Nick Boström - Kas me elame arvutisimulatsioonis?

Urmas Pitsi, märts 2020

Kokkuvõte

Boströmi argumendi kohaselt on *vähemalt üks* järgnevatest väidetest tõene:

1. Inimkond sureb välja enne, kui jõuab “posthuman” arengutasemeni;
2. On äärmiselt ebatõenäoline, et ükski posthuman tsivilisatsioon simuleerib enda evolutsioonilist ajalugu (või mõnd selle variatsiooni);
3. Me elame paaegu kindlasti arvutisimulatsioonis.

Sellest tulenevalt on uskumus sellest, et on märkimisväärne võimalus posthuman-itest, kes teevad eellaste simulatsioone on väär, välja arvatud juhul, kui me juba praegu elame arvutisimulatsioonis. Lisaks veel mitmed muud järeldused, mida arutleme allpool.

Boströmi artikkel on jaotatud 7-ks paragrahviks:

I. Sissejuhatus
II. Eeldus
III. Arvutatavuse tehnoloogilised piirid
IV. Simulatsiooniargumendi tuum
V. Ükskõiksusprintsiip (võrdse tõenäosuse printsiip)
VI. Tõlgendus
VII. Järeldus

I. Sissejuhatus

Paljud ulmekirjanikud, tehnoloogid, futuroloogid ennustavad, et kunagi tulevikus saabub aeg, kui inimkonna kasutatav arvutusvõimsus on enneolematult suur. Nii suur, et selline arvutusvõimsus (arvuti) lubab meil simuleerida reaalset tegelikkust, inimese ajutegevust, mõistust, teadvust jms. Nii võib ette kujutada, et enamus mõistuslikke olendeid on hoopis simuleeritud väga kaugelearenenud eellaste poolt. Sellest tulenevalt võib argumenteerida, et me oleme pigem simuleeritud mõistuslikud olevused, mitte bioloogilised. Seega, kui me usume, et me ei ela praegu simulatsioonis, siis meil pole alust uskuda, et meil saab olema järeltulevaid põlvkondasid, kes suudaksid teha eellaste simulatsioone.

Struktuur on järgnev:

(1) Formuleerime eelduse, mille me võtame mõistuse filosoofiast (philosophy of mind);
(2) Vaatleme kogemuslikult millistel alustel me saame mõelda, et tuleviku inimkond on võimeline simuleerima inimmõistuseid. Arvestades, et see oleks kooskõlas meile teadaolevate füüsikaseadustega ja inseneeria piirangutega. See osa ei ole filosoofiliselt vajalik, küll aga pakub motivatsiooni järgnevat tähelepanelikumalt jälgima.
(3) Argumendi tuum: kasutab lihtsat õenäosusteooriat ja ükskõiksuse printsiipi. Ükskõiksuse printsiip: (võrdse tõenäosuse printsiip, principle of indifference) tõenduse puudumisel jaotame sündmuste toimumiste tõenäosused võrdselt kõigi võimaluste vahel.
(4) Lõpetuseks arutleme tulemuseks saadud 3-väitelise järelduse tõlgendusi.

II. Eeldus

Tavaline eeldus mõistusefilosoofias on *ainest-sõltumatuse* (*substrate-independence*) eeldus. Idee seisneb selles, et mentaalsed olekud on sõltumatud füüsikalisest ainest, milles need parasjagu tekivad. Sellest tulenevalt, saab inimaju tegevust modelleerida arvutiga, teostades korrektsed arvutuslikud struktuurid ja protsessid, mis on seotud teadvusega (teadlike kogemustega). Arvuti ei pea olema ilmtingimata moodustatud bioloogilistest materjalidest, vaid võib ka olla muust materjalist, peaasi, et suudab teostada samasuguseid tegevusi, nagu seda teeb inimaju.
Ainest-sõltumatuse eeldus ei ole range, st see pole *vajalik*. Piisab, et arvutusi teostav masin omaks teadvust. Veel enam, mõistuse loomiseks pole vaja isegi inimkäitumise täielikku imiteerimist. Meil on vaja täita ainult nõrgem eeldus, mis seisneb järgnevas. Piisab subjektiivsete kogemuste genereerimisest, nii et inimaju protsessid on taasesitatud piisava täpsusega, näiteks üksikute sünapsite (aju-närvirakkude vahelised ühendused) tasemel. Taoline leebem versioon ainest-sõltumatusest on üsna laialt aktsepteeritud.
Ajus on ka sünapsitest väiksemal skaalal osakesi, aga me eeldame, et suudame modelleerida subjektiivseid kogemusi sünapside tasandil.

II. Kriitika:

- Eeldame, et tegemist on arvutatava probleemiga (Universaalse Turingi Masina mõistes)
- Eeldame, et saame lahenduse realiseerida digitaal-arvutil
- Eeldame, et aju piisavalt täpseks simuleerimiseks piisab neuronite/sünapsite tasemest. Võib-olla peab arvestama ka neuronisiseste mehhanismidega?
- Eeldame, et sünapsite arv on ajas muutumatu: päriselus sünapsite arv ja intensiivsus muutub ajas.
- Eeldame, et neuron/sünapsi simuleerimine toimib 2-oleku süsteemis: aktiivne/mitte-aktiivne olek. Päriselus on rohkem olekuid: ajaliselt kumulatiivne aktiveerimispotentsiaali kogumine, aktiveerumine vastavalt sisendite intensiivsusele jne.
- Eeldame, et meil on teada *piisavalt efektiivne* algoritm, mis määrab mudeli olekute üleminekud. Üks asi on loendada algosad, hoopis midagi muud on teada milline algosade kombinatsioon järgneb etteantud algosade kombinatsioonile. Kõik olekud ja üleminekud peavad olema kooskõlas “normaalse” maailmaga ühe aju piires ning samuti ka kõikide ajude vahel, sest ajud vaatlevad samu väliseid nähtusi.

III. Arvutatavuse tehnoloogilised piirid

Käesoleval ajal ei ole meil piisava jõudlusega riistvara ega ka eeldatavat tarkvara, mille abil saaks arvutites luua teadvust. Meil on veenvaid argumente, et *kui* tehnoloogia areng jätkub samas tempos, *siis* me ületame selle takistuse. Käesolev argument ei sea ajalisi piiranguid ega eeldusi, mis tähendab, et samahästi võib “posthuman” tasemeni jõudmiseks kuluda ka sadu tuhandeid aastaid. Selleks ajaks on inimkond omandanud enamiku tehnoloogilisest võimekusest, mis on vastavuses füüsika seadustega ning materjali- ja energiapiirangutega.

Kuna meil hetkel puudub “kõige teooria” (theory of everything), siis ei saa välistada uusi füüsikalisi nähtusi, mis praeguses füüsikateoorias on võimatud. Näiteks Bremermann-Bekensteini piir ja Musta Augu limiit (? U.P.: kas see pole eelmise lausega otseses vastuolus?).
Bremermanni limiit [5]: arvutusvõimsuse maksimummäär, mida saab välja “pigistada” 1kg mateeriast meie Universumis. $\approx 1.36 \times 10^{50}$ bitti sekundis kilogrammi kohta.
Beckensteini piir [6]: kui palju informatsiooni saab maksimaalselt salvestada ühte ruumiossa $\approx 2.58 \times 10^{43}$ bitti kilogrammi ja meetri kohta (meeter on siin vaadeldava ruumi raadius).
Erinevad autorid on arvutanud välja erinevaid arvutusvõimsuse hinnanguid.
Lloyd [7]: 5×10^{50} loogilist operatsiooni sekundis.
Boström järeldab, et inimaju simuleerimiseks piisab arvutuskiirusest $\sim 10^{16} - 10^{17}$ operatsiooni sekundis.
Selline tuletuskäik: Neuronite arv * sünapsid neuroni kohta * 10 korda sekundis: $\sim 10^{10} \times 10^{5} \times 10$
Maailma simuleerimisel on analoog arvutimängudega: kuvatakse ainult seda pilti, mida osaleja parasjagu vaatab. Näiteks Maakera sisemust polegi vaja kunagi välja joonistada, ega ka rakkusid ja kosmost. Piisab hägusest pildist. Vajadusel saab simuleerija teha ka parandusi ja simulatsiooni tagasi kerida, näiteks mõne vastuolu ilmnemisel. Simulatsiooni juhendaja saab vastuolulisi kohti vahele jätta. Seega on igati tõenäoline, et inimõistuse simuleerimiseks piisab neuronite või neuronite sisese taseme simuleerimisest.
Hinnanguline realistliku inimajaloo simulatsiooni kulu on seega $\sim 10^{33} - 10^{36}$ operatsiooni (100 milardit inimest * 50 aastat/inimene * 30 milj sek/aastas * [$10^{14}, 10^{17}$] operatsiooni igas inimajus sekundis $\approx 10^{33} - 10^{36}$).
Ja nagu juba öeldud, saab juba teadaoleva nanotehnoloogia baasil luua planeedi suuruse arvuti, mis teeb 10^{42} operatsiooni sekundis ning võib eeldada, et see on kaugel optimaalsest [8]. (! U.P.: tuletan meelde, et oleme musta augu piiril ja selline väide on ekstreemselt naiivne ja jabur).
Üks taoline arvuti suudab teostada kogu inimkonna ajaloo simulatsiooni, kasutades vähem kui miljondikku oma arvutusvõimsusest ühes sekundis. Tuleviku inimesed võivad ehitada astronoomilise hulga selliseid arvuteid. Järeldus:

- Tulevikuinimestel on piisavalt arvutusvõimsust, et teostada ülipalju eellaste simulatsioone, kasutades seejuures vaid tibatillukest osa saadaolevast arvutusvõimsusest.

III. Kriitika:

- Bostrom väidab samaaegselt, et lähtume füüsikaseadustest ning seda, et me ei välista füüsikateoorias võimatuid sündmusi. Kuidas sellest aru saada? Minu arvates ilmne vastuolu.
- Lloyd [7] kirjutab muuhulgas, et arvuti töötemperatuur on 1 miljard kraadi K; praeguste füüsikaseaduste kohaselt ei saa selline arvuti täisvõimsusel töötada. Tänaseks arvatakse, et maksimumpiir on 10^{43} operatsioonile sekundis. Sellise kiiruse saavutamisel läheb arvutaja musta auku. (vt. S.Aaronson)
- Arvutuste maht $\approx 10^{33} - 10^{36}$ operatsiooni tähendab, et me loendame üle ja esitame nii mitu olekut. See aga eeldab “super”-algoritmi olemasolu, mis söödab arvutile ette neid olekuid, mida kuvada. Kust tuleb algoritm ehk: ***Kes kirjutab stsenaariumit?***
- Elektronide tasemel arvuti maksimaalne töökiirus on 10^{22} operatsiooni sekundis (viide?)
- Kohe valmiv planeedisuurune arvuti, millele Bostrom viitab: allikas [8]. Tegemist on suvalise inimese poolt kokku pandud fantaasiaga, mida on küll põnev lugeda, aga mida ei saa pidada piisavalt kaalukaks ja autoriteetseks allikaks. Isegi 12.sep.2019 antud videointervjuus [10] viitab Bostrom antud allika järeldustele, kui kindlale teadusfaktile!

- Arvestades, et ajus on 10^{14} sünapsi, siis on kogu olekute hulga suuruseks $2^{10^{14}}$, ehk $\approx 10^{10^{13}} \approx 10^{10\,000\,000\,000\,000}$ erinevat moodust kuidas valida sünapse, mis võiksid esindada mingit teadvuse olekut. On üsna ilmne, et see arv on nii kolossaalselt suur, et ilma “ülihea” algoritmita pole reaalset lootust leida sobivaid olekuid ja üleminekuid ühest olekust teise.
- Üksteise sees toimuvad simulatsioonid: kõrgemal tasandil “inimese” vaatenurgast kasvavad arvutusmahud eksponentsiaalselt, seega ka arvutuste kiirus. Lõpetavad mustas augus pigem varem, kui hiljem!
  Simulatsioon võiks olla kiirem reaalajast näiteks, 1 reaal-aasta sekundis (10 miljonit sekundit aastas): $10^{14} \times 10^{7} = 10^{21}$ operatsiooni sekundis. Sel juhul piisab, et tekib ainult 1 lisa sisemine tasand, kui kogu kaadervärk lendab musta auku! St iga simuleeritav maailm simuleerib samadel tingimustel edasi (nagu Bostrom pakub): $10^{21} \times 10^{21} = 10^{42}$.
  Ja mitte unustada ka arvuti töötemperatuuri suurusjärgus 1 miljard kraadi.

IV. Simulatsiooniargumendi tuum

Põhiidee seisneb järgnevas: kui oleks märkimisväärne võimalus, et meie tsivilisatsioon areneb posthuman tasemele ja teostab palju eellaste simulatsioone, siis kuidas saab olla nii, et sina ei ela ühes sellises simulatsioonis?
Teiste sõnadega: kui on väga tõenäoline, et meie tsivilisatsioon areneb posthuman tasemele ja teostab palju eellaste simulatsioone. Võiks järeldada, et ka meie elame juba ühes sellises simulatsioonis.

f_p : inim-tasemel tehnoloogilisel arengutasemel tsivilisiatsioonide osakaal, mis jõuavad posthuman tasemeni.

$\overline{N}$: Keskmine arv eellaste simulatsioone, mida teostab üks posthuman tsivilisatsioon.

$\overline{H}$: Keskmine arv indiviide, kes on elanud tsivilisatsioonis enne, kui see jõuab posthuman tasemeni.

Tegelik osakaal kõikidest inimtüüpi kogemustega jälgijatest, kes elavad simulatsioonides:

$$f_{sim} = \frac{f_p \overline{N}\,\overline{H}}{f_p \overline{N}\,\overline{H} + \overline{H}}$$

Olgu:

f_i : eellaste simulatsioonidest huvitatud tsivilisatsioonide osakaal.

$\overline{Ni}$: keskmine arv eellaste simulatsioone mida huvitatud tsiviliatsioonid teostavad.

Saame:

$$\overline{N} = f_i \overline{Ni}$$

Seega:

$$f_{sim} = \frac{f_p f_i \overline{\overline{Ni}}}{f_p f_i \overline{\overline{Ni}} + 1} \qquad (*)$$

Kuna posthuman tsivilisatsioonil on erakordselt võimsad arvutid, siis ni on väga suur, nagu me nägime eelmises peatükis. Vaadeldes (*) võime näha, et *vähemalt üks* järgnevast kolmest väitest peab olema tõene:

(1) $f_p \approx 0$

(2) $f_i \approx 0$

(3) $f_{sim} \approx 1$

IV. Kriitika:

- Valem (*) on ebakorrektne. Autorid märkasid seda ja tegid paranduse 2011.aastal [9]! Seda asjaolu arvestades tundub, et mitte keegi ei vaevunud 10 aasta jooksul süübima argumendi *tuuma* valemisse. Autorite arvates on viga valemis nüüdseks parandatud [9] ja kõik on hästi.
- Argumendi kogutõenäosus on pigem tinglik tõenäosus. Boström esitab argumendi tõenäosuse, eeldades, et kõik sellele eelnevad vajalikud ja piisavad sündmused toimuvad tõenäosusega 100%. Arvestades ülaltoodud kriitikat võiks argumenteerida, et “tegelik” tõenäosus on P(Simulatsiooniargument) tingimusel, et P(Eeldused on täidetud). Eelduste täitmise tõenäosus võiks olla midagi järgnevat:
  P(Eeldused) = P(Töötame välja vastava matemaatilise mudeli) * P(Arvutusmudel simuleeritav (digitaal)arvutil) * P(Vastav arvuti on realiseeritav, st suudame reaalselt ehitada) * P(Avastame piisavalt hea otsingu/õppimis- algoritmi).
  P(Eeldused) = 0.5 * 0.5 * 0.5 * 0.5 = 0.0625
  Seega võiks väita, et tõenäosus, et oleksid täidetud eeldused simulatsiooni-võimeks, on maksimaalselt 0.0625. Lähtume siin samast “võrdse tõenäosuse printsiibist” - kuna meil pole täpsemaid tõendeid, siis võib iga eeldus eraldivaadatuna täituda 50% tõenäosusega.

V. Ükskõiksusprintsiip (võrdse tõenäosuse printsiip)

Kuna meil puuduvad selged tõendid, kuidas jaotada tõenäosusi alternatiivide vahel, siis jaotame sündmuste toimumiste tõenäosused võrdselt kõigi võimaluste vahel. Argumendi 3 osa saavad igaüks tõenäosuseks 1/3.

VI. Tõlgendus

Kui (1) on tõene, siis inimkon(na)d ei jõua posthuman tasemeni. Seega on üsna keeruline õigustada miks peaks meie tsivilisatsioon olema priviligeeritud seisuses ja vältides katastroofe jõudma posthuman tasemeni. Võime määrata inimkonna hukkumise tõenäosuseks ligikaudu 100%. Alternatiivselt võib olla, et madalama taseme tsivilisatsioonid jäävad Maal ellu alati, samas kui teatud arengutasemeni jõudvad tsivilisatsioonid hukkuvad.
Alternatiivi (2) tõesuse korral peaks eksisteerima mingi veenev põhjus, miks kaugelearenenud tsivilisatsioonid ei tahaks simuleerida maailmasid. Võib-olla on jõukus hajutatud, et kellelgi üksikuna pole vastavat ressurssi. Või ei anna selline simulatsioon piisavalt uut informatsiooni ja on seetõttu ebahuvitav. Või moraalsetel põhjustel. jne, jne.

Alternatiiv (3) on kõige intrigeerivam. Maailm, mida me näeme, pole “reaalne”, see pole reaalsuse algtasand. Sellised simulatsioon-maalilmad on nagu “virtuaalmasinad” arvutimaailmas, kus üks programm jookseb teise - simuleeritud - arvuti sees. Niimoodi võib simulatsioone aheldada lõpmatult üksteise sisse. Kui me suudame luua eellaste simulatsioone, siis see oleks tugev tõendus (1) ja (2) vastu, millest peaksime järeldama, et me elame simulatsioonis. Veel enam, peaksime kahtlustama, et meie simuleerijad on ise ka simuleeritud olevused; ja nende loojad omakorda võivad samuti olla simuleeritud. Reaalsusel on seega mitu tasandit. Tasandite arvu võib piirata ressursside piiratud maht. Näiteks võime eeldada, et meie simulatsioon võidaks peatada kui saavutame posthuman taseme.
Eeldades, et elame simulatsioonis, millised võiksid olla järeldused meile, inimestele? Jätkame elu ja planeerimist, nagu oleme siiani teinud. See tundub kõige mõistlikum alternatiiv.

VII. Järeldus

Tehnoloogiliselt edasiarenenud (mature) “posthuman” tsivilisatsioonil on kasutada ülimalt suur arvutusvõimsus. Põhinedes sellele kogemuslikule faktile, näitab simulatsiooniargument, et *vähemalt üks* järgnevatest väidetest on tõene:

(1) “Posthuman” tasemeni arenenud inimtsivilisatsioonide osakaal on ligikaudu null.
(2) “Posthuman” tasemeni arenenud ja eellaste simulatsioonist huvitatud inimtsivilisatsioonide osakaal on ligikaudu null.
(3) Meiesarnaste kogemustega inimeste osakaal, kes elavad simulatsioonis on ligikaudu üks.

Kui (1) on tõene, siis peaaegu kindlasti me sureme välja enne “posthuman” tasemeni jõudmist. Kui (2) on tõene, siis peavad edasijõudnud tsivilisatsioonid tugevalt koonduma niimoodi, et neis poleks ühtegi suhteliselt jõukat indiviidi, kes sooviks teostada eellaste simulatsiooni. Kui (3) on tõene, siis me elame peaegu kindlasti simulatsioonis. Kuna meil puudub täpsem informatsioon, siis on mõistlik jaotada hinnangulised tõenäosused võrdselt (1), (2) ja (3) vahel.
Kui on täidetud tingimus, et me praegu ei ela simulatsioonis, siis meie järeltulijad ei teosta peaaegu kindlasti mitte kunagi eellaste simulatsiooni.

Kasutatud kirjandus

[1] Bostrom, Nick, “ARE YOU LIVING IN A COMPUTER SIMULATION?”, Philosophical Quarterly (2003) Vol. 53, No. 211, pp. 243-255. (First version: 2001)]; https://www.simulation-argument.com/simulation.html
[2] https://en.wikipedia.org/wiki/Principle_of_indifference
[3] http://mathworld.wolfram.com/PrincipleofInsufficientReason.html
[4] Bostrom, Nick “The Simulation Argument: Why the Probability that You Are Living in a Matrix is Quite High”, 2003, https://www.simulation-argument.com/matrix.html
[5] Bremermanni limiit, https://en.wikipedia.org/wiki/Bremermann%27s_limit
[6] Beckensteini piir, https://en.wikipedia.org/wiki/Bekenstein_bound
[7] S. Lloyd, “Ultimate physical limits to computation.”, 2000, https://arxiv.org/abs/quant-ph/9908043
[8] Bradbury, J. Robert, “Matrioshka Brains”, 1999, https://www.gwern.net/docs/ai/1999-bradbury-matrioshkabrains.pdf
[9] Bostrom, Nick; Kulczycki, Marcin; “A Patch For The Simulation Argument”, 2011; [Published in: Analysis, Vol. 71, No. 1 (2011): 54-61] https://www.simulation-argument.com/patch.pdf
[10] Nick Bostrom: Why Our Brains Themselves May Be Simulated, JRE #1350; 12.sep.2019; https://youtu.be/Td_qaNy1W9U