sjálfvirkar mannaveiðar og dráp möguleg jafnvel fyrir fámenn hryðjuverkasamtök eða brjálaða einstaklinga. Við þetta má bæta að sá möguleiki er þegar í sjónmáli að mennsk yfirvöld setji bara markmið með hernaði og láti vélar um að ná þeim.[iii] Fyrir fólk sem dreymir um frið og farsæld er það ekki tilhlökkunarefni ef gervigreind sem enginn maður hefur roð við fær fyrirmæli um að vinna þriðju heimstyrjöldina hvað sem það kostar.

Eftirlit og einstaklingsfrelsi

Ein afleiðing af þeim hættum sem fylgja sjálfvirkum hernaði og samþættingu gervigreindar og líftækni er stóraukin þörf fyrir eftirlit og njósnir. Suleyman minnir á að kóvid faraldurinn fékk almenning víða um heim til að sætta sig við skerðingu á ferðafrelsi og fleiri mannréttindum og ályktar að vitneskja um getu glæpamanna til að búa til farsóttir fái fólk til að kalla eftir lögregluríki þar sem fylgst er nákvæmlega með okkur öllum.

Gervigreind með vélum sem þekkja andlit og myndavélar við hvert horn gerir yfirvöldum kleift að fylgjast betur með fólki en Stóri bróðir gerði í bókinni *Nítján hundruð áttatíu og fjögur* eftir George Orwell (1949). Stóraukið rafrænt eftirlit með fólki er þegar orðið að veruleika eins og Cathy O‘Neil (2016) gerir grein fyrir í bókinni *Weapons of Math Destruction* sem markaði þáttaskil í umræðu um áhrif upplýsingatækni á líf okkar. Hún vann á tímabili við að forrita fjáröflun vogunarsjóða og síðan við að reikna út hvernig best borgaði sig að bauna auglýsingum á fólk gegnum samskiptamiðla. Vegna þessara starfa fór hún að hugsa um hvernig sjálfvirk söfnun og úrvinnsla gagna er notuð í sívaxandi mæli til að fylgjast með fólki, hafa áhrif á það og úthluta lífsgæðum. Sá bandaríski „Stóri bróðir“ sem hún lýsir er ekki mennskur einstaklingur heldur fjöldi

meira og minna samtengdra upplýsingarkerfa sem enginn hefur yfirsýn yfir og fáir vita hvernig virka. Sum þeirra ráða því hvort einstaklingur fær lán og þá á hvaða kjörum, önnur hvað tryggingar hans kosta, enn önnur hvort hann fær vinnu, skólavist eða reynslulausn úr fangelsi. Almennt eru þessi kerfi hvorki réttlát né áreiðanleg. Eins og tölvufræðingurinn og listamaðurinn Joy Buolamwini (2023) gerir grein fyrir í nýlegri bók eru þau meira og minna lituð af þeim fordómum sem finna má í efninu sem þau eru mötuð á og menningunni sem mótaði það.

Suleyman segir að með samtengingu eftirlits og upplýsingarkerfa sem þegar eru til geti ríkisvald nútímans gengið mun lengra í eftirliti með borgurunum en nokkurt alræði sem við þekkjum úr sögu fyrri ára og hann lætur að því liggja að stjórnvöld í Kína geri þetta nú þegar.

Ríkisvald og lýðræði

Þegar best lætur, segir Suleyman, beita þjóðríki nútímans ríkisvaldinu til að gera fólki mögulegt að lifa við frið og velmegun en takmarka jafnframt yfirráð þess með ýmsum hætti. Hann segir að í lýðræðisríkjum nútímans ógni gervigreindin viðkvæmu jafnvægi milli einstaklingsfrelsis og öryggis, persónuverndar og eftirlits, stjórnleysis og ofríkis. Hættan sé ekki aðeins sú að ríkið öðlist alræðisvald eins og í Kína heldur líka sú að það missi öll tök á samfélaginu eins og í Líbanon.

Suleyman er ættaður frá Sýrlandi í föðurætt og þess sér víða stað í bókinni að hann er mjög minnugur þess að friður og farsæld eru háð ríki sem virkar en kann sér samt hóf í valdbeitingu.

Til að bregðast við hættunum sem fylgja gervigreind með skynsamlegri lagasetningu og samstilltu átaki þurfa þjóðríkin, segir hann, traust almennings til að ná samstöðu bæði innanlands og á alþjóðavettvangi. Þetta er erfitt meðal annast vegna þess ríki nútímans einkennast af

vantrausti sem er að nokkru vegna þess hvernig gervigreind mótar samskiptamiðla og leitarvélar. Slík „tæki“ safna upplýsingum til að marka okkur og draga í dilka sem sumir kalla bergmálshella. Þar fær hver hópur sína ýktu og afskræmdu mynd af öðrum og út verður skautun og skortur á trausti.[iv] Vegna þessa eru samfélög nútímans í erfiðri stöðu og eiga bágt með að koma böndum á þann Fenrisúlf nútímans sem gervigreindin er. Uppgjöf frammi fyrir því verkefni kemur samt ekki til greina. Of mikið er í húfi.

Bönd á úlfinn

Suleyman notar orðið „containment“ yfir að ná stjórn á þróun og útbreiðslu gervigreindar. Hann gerir sér ljóst að margt af þessari tækni er þegar aðgengilegt almenningi og það er afar erfitt að stjórna notkun hennar með eintómri lagasetningu og regluverki. Hann segir að einnig þurfi hugarfarsbreytingu og bendir á jákvæðar fyrirmyndir úr fluginu þar sem flugfélög og yfirvöld standa saman um að allir eigi þess kost að læra af slysum og atvikum þar sem hætt var við slysum. Í fluginu, segir hann, hefur byggst upp í senn regluverk og alþjóðleg öryggismenning sem virkar og því séu flugferðir nútímans einhver hættuminnsti samgöngumáti sem sögur fara af.

Hann kallar eftir að ýmislegur búnaður bæði á sviði gervigreindar og líftækni verði háður ströngu eftirliti og stöðlum sem tryggja öryggi. Þetta þurfi að taka jafn alvarlega og viðleitni til að hemja útbreiðslu kjarnorkuvopna og jafnframt þurfi að byggja upp menningu, vitund og samábyrgð eins og tekist hefur í fluginu.

Það er mikil fáviska að halda að gervigreind sé eins og hvert annað verkfæri sem léttir okkur lífið. Veruleikinn er sá að hún felur í sér hættur og þær mjög af stærri gerðinni. Bók Suleymans á því erindi við alla sem þora að hugsa um brýnustu úrlausnarefni samtímans.

Höfundur er prófessor við Menntavísindasvið Háskóla Íslands.

Rit sem vísað er til

Buolamwini, J. (2023). *Unmasking AI: My Mission to Protect What Is Human in a World of Machines*. Random House.

Mollick, E. (2024). *Co-Intelligence: Living and Working with AI*. Penguin, Random House.

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown publishers.

Orwell, G. (1949). *Nineteen eighty-four*. Secker & Warburg.

Suleyman, M. (2023). *The Coming Wave: AI, Power and Our Future*. Crown publishers.

Vaidhyanathan, S. (2018). *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy*. Oxford University Press

Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman & co.

---

[i] Sjá https://www.safe.ai/work/statement-on-ai-risk

[ii] Sjá https://www.numenta.com/blog/2023/08/10/ai-is-harming-our-planet-2023

[iii] Sjá https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/May-June-2017/Pros-and-Cons-of-Autonomous-Weapons-Systems

[iv] Sjá t.d. Vaidhyanathan (2018) og umsögn mína um bók hans á https://atlivh.com/textar/ymislegt/Umsogn_um_bokina_Antisocial_Me