Garrison Lovely:
Kan mänskligheten överleva AI?

[Ur Jacobin magazine, 22 januari 2024. Översättning från engelska, Göran Källqvist.]

Utvecklingen av konstgjord intelligens (AI) går oerhört fort, och några av världens rikaste män kanske just nu avgör mänsklighetens öde.

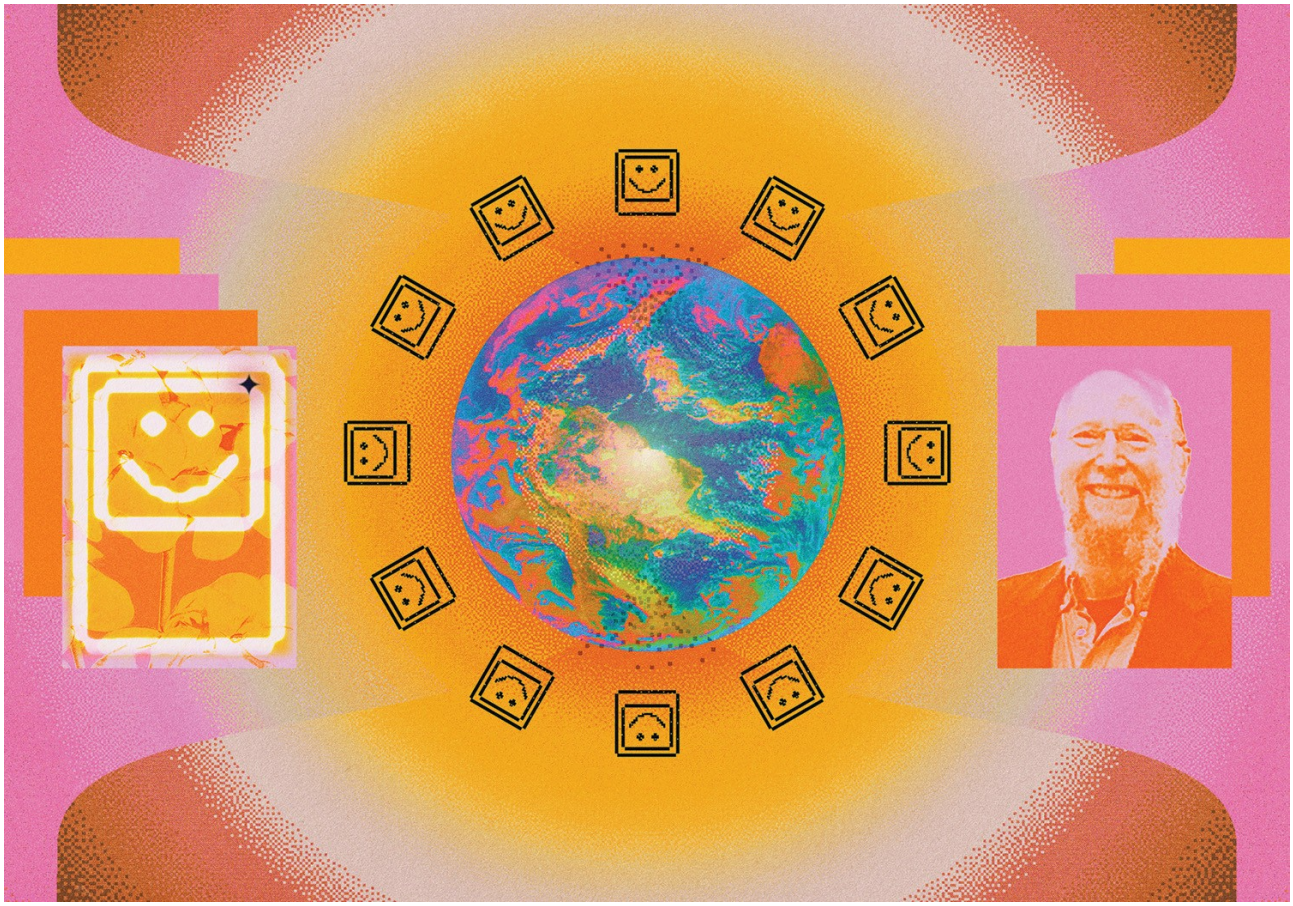

Illustration av Ricardo Santos.

Googles medgrundare Larry Page tror att ett superintelligent AI ”bara är nästa steg i evolutionen”. I själva verket har Page, som är värd omkring 120 miljarder dollar, hävdat att alla försök att förhindra en av AI pådriven utrotning och skydda mänsklighetens medvetenhet är ”artdiskriminering” och ”sentimentalt nonsens”.[1]

I juli sa Google DeepMinds ledande forskare Richard Sutton – en av pionjärerna bakom förstärkt inlärande, ett av AI:s viktigare underområden – att teknologin ”skulle kunna tränga undan oss ur existensen”, och att ”vi inte bör göra motstånd mot tronföljden”. I ett tal 2015 frågade Sutton, att om ”allting misslyckas” och AI ”dödar oss alla”, ”Är det så illa att människan inte är den slutliga formen för intelligent liv i universum?”[2]

1 Vanity Fair, oktober 2023; New York Times, 3 december 2023; Time, 6 september 2023.
2 YouTube, juli 2023; YouTube, 2015.

Poängen är inte biologisk utrotning”, sa Sutton, 66 år, till mig. ”Mänsklighetens klarhet och vår förståelse, vår intelligens – vårt medvetande om du så vill – kan fortsätta utan människor av kött.”

Yoshua Bengio, 59, är den näst mest citerade nu levande vetenskapsmannen, och uppmärksammad för sitt grundläggande arbete om djup inlärning. Som ett svar till Page och Sutton, sa Bengio till mig: ”Jag tycker att det de vill är att spela tärning med mänsklighetens framtid. Personligen tycker jag att det borde kriminaliseras.” Något överraskad frågade jag exakt vad han ville förbjuda, och han sa, försöken att bygga ”AI-system som skulle kunna övermanna oss och ha ett eget inbyggt egenintresse.” I maj började Bengio skriva och tala om hur avancerade AI-system skulle kunna löpa amok och utgöra en risk för att ta död på mänskligheten.[3]

Bengio hävdar att framtida, AI-system på verkligt mänsklig nivå skulle kunna förbättra sina egna förmågor, och funktionellt skapa en ny, intelligentare art. Mänskligheten har utrotat hundratals andra arter, till stor del av en händelse. Han fruktar att vi kan bli den nästa – och han är inte ensam.[4]

Bengio delade 2018 års Turingpris, datorvärldens Nobelpris, med två andra pionjärer inom djup inlärning,Yann LeCun och Geoffrey Hinton. Hinton är den mest citerade nu levande vetenskaps-mannen, och skapade svallvågor i maj när han avgick från sin ledande roll på Google för att mer fritt tala om möjligheten att framtida AI-system skulle kunna utrota mänskligheten. Hinton och Bengio är de två mest framstående AI-forskarna som har anslutit sig till gruppen ”x-risk”. Denna löst sammansatta grupp kallas ibland förespråkare för AI-säkerhet eller ”domedagsprofeter”, och oroar sig för att AI utgör en existentiell risk för mänskligheten.

Samma månad som Hinton lämnade Google skrev hundratals AI-forskare och framträdande figurer under ett öppet brev, som sa: ”Det borde vara en global prioritet att minska risken för att utrotas av AI, jämte andra risker i samhällelig skala som pandemier och kärnvapenkrig.”[5] Hinton och Bengio var ledande undertecknare, följda av OpenAI:s VD Sam Altman och chefer för andra ledande AI-laboratorier.

Hinton och Bengio var också de första författarna till ett strategidokument från oktober om risken för att ”mänskligheten fullständigt tappar kontrollen över autonoma AI-system”,[6] till vilket berömda akademiker som Nobelpristagare Daniel Kahneman och författaren till *Sapiens*, Yuval Noah Harari anslöt sig.

LeCun, som sköter AI på Meta, håller med om att AI på mänsklig nivå kommer att komma, men sa i en offentlig debatt med Bengio om att AI kan utrota oss: ”Om det är farligt så kommer vi inte att bygga det.”[7]

De mest avancerade AI-systemen i världen, från DeepMinds modell för proteinvikning till stora språkmodeller (LLM) som OpenAI:s ChatGPT, drivs av djup inlärning. Ingen förstår egentligen hur djupa inlärningssystem fungerar, men deras prestationsförmåga har ändå fortsatt att bli bättre. Dessa system är inte tänkta att fungera enligt en rad begripliga principer, utan ”tränas” istället i att analy-sera mönster i stora datamängder, där det som en konsekvens uppstår ett komplicerat beteende –

3 Yoshua Bengios webbplats, 22 maj 2023.
4 Ibid, 24 juni 2023.
5 ”Statement on AI Risk”, *Center for AI Safety*.
6 ”Managing AI Risks in an Era of Rapid Progress”, oktober 2023.
7 YouTube, juli 2023.

som språkförståelse.[8] AI-utvecklaren Connor Lehay berättade för mig att ”det är mer som att peta i något i en petriskål” än att skriva kod. Strategidokumentet från oktober varnar för att ”ingen för närvarande vet hur AI:s agerande på ett pålitligt sätt ska anpassas till komplicerade värderingar”.

Trots all denna osäkerhet ser AI-företag sig som deltagare i en tävling om att göra dessa system så kraftfulla som de kan – utan någon fungerande plan för att förstå hur de saker de skapar faktiskt fungerar, och samtidigt som de fuskar med säkerheten för att vinna marknadsandelar. Artificiell allmän intelligens (AGI) är den heliga graal som ledande AI-laboratorier uttryckligen arbetar för. AGI definieras ofta som ett system som är minst lika bra som människor på nästan vilken intellektuell uppgift som helst. Det är också det som Bengio och Hinton tror kan bli slutet för mänskligheten.

Bisarrt nog tror många av de personer som aktivt verkar för AI-förmågor att det är en stor chans att det till slut kommer att orsaka en apokalyps. En enkät bland forskare av maskininlärning visade att nästan hälften av dem trodde att det var minst 10% risk att avancerad AI skulle kunna leda till att ”mänskligheten utrotas eller permanent och allvarligt sätts ur spel”.[9] Bara några månader innan Altman var med och grundade OpenAI, sa han: ”AI kommer troligen att leda till världens slut, men under tiden kommer det att finnas mycket fina företag.”[10]

Allmänhetens åsikter om AI har blivit mer negativa, speciellt under året efter att ChatGPT släpptes. I alla undersökningar 2023 utom en trodde de flesta amerikaner att AI kunde utgöra ett existentiellt hot mot mänskligheten. De sällsynta tillfällen då opinionsundersökarna frågade folk om de ville ha AI på mänsklig nivå eller högre sa starka majoriteter i USA och Storbritannien att de inte ville det.[11]

När socialister har diskuterat AI har det hittills varit för att belysa AI-driven diskriminering eller varna för möjliga negativa effekter av automatisering i en värld med svaga fackföreningar och mäktiga kapitalister. Men vänstern har varit påtagligt tyst om Hintons och Bengios mardrömsscenario – att avancerad AI skulle kunna döda oss alla.

Oroande förmågor

Även om mycket av gruppen x-risks uppmärksamhet riktar in sig på tanken att mänskligheten så småningom skulle kunna tappa kontrollen över AI, är många också oroade för att mindre kapabla system kan göra dåliga aktörer med mycket korta tidsplaner starkare.

Tack och lov är det svårt att göra biologiska vapen. Men det kan snart förändras.

Anthropic är ett ledande AI-laboratorium som har grundats av en säkerhetsmedveten tidigare anställd vid OpenAI, och har nyligen arbetat med experter på biologisk säkerhet för att se hur mycket LLM skulle kunna hjälpa en bioterrorist. I ett vittnesmål inför en underkommitté i senaten i juli, rapporterade Anthropics VD Dario Amodei att vissa steg i tillverkningen av biologiska vapen inte går att hitta i läroböcker eller sökmotorer, men att ”dagens AI-verktyg kan fylla i en del av de här stegen, om än ofullständigt”, och att ”en enkel extrapolering av dagens system till de vi förväntar oss se om två eller tre år antyder en avsevärd risk för att AI-system kommer att kunna fylla i de

8 Ars Technica, 31 juli 2023.
9 AI Impacts, 3 augusti 2022.
10 Future of Life, 6 juni 2015.
11 AIPI Survey, november 2023.

saknade delarna.”[12]

I oktober rapporterade *New Scientist* att Ukraina för första gången använde dödliga självstyrande vapen (LAW) – bokstavligt talat mördarrobotar. USA, Kina och Israel utvecklar egna LAW. Ryssland har tillsammans med USA och Israel gått mot ny internationell lagstiftning om LAW.[13]

Men många kritiserar den mer omfattande tanken att AI utgör en existentiell risk, och den grumliga diskussionen om AI är svår att tolka: folk med lika höga vitsord gör motsatta påståenden om riskerna för utrotning med AI är verkliga, och riskkapitalister skriver under öppna brev tillsammans med progressiva AI-etiker. Och även om tanken på utrotningsrisk verkar vinna mark snabbast, så publicerar stora tidskrifter nästan varje vecka artiklar som hävdar att risken för utrotning distraherar från existerande risker. Samtidigt ägnar sig oändligt mycket mer pengar och människor i tysthet åt att göra AI-systemen kraftfullare än åt att göra dem säkrare eller mindre partiska.

Vissa är inte rädda för det ”science fiction”-scenario där AI-modeller blir så kunniga att de tar kontrollen ur våra svaga händer, utan istället att vi kommer att ge ensidiga, sköra och hallucinerande system alltför mycket ansvar och öppnar en mer alldaglig Pandoras ask full av hemska men bekanta problem, som förstoras av de algoritmer som orsakar dem. Denna grupp av forskare och förespråkare – ofta kallade ”AI-etiker” – lutar åt att koncentrera sig på de omedelbara skador som AI orsakar, och utforskar lösningar inklusive modellernas ansvarsskyldighet, algoritmernas transparens och maskininlärningsrättvisa.

Jag talade med några av de mest framträdande rösterna från den AI-etiska gruppen, som datorvetenskapsmännen Joy Buolamwini och Inioluwa Deborah Raji. Båda har genomfört banbrytande forskning om existerande skador som har orsakats av diskriminerande och bristfälliga AI-modeller, vars effekter enligt deras åsikt döljs ena dagen och överdrivs nästa. Liksom många AI-etiska forskares arbete är deras arbete en blandning av vetenskap och aktivism.

De jag talat med bland AI-etiker uttryckte till största delen en åsikt att AI:s framtid inte står inför i grunden nya utmaningar, som en framtidsutsikt av fullständig teknologisk arbetslöshet eller utrotning, utan mer ser ut som ökad rasdiskriminering vid fängslanden och beslut om lån, en förändring av arbetsplatserna till något som liknar Amazon, angrepp på arbetande fattiga, och ett fortsatt befästande och berikande av teknoeliten.[14]

Ett ofta förekommande argument från denna grupp är att talet om utrotning överskattar vad Big Techs produkter är kapabla till, och på ett farligt sätt ”distraherar” från AI:s omedelbara skador. I bästa fall, säger de, är det ett slöseri med tid och pengar att överväga tanken på risk för utrotning. I värsta leder det till katastrofala politiska idéer.

Men många av de som tror på en risk för utrotning betonade att ståndpunkterna ”AI orsakar skador nu” och ”AI skulle kunna göra slut på världen” inte utesluter varandra. Vissa forskare har uttryckligen försökt överbrygga klyftan mellan de som riktar in sig på existerande skador och de som fokuserar på utrotning, och framhäver möjliga delade politiska mål. AI-professorn Sam Bowman är en annan person vars namn finns på brevet om utrotning. Han har forskat för att hitta och minska

12 YouTube, juli 2023.
13 New York Times, 21 november 2023.
14 Technology Review, 21 januari 2019; The Markup, 25 augusti 2021; Jacobin, juli 2023; Technology Review, 18 april 2023; The Conversation, 27 april 2023.

algoritmernas ensidighet och granskar bidrag till de största AI-etiska konferenserna. Samtidigt har Bowman uppmanat fler forskare att arbeta på AI:s säkerhet och skrev om ”faran att underskatta” LLM:s förmåga.

X-risk-gruppen jämför sig vanligtvis med klimatdebattörer, och frågar om fokus på att minska klimatförändringarnas långsiktiga skador på ett farligt sätt avleder från skador på kort sikt av luftföroreningar och oljeutsläpp.

Men de medger själva att alla från x-risk-gruppen inte har varit så diplomatiska. I en rad kryddade politiska inlägg i augusti 2022 om AI, twittrade medgrundaren av Anthropic, Jack Clark, att ”vissa personer som arbetar med en långsiktig/AGI-liknande politik tenderar att bortse från, minimera eller helt enkelt inte överväga de omedelbara problemen med spridning/skador av AI.”[15]

”AI kommer att rädda världen”

Ett tredje läger oroar sig för att vi när det gäller AI faktiskt inte rör oss tillräckligt snabbt. Framträdande kapitalister som miljardären Marc Andreessen håller med säkerhetsfolk om att AGI är möjligt, men hävdar att det istället för att döda oss alla kommer att ge upphov till en oöverskådlig gyllene tidsålder med omvälvande överflöd och teknologier på gränsen till det magiska. Denna grupp kommer till största delen från Silicon Valley och kallas vanligtvis för AI-boosters, och har en benägenhet att oroa sig mycket mer för att överdrivna regleringar av AI kommer att hämma en omvandlande, världsräddande teknologi i sin linda och döma mänskligheten till ekonomisk stagnation.

En del teknooptimister föreställer sig en AI-driven utopi som får Karl Marx att verka fantasilös. *Guardian* publicerade nyligen en minidokumentär med intervjuer från 2016 till 2019 med OpenAI:s främste vetenskapsman, Ilja Sutskever, som djärvt förkunnar: ”AI kommer att lösa alla problem vi har idag. Det kommer att lösa sysselsättning, sjukdomar, fattigdom. Men det kommer också att skapa nya problem.”[16]

Andreessen håller med Sutskever – ända fram till ”men”. I juni publicerade Andreessen en artikel med titeln ”Varför AI kommer att rädda världen”,[17] där han förklarar hur AI kommer att göra ”allt vi bryr oss om bättre”, så länge vi inte reglerar det till döds. Han följde upp det i oktober med sitt ”Teknooptimistiska manifest”, som utöver att hylla en av den italienska fascismens grundare, kallade tankar om ”existentiell risk”, ”hållbarhet”, ”ansvar och säkerhet” och ”tekniketik” för fiender till framsteg. Andreessen skräder inte med orden, och skriver: ”Vi tror att varje bromsande av AI kommer att kosta liv. Dödsfall som hade gått att förhindra med AI som hindrades existera [är] en sorts mord.”

Tillsammans med sin ”bror inom läkemedelsindustrin”, Martin Shkreli, är Andreessen den mest berömda förespråkare för ”effektiv accelerationism”, också kallat ”e/acc”, ett till största delen internetbaserat nätverk som blandar kultartad scientism, hyperkapitalism och naturalistiska vanföreställningar. E/acc blev viralt i somras, och bygger på den reaktionära författaren Nick Lands teori om accelerationism, som hävdar att vi måste intensifiera kapitalismen för att pressa oss framåt till en posthuman, AI-driven framtid. E/acc tar denna tanke och lägger till ett lager av fysik och

15 Twitter, 6 augusti 2022.
16 The Guardian, 2 november 2023.
17 ”Why AI Will Save the World”, 6 juni 2023.

internetfenomen, som anpassar den till en viss undergrupp inom Silicon Valleys eliter. Den bildades som reaktion på krav från ”decels” att bromsa AI, krav som till stor del har kommit från gruppen Effektiv altruism (EA), från vilken e/acc tar sitt namn.

AI-boostern Richard Sutton – vetenskapsmannen som är beredd att säga adjö till ”köttmänniskor” – arbetar nu vid KeenAGI, ett företag som nyss startats av John Carmack, den legendariske programmeraren bakom 1990-talets datorspel *Doom*. Företagets uppdrag är enligt Carmack: ”AGI eller gå under, med hjälp av galen vetenskap!”[18]

Kapitalismen gör det värre

I februari twittrade Sam Altman att Eliezer Yudkowsky så småningom kanske ”förtjänar Nobels fredspris”. Varför? Eftersom Altman ansåg att den självlärda forskaren och författaren av Harry Potter-fanfiction hade gjort ”mer för att skynda på AGI än någon annan”.[19] Altman åberopade hur Yudkowsky hjälpte DeepMind att säkra väsentlig tidig finansiering från Peter Thiel, liksom Yudkowskys ”avgörande” roll ”för beslutet att starta Open AI”.[20]

Yudkowsky var accelerationist innan begreppet ens hade myntats. Sjutton år gammal publicerade han – trött på diktaturer, världssvält och till och med själva döden – ett manifest som krävde att det skulle skapas en digital superintelligens för att ”lösa” mänsklighetens alla problem.[21] Under sitt livs följande årtionde förvandlades hans ”teknofili” till fobi, och 2008 skrev han om sin förvandling, och medgav: ”att säga *jag förstörde nästan världen!* skulle ha varit alltför högfärdigt.”

Yudkowsky är nu berömd för att popularisera tanken att AGI kan döda alla, och har blivit den största domedagsprofeten bland AI-domedagsprofeterna. En hel generation teknikintresserade läste Yudkowskys blogg när de växte upp, men fler av dem (mest konsekvent kanske Altman) anammade hans påstående att AGI skulle bli det viktigaste någonsin än hans uppfattningar om hur svårt det skulle bli att undvika att det dödar oss.

Under vårt samtal jämförde Yudkowsky AI med en maskin som ”trycker guld” ända fram tills den ”sätter eld på atmosfären”.

Och oavsett om den kommer att sätta eld på atmosfären eller ej, så trycker denna maskin nu guld fortare än någonsin. Uppsvinget för ”generativt AI” gör en del personer mycket, mycket rika. Sedan 2019 har Microsoft investerat 13 miljarder dollar i OpenAI. Som en följd av ChatGPT:s galna framgång ökade Microsoft nästan 1 biljon dollar i värde under året efter att produkten släpptes. Idag är det nästan 50 år gamla företaget värt mer än Google och Meta tillsammans.

Under jakten på rikedomar – eller helt enkelt för äran att ha skapat en digital superintelligens, som, twittrade Sutton, ”kommer att bli tidernas största intellektuella bedrift … vars betydelse går utöver mänskligheten, utöver livet, utöver bra och dåligt”[22] – kommer aktörer som vinstmaximerar att fortsätta att pressa på framåt, och lägga riskerna på alla oss andra, som aldrig har gått med på att ta på oss dessa risker. Troligen kommer trycket från marknaden att pressa företagen att föra över alltmer makt och självständighet till AI-systemen vartefter de förbättras.

18 Twitter, 19 augusti 2022.
19 Twitter, 3 februari 2023.
20 Cade Metz, Genius Makers, Penguin Random House, februari 2022.
21 ”Staring into the Singularity”, 1996.
22 Twitter, 30 september 2022.

En forskare på Google AI skrev till mig: ”Jag tror att storföretagen har så bråttom att erövra marknadsandelar att [AI-] säkerheten betraktas som en sorts löjlig störande faktor.” Bengio sa till mig att han ser ”en farlig tävlan mellan företag”, som skulle kunna bli ännu värre.

I ett panikartat svar på Bings OpenAI-drivna sökmotor utropade Google ”kristillstånd”, ”kalibrerade om” sin riskbenägenhet, och skyndade sig att släppa sin LLM, mot personalens motstånd. I interna diskussioner kallade anställda Bard för ”en patologisk lögnare” och ”pinsam”. Google publicerade den ändå.[23]

Chefen för Center for AI Safety, Dan Hendrycks, sa att ”AI-utvecklingen till största delen drivs av … att fuska med säkerheten… Jag tror faktiskt inte att avsikter spelar särskilt stor roll i närvaro av detta intensiva konkurrenstryck.” Ironiskt nog är Hendrycks också säkerhetsrådgivare åt Elon Musks senaste satsning, xAI.

De tre ledande AI-laboratorierna började samtliga som oberoende, uppdragsorienterade organisationer, men är nu antingen helt underordnade teknologijättar (Google DeepMind) eller har tagit emot så många miljarder dollar i investeringar från biljondollarföretag att deras altruistiska uppdrag kan inordna sig under den oändliga jakten på värde för aktieägarna (Anthropic har tagit emot 6 miljarder dollar från Google och Amazon tillsammans, och Microsofts 13 miljarder dollar gav dem 49% av OpenAI:s vinstdrivna del). *New York Times* rapporterade nyligen att DeepMinds grundare har blivit ”alltmer oroliga över vad Google kommer att göra med deras uppfinningar. 2017 försökte de bryta sig loss från företaget. Google svarade med att öka DeepMinds grundares och deras personals löner och aktiebonusar. De stannade kvar.”[24]

En utvecklare vid ett ledande laboratorium skrev till mig i oktober, att eftersom ledningarna för dessa laboratorier i typfallet verkligen tror att AI kommer att undanröja behovet av pengar, är vinstorientering ”till största delen ett hjälpmedel” i insamlingssyfte. Men ”sedan pressar investerarna (oavsett om det är ett riskkapitalföretag eller Microsoft) på för vinststrävan”.

Mellan 2020 och 2022 strömmade mer än 600 miljarder dollar i företagsinvesteringar in i industrin , och en enda AI-konferens 2021 var värd för nästan 30.000 forskare. Samtidigt visade en bedömning i september 2022 att det bara fanns 400 heltidsanställda forskare i AI-säkerhet, och den första konferensen i AI-etik 2023 hade färre än 900 deltagare.

Med tanke på det sätt på vilket mjukvara ”åt upp världen”, bör vi förvänta oss att AI kommer att uppvisa en liknande vinnaren-tar-allt-dynamik, som kommer att leda till allt större koncentration av välstånd och makt. Altman har förutspått att ”kostnaden för intelligens” kommer att sjunka till nästan noll som ett resultat av AI, och 1921 skrev han att ”ännu mer makt kommer att gå från arbete till kapital”. Han fortsatte: ”Om den offentliga politiken inte anpassar sig efter det så kommer de flesta människor att få det sämre än idag.”[25] Och i sin tråd ”spicy take” [ungefär ”kryddstarka åsikter”] skrev Jack Clark: ”kapitalistisk stordrift är till sin natur antidemokratisk, och därmed är kapitalintensiv AI antidemokratisk.”[26]

Markus Anderljung är politisk chef för den ledande tankesmedjan för AI-säkerhet, och första

23 New York Times, 20 januari 2023; Bloomberg, 19 april 2023.
24 New York Times, 3 december 2023.
25 Sam Altman, ”Moores Law for Everything”, 16 mars 2021.
26 Twitter, 6 augusti 2022.

författare till en inflytelserik vitbok som är inriktad på att reglera ”AI:s gränsområden”. Han skrev till mig och sa: ”Om ni är oroliga för kapitalismen i sin nuvarande form så borde ni vara ännu mer oroliga för en värld där mycket stora delar av ekonomin sköts av AI-system som uttryckligen är upplärda att maximera profiterna.”

I juni 2021 höll Sam Altman med, och berättade för Ezra Klein om filosofin bakom grundandet av OpenAI: ”En av drivkrafterna som vi var mycket oroliga för var motivet att få obegränsad profit, där mer alltid är bättre…. Och jag tror att man i synnerhet med dessa mycket kraftfulla allmänna AI-system inte vill ha en drivkraft att maximera profiterna i all oändlighet.”[27]

Med en häpnadsväckande åtgärd som vida omkring har setts som det hittills viktigaste ögonblicket i debatten om AI-säkerhet, avskedade OpenAI:s ideella styrelse den 17 november 2023, fredagen före Thanksgiving, VD:n Sam Altman. Enligt OpenAI:s ovanliga stadgar har styrelsen sin skyldighet mot ”mänskligheten” istället för mot investerare eller anställda.[28] Som rättfärdigande åberopade styrelsen vagt Altmans brist på ärlighet, men höll sedan ironiskt nog till stor del tyst om sitt beslut.

Vid 15-tiden påföljande måndag tillkännagav Microsoft att Altman skulle sätta upp ett avancerat forskningslaboratorium med anställning för samtliga anställda på OpenAI, varav den överväldigande majoriteten hade skrivit under ett brev där de hotade att anta Microsofts erbjudande om inte Altman återanställdes. (Även om han verkar vara en populär VD, så är det värt att notera att avskedet störde en planerad försäljning av de av OpenAI:s aktier som ägdes av personalen och värderades till 86 miljarder dollar.) Strax efter kl 13 på onsdag tillkännagav OpenAI att Altmans återvänt som VD och att det kommit med två nya medlemmar i styrelsen: Twitters tidigare styrelseordförande, och den tidigare finansministern Larry Summers.

På mindre än en vecka hade OpenAI:s chefer och Altman samarbetat med Microsoft och företagets personal för att arrangera hans framgångsrika återkomst och avsätta större delen av de styrelsemedlemmar som låg bakom avskedet av honom. Microsoft ville helst att Altman skulle återkomma som VD. Det oväntade avskedet fick teknologigigantens aktie att rasa 5% (140 miljarder dollar) och tillkännagivandet av Altmans återinsättande tog den till en rekordnivå. Ovilliga att ”överrumplas” en gång till har Microsoft nu en plats utan rösträtt i den icke vinstdrivande styrelsen.

Omedelbart efter avskedet av Altman exploderade X, och det uppstod en berättelse som till största delen underblåstes av rykten och artiklar med anonyma källor, att säkerhetsinriktade oegennyttiga personer i styrelsen hade avskedat honom på grund av hans aggressiva kommersialisering av OpenAI:s modeller på bekostnad av säkerheten. Tonen i svaret som till överväldigande delen kom från e/acc sattes av grundaren med pseudonymen @BasedJeffBezos som skrev: ”EA-anhängare är i grund och botten terrorister. Att på en natt förstöra värden på 80B [biljoner dollar] är en terroristhandling.”[29]

Den bild som framträder från den senare rapporteringen var att det var en grundläggande misstro mot Altman, inte omedelbara dubier om AI-säkerhet, som låg bakom styrelsens val. *Wall Street Journal* konstaterade att ”det inte var en enstaka händelse som ledde till deras beslut att kasta ut Altman, utan en stadig, långsam urholkning av tilltron som gjorde dem alltmer betänkliga.”[30]

27 New York Times, 11 juni 2021.
28 Ars Technica, 18 november 2023.
29 Twitter, 20 november 2023.
30 Wall Street Journal, 22 november 2023.

Flera veckor innan avskedet rapporteras Altman ha använt en oärlig taktik för att försöka avlägsna styrelsemedlemmen Helen Toner för en akademisk uppsats som hon varit medförfattare till, och som han ansåg vara kritisk mot OpenAI:s förpliktelser mot AI-säkerhet. I uppsatsen berömde Toner, en till EA ansluten forskare i AI-styrning, Anthropic för att ha undvikit ”den sorts desperata genvägar som publiceringen av ChatGPT verkade stimulera till.”[31]

New Yorker rapporterade att ”en del av styrelsens 6 medlemmar ansåg att Altman var manipulativ och intrigant”.[32] Några dagar efter avskedet skrev en forskare i AI-säkerhet på DeepMind som tidigare hade arbetat för OpenAI, att Altman ”ljög för mig vid olika tillfällen” och ”var bedräglig, manipulativ och värre än så mot andra”,[33] en uppfattning som nyligen upprepats i reportage i *Time*.

Detta var inte första gången som Altman avskedades. 2019 avsatte Y Combinators grundare Paul Graham Altman från teknikacceleratorns ledning på grund av oro för att han prioriterade sina egna intressen över organisationens. Graham har tidigare sagt att ”Sam är otroligt bra på att bli mäktig”.[34]

OpenAI:s ovanliga modell för styrning upprättades specifikt för att hindra ett fördärvligt inflytande från vinstsyften, men som Atlantic med rätta förkunnade ”vinner pengar alltid”. Och mer pengar än någonsin går till att befrämja AI:s kapacitet.

Full fart framåt

AI:s utveckling har på senare tid drivits på av en kulmen på många decennielånga trender: ökning av beräkningskraften (kallat ”datorkraft”) och data för att lära upp AI-modeller, som själva har ökat i storlek genom betydande förbättringar av algoritmernas effektivitet. Sedan 2010 har den mängd datorkraft som används för att lära AI-modeller ökat ungefär *100 miljoner gånger*.[35] De flesta framsteg vi ser nu är resultatet av det som på den tiden var ett mycket mindre och fattigare område.

Och även om det senaste året förvisso har innehållit mer än sin skäliga andel av AI-hajp, så har dessa tre trender kombinerats och lett till mätbara resultat. Den tid det tar för AI-system att uppnå resultat på mänsklig nivå i många prestandatester har under det senaste årtiondet förkortats drastiskt.

Det är förstås möjligt att AI:s kapacitetsökning kommer att nå en gräns. Forskare kanske får slut på bra data att använda. Moores lag – observationen att antalet transistorer i mikrochips fördubblas vartannat år – kommer så småningom att bli historia. Politiska händelser kanske stör tillverkning och leveranskedjor och pressar upp kostnaderna för beräkningarna. Och större system kanske inte längre leder till bättre kapacitet.

Men i själva verket är det ingen som vet var de verkliga gränserna för de befintliga metoderna ligger. Ett urklipp ur en intervju med Yann LeCun i januari 2022 dök upp på Twitter i år. LeCun sa: ”Jag tror inte att vi kan lära en maskin att bli intelligent bara utifrån text, för jag tror att den mängd information om världen som finns i text är ynkligt jämfört med vad vi behöver veta.” För att illustrera denna punkt gav han ett exempel: ”Jag tar ett objekt och lägger det på bordet och knuffar iväg bordet. För mig är det helt uppenbart att objektet kommer att åka med bordet.” Men ”om man lär

31 New Yorker, 1 december 2023; ”Decoding Intentions”, CSET, oktober 2023.
32 New Yorker, op cit.
33 Twitter, 21 november 2023.
34 New Yorker, 3 oktober 2016.
35 Asterisk magazine, juni 2023.

upp en maskin, oavsett hur kraftfull den är, med en textbaserad modell, så kommer din ’GPT-5000’ … aldrig att lära sig det.”[36]

Men om man ger ChatGPT-3.5 det exemplet spottar det genast ur sig rätt svar.

I en intervju som publicerades 4 dagar innan han avskedades, sa Altman: ”Fram tills dess vi lär upp den modellen [GPT-5] är det som en rolig gissningslek för oss. Vi försöker bli bättre på det, för jag tror att det ur ett säkerhetsperspektiv är viktigt att förutsäga förmågan. Men jag kan inte här säga exakt vad den kommer att göra som GPT-4 inte gjorde.”[37]

Historien är full av dåliga förutsägelse om uppfinningstakten. En ledare i *New York Times* hävdade att det skulle kunna ta ”en miljon till tio miljoner år” att utveckla ett flygplan – 69 dagar innan bröderna Brother flög för första gången. 1933 avvisade ”kärnfysikens fader” Ernest Rutherford självsäkert möjligheten till en neutronutlöst kedjereaktion, vilket inspirerade fysikern Leo Szilard att *redan nästa dag* göra en hypotes om en fungerande lösning – som till slut blev grundläggande för att skapa atombomben.

En slutsats som verkar svår att undvika är att de personer som är bäst på att bygga AI-system på senare tid har börjat tro att AGI både är möjlig och omedelbart förestående. De två kanske ledande AI-laboratorierna, OpenAI och DeepMind, har ända sedan de började arbetat fram mot AGI, redan när man blev utskrattad om man trodde att det var möjligt inom en nära framtid. (Ilja Sutskever ledde taktfasta rop ”Känn AGI” vid OpenAI:s semesterfest 2022.)

Perfekta arbetare

Arbetsköpare använder redan AI för att övervaka, kontrollera och utnyttja arbetare. Men den verkliga drömmen är att utesluta människor ur systemet. Trots allt skrev Marx: ”Maskineriet är ett medel för att producera mervärde.”

AI-riskforskaren Ajeya Cotra på Open Philanthropy (OP) skrev till mig att ”den logiska slutpunkten för en maximalt effektiv kapitalistisk eller marknadsekonomi” inte kommer att innehålla människor, eftersom ”människor ju är väldigt ineffektiva varelser för att tjäna pengar”. Vi värdesätter alla dessa ”kommersiellt improduktiva” känslor, skriver hon, ”så om vi till slut får det bra och gillar resultatet, så kommer det att vara på grund av att vi började med makten och formade systemet så att det anpassar sig till mänskliga värden.”

OP är en EA-inspirerad stiftelse som finansieras av Facebooks medgrundare Dustin Moskovitz. Den är ledande finansiär av organisationer för AI-säkerhet, varav många nämns i denna artikel. OP gav också 30 miljoner dollar till OpenAI för att stöda arbetet med AI-säkerhet, två år innan laboratoriet startade en vinstdrivande del 2019. Jag har tidigare fått en engångssumma från EA Funds, som själv får stöd från OP, för att stöda förlagsarbete på *New York Focus*, ett ideellt undersökande nyhetsorgan som täcker politiken i New York. Efter att jag hade stött på EA 2017 började jag donera 10-20% av mina inkomster till ideella organisationer för global hälsa och icke industriella jordbruk, arbetade frivilligt som lokalorganisatör, och arbetade vid en närliggande ideell grupp mot global fattigdom. EA var en av de tidigaste grupperna som på allvar tog upp AI:s existentiella risker, men jag betraktade AI-folket med viss försiktighet, med tanke på hur osäkert problemet var och det

36 Twitter, 19 maj 2023.
37 Financial Times, 13 november 2023.

enorma, undvikbara lidande som äger rum just nu.

Ett medgörligt AGI vore en arbetare som kapitalister bara kan drömma om: outtröttlig, motiverad och inte tyngd av behovet av toalettbesök. Företagare från Frederick Taylor till Jeff Bezos avskyr de olika sätt på vilket människor inte är optimerade för produktion – och därmed sina arbetsköpares vinster. Redan före tiden med Taylors vetenskapliga styrning försökte industrikapitalismen få arbetarna att arbeta mer som de maskiner de arbetar bredvid och alltmer ersätts av. Som *Kommunistiska manifestet* så förutseende observerade, förvandlar kapitalisternas omfattande användning av maskiner arbetarna till ”ett tillbehör till maskinen”.

Men enligt AI-säkerhetsgruppen gör just de omänskliga egenskaper som skulle få Bezos att drägla AGI till en dödsfara för människor.

Explosionen: frågan om utrotning

De som hävdar att det finns existentiella risker resonerar vanligtvis så här: när AI-systemen väl når en viss tröskel kommer de att kunna förbättra sig själva och starta en ”intelligensexplosion”. Om ett nytt AGI-system blir tillräckligt smart – eller bara tillräckligt stort – så kommer det att kunna göra mänskligheten maktlös för gott.

Uppsatsen ”Hantera AI-risker” från oktober säger:

> Det finns inget grundläggande skäl att AI:s utveckling skulle sakta in eller stanna när den når en förmåga på mänsklig nivå…. Jämfört med människor kan AI-system agera fortare, ta till sig mer kunskap och kommunicera med en mycket högre bandbredd. Dessutom kan de utökas för att använda enorma datorresurser och kan kopieras i miljoner.[38]

Dessa kännetecken har redan möjliggjort övermänskliga förmågor: LLM kan ”läsa” en stor del av internet på månader, och DeepMinds AlphaFold kan utföra åratals av mänskligt laboratoriearbete på några dagar.

Här är en stiliserad version av tanken att tillväxt av ”datorbefolkningen” driver på en intelligensexplosion: om AI-system konkurrerar med mänskliga vetenskapsmän vid forskning och utveckling, kommer systemen att föröka sig snabbt och leda till motsvarigheten till att ett enormt antal nya, ytterst produktiva arbetare skulle träda in i ekonomin. Uttryckt annorlunda: om GPT-7 kan utföra de flesta av en mänsklig arbetares uppgifter, och det bara kostar lite extra att sätta en upplärd modell på en dags arbete, så skulle varje modell vara oerhört lönsam och utlösa en positiv återkoppling. Det skulle kunna leda till en virtuell ”befolkning” på miljarder eller fler digitala arbetare, var och en värd mycket mer än energikostnaderna för att driva dem. Sutskever tror att det är sannolikt att ”hela jorden kommer att vara täckt av solpaneler och datacentraler.”[39]

Dessa digitala arbetare kanske kan förbättra vår AI-design och skaffa sig sätt att skapa ”superintelligenta” system, vars kapacitet Alan Turing 1951 spekulerade om snart skulle överträffa våra svaga krafter”. Och, som en del förespråkare för AI-säkerhet hävdar, måste inte en enskild AI-modell vara superintelligent för att utgöra ett existentiellt hot. Det kanske bara behöver finnas tillräckligt många kopior av den. Många av mina källor liknade företag med superintelligenser, vars förmågor överträffar de ingående medlemmarnas förmågor.

38 ”Managing AI Risks in an Era of Rapid Progress”, oktober 2023.
39 YouTube, november 2023.

Den vanligaste invändningen är ”dra bara ur kontakten”. Men när en AI-modell är kraftfull nog att hota mänskligheten kommer den troligen att vara den mest värdefulla sak som finns. Det är nog lättare att ”dra ur kontakten” på börsen i New York eller Amazons webbtjänster.

En lat superintelligens kanske inte utgör någon större risk, och skeptiker som Allen Institute for AI:s VD Oren Etzioni, professorn i komplexitet Melanie Mitchell, och AI Now Institutes verkställande direktör Sarah Myers West sa alla till mig att de inte har sett övertygande bevis för att AI-system har blivit mer självstyrande. Anthropics Dario Amodei verkar vara överens om att nuvarande system inte uppvisar någon oroande kapacitetsnivå. Men ett helt passivt men tillräckligt kraftfullt system som sköts av en omoralisk aktör räcker för att oroa personer som Bengio.

Dessutom ökar både akademiker och industriföretagare ansträngningarna att göra AI-modeller mer självstyrande. Dagar före sitt avsked sa Altman till *Financial Times*: ”Vi kommer att göra dessa medel allt mäktigare … och härifrån kommer handlingarna att bli alltmer komplicerade…. Den mängd affärsvärde som kommer från att kunna göra det i varje kategori tror jag är rätt stor.”[40]

Vad ligger bakom hajpen?

Den oro som håller många av de som talar om existentiella risker vakna på nätterna är inte att en avancerad AI kan ”vakna upp”, ”bli ond” och besluta sig för att döda alla av ondska, utan snarare att den kommer att se oss som ett hinder för de mål den har, oavsett vilka de är. I sin sista bok, *Korta svar på stora frågor*, uttryckte Stephen Hawking detta, och skrev: ”Du är troligen inte en ond myrhatare som trampar på myror av ondska, men om du har ansvar för ett grönt vattenkraftsprojekt och det finns en myrstack i området som ska översvämmas så blir det tråkigt för myrorna.”[41]

Oväntade och oönskade beteenden kan bero på enkla mål, vare sig det är vinster eller ett AI:s belöningssystem. På en ”fri” marknad leder vinstjakt till monopol, marknadsföring på flera nivåer, förgiftning av floder och luft och oräkneliga andra skador.

Det finns ett överflöd av exempel på AI-system som uppvisar överraskande och oönskade beteenden. Ett program som var avsett att eliminera sorteringsfel i en lista raderade hela listan. En forskare överraskades av att hitta en AI-modell som ”spelade död” för att undvika att bli identifierad under säkerhetstester.[42]

Andra ser en konspiration från Tech-giganterna bakom denna oro. En del personer som är inriktade på omedelbara skador av AI hävdar att industrin aktivt främjar tanken på att deras produkter till slut kan göra slut på världen. Exempelvis säger Myers West från AI Now Institute att hon ”ser berättelserna om så kallade existentiella risker som en lek för att tömma rummet på luft och se till att det inte blir någon meningsfull rörelse just nu.” Märkligt nog säger sig Yann LeCun och vetenskapschefen på Baidu AI, Andrew Ng, hålla med.

När jag presenterar denna tanke till de som tror på existentiella risker svarade de ofta med en blandning av förvirring och förbittring. Ajeya Cotra från OP skrev tillbaka: ”Jag önskar att det vore mindre av en industriförknippad sak att oroa sig för existentiella risker, för jag tycker att det på egna

40 Financial Times, november 2023.
41 Stephen Hawking, *Korta svar på stora frågor*, Stockholm : Mondial, 2018. [Här översatt från engelska – öa.]
42 ”Emergent Abilities of Large Language Models”; ”The Surprising Creativity of Digital Evolution”, *MIT Press*, våren 2020.

förtjänster i grund och botten är en mycket industrifientlig åsikt att ha…. Om företagen bygger saker som kommer att döda oss alla så är det riktigt illa, och de borde begränsas mycket strikt av lagstiftning.”

GovAI:s Markus Anderljung kallade rädsla för obalans i maktförhållanden mellan myndigheter och företag för ”en normal reaktion som folk har”, men betonade att den politik han föredrog mycket väl kan skada industrins största aktörer.

En förståelig källa till misstankar är att Sam Altman nu är en av de personer som är med förknippad med tanken på existentiella risker, medan hans företag har gjort mer än något annat för att flytta fram gränsen för allmänt AI.

När OpenAI närmade sig att bli lönsamt och Altman kom närmare makten, förändrade VD:n dess-utom hur han talade offentligt. När han under en frågestund i januari 2023 fick frågan om vad AI i värsta fall kunde leda till, svarade han: ”Ljuset slocknar för alla.” Men när han besvarade en liknande fråga under ed vid en senatsutfrågning i maj nämner han inte utrotning. Och i sin kanske sista intervju innan han avskedades sa Altman: ”Jag tror faktiskt inte att vi alla kommer att utrotas. Jag tror att det kommer att bli härligt. Jag tror att vi är på väg mot den bästa världen någonsin.”[43]

I maj anmodade Altman kongressen att styra AI-industrin, men en undersökning i november visade att OpenAI:s halvt moderbolag Microsoft hade stort inflytande i det i slutändan fruktlösa lobbyarbe-tet för att utesluta ”basmodeller” som ChatGPT från reglering i EU:s kommande AI-lagstiftning. Och Altman gjorde själv en hel del eget lobbyarbete i EU, och hotade till och med att dra sig ur regionen om reglerna blev alltför besvärliga (hot som han snabbt drog tillbaka). I ett tal inför en panel av VD:ar i San Francisco några dagar innan han avsattes, sa Altman att ”de nuvarande model-lerna är utmärkta. Vi behöver ingen styrning här. Troligen inte ens för kommande generationer.”[44]

President Joe Bidens ”genomgripande” dekret om AI verkar hålla med: dess krav på att dela med sig information om säkerhetstester gäller bara modeller som är större än någon av de som troligen har lärts upp hittills. Myers West kallade den sortens ”storlekströsklar” för ett ”massivt undantag”. Anderljung skrev till mig att regleringen borde anpassas till ett systems kapacitet och användning, och sa att han ”skulle vilja ha en viss reglering av dagens mest kompetenta och mest använda modeller”, men han tror att det kommer att ”bli betydligt mer politiskt genomförbart att genomdriva krav på system som inte är utvecklade ännu”.

Inioluwa Deborah Raji dristade sig att säga, att om Tech-giganterna ”vet att de måste vara boven i någon mening … så föredrar de att det är abstrakt och på lång sikt”. Det låter för mig långt mer troligt än tanken att Big Tech faktiskt vill främja tanken på att deras produkter har en rimlig chans att *bokstavligen döda alla*.

Nästan 700 personer skrev under brevet om utrotning, de flesta av dem akademiker. Bara en av dem driver ett börsnoterat företag: OP:s grundare Moskovitz, som också är medgrundare och VD för Asana, en produktivitetsapp. Det fanns inga anställda från Amazon, Apple, IBM eller något ledande AI-hårdvaruföretag. Inga chefer från Meta skrev under.

Om Big Tech-företagens chefer vill förstärka skildringen av utrotning, varför har de då inte lagt sina

43 YouTube, januari 2023; TechPolicy, 17 maj 2023; New York Times, 20 november 2023.
44 New York Times, 16 maj 2023; Corporate Europe Observatory, 17 november 2023; Time, 22 november 2023; Politico, 20 maj 2023; Twitter, 26 maj 2023; San Fransisco Standard, 17 november 2023.

namn till listan?

Varför bygga en ”domedagsmaskin”?

Om AI verkligen räddar världen kan den som skapade den hoppas att lovprisas som en modern Julius Caesar. Och även om den inte gör det, kommer vem som än är först att bygga ”den sista uppfinningen som människan någonsin behöver göra”[45] inte att behöva oroa sig för att glömmas bort av historien – om nu inte historien plötsligt tar slut efter deras uppfinning.

Connor Leahy tror att vår nuvarande väg leder till att historiens slut kommer att följa strax efter att AGI uppstår. Med sitt böljande hår och misskötta pipskägg skulle han troligen passa ihop med en skylt där det står ”Slutet är nära” – även om det inte har hindrat honom från att bli inbjuden att tala i brittiska överhuset och CNN. Den 28-årige VD:n för Conjecture och medgrundare av det inflytelserika kollektivet för öppen källkod, EleutherAI, berättade en hel del för mig om vad motivet att bygga AI handlar om: ”Så du bygger den ultimata domedagsmaskinen som tjänar miljardtals dollar till dig och gör dig till kung över jorden eller dödar alla? Javisst, det är den maskulina drömmen. Du säger: ’Ja, för f-n, jag är domedagskungen’.” Han fortsätter: ”Jag gillar det. Det är väldigt likt Silicon Valleys estetik.”

Leahy förmedlar också något som inte kommer att överraska folk som har tillbringat någon längre tid i Bay Area eller vissa hörn av internet:

> Det finns verkliga, helt ansvarslösa, ej valda, tekno-utopiska affärsmännniskor och teknologer, till största delen boende i San Francisco, som är beredda att riskera era liv, era barn, era barnbarn och mänsklighetens hela framtid bara för att de kanske har en chans att leva för evigt.

I mars rapporterade *MIT Technology Review* att Altman ”säger att han tömde sitt bankkonto för att finansiera två … mål: obegränsad energi och förlängd livslängd.”[46]

Med tanke på allt detta kan man förvänta sig att etikgruppen skulle se säkerhetsgruppen som en naturlig allierad i en gemensam kamp för att få makt över de ansvarslösa teknologieliter som ensidigt bygger farliga och skadliga produkter. Och vi såg tidigare att många säkerhetsförespråkare har gjort inviter till AT-etikerna. Det är också sällsynt att personer i x-risk-gruppen öppet angriper AI-etiker (medan motsatsen … inte är sant), men verkligheten är att säkerhetsförespråkare ibland har varit svåra att stå ut med.

AI-etiker och de personer de förespråkar rapporterar ofta att de känner sig utstötta och avskurna från den verkliga makten, att de kämpar i motvind mot teknologiföretagen, som ser dem som ett fikonlöv snarare än som en verklig prioritering. De stora nedskärningarna av AI-etikgrupperna på många av Big Techs företag de senaste åren (eller dagarna) ger trovärdighet åt denna känsla. Och i ett antal fall har dessa företag hämnats på etikinriktade visselblåsare och fackliga organisatörer.[47]

Det innebär inte nödvändigtvis att dessa företag istället på allvar prioriterar existentiella risker. Etikgruppen på Google DeepMind, med bland andra Larry Page och den framstående forskaren av existentiella risker, Toby Ord, hade sitt första möte 2015, men höll aldrig något fler. En av Googles AI-forskare skrev till mig att de ”inte talar om långsiktiga risker … på kontoret”, och fortsatte: ”Google

45 Quote Investigator, 4 januari 2022.
46 MIT Technology Review, 8 mars 2023.
47 The Guardian, 9 0ktober 2021; New York Times, 20 april 2019.

är mer inriktat på att bygga teknologi och säkerhet i meningen laglighet och anstötlighet.”

Programvaruingenjören Timnit Gebru var med i ledningen för Googles AI-etiska grupp innan hon i slutet av 2020 tvingades bort från företaget efter en dispyt om ett artikelutkast – nu en av de mest berömda publikationerna om maskininlärning någonsin. I artikeln om ”slumpmässiga papegojor”[48] hävdar Gebru och hennes medförfattare att LLM skadar miljön, ökar de sociala fördomarna och använder statistik för att ”slumpmässigt” sätta ihop språk ”som saknar koppling till någon mening”.

Gebru är ingen beundrare av AI-säkerhetsgruppen, och har krävt ökat skydd för visselblåsare bland AI-forskarna, vilket är också är en av de viktigaste rekommendationerna i GovAI:s vitbok. Efter att Gebru drevs bort från Google har nästan 2.700 anställda skrivit under ett solidaritetsbrev, men dåvarande Google-medarbetaren Geoff Hinton var inte en av dem. När han på CNN fick frågan varför han inte stödde en annan visselblåsare, svarade Hinton att Gebrus kritik av AI grundades på ”helt andra farhågor än mina” som ”inte är så existentiellt allvarliga som tanken att de här sakerna kan bli intelligentare än oss och ta över.”

Raji sa till mig att ”en stor anledning till frustrationen och fientligheten” mellan etik- och säkerhetslägren är att ”ena sidan har så mycket mer pengar och makt än andra sidan”, vilket ”gör att de kan driva sin agenda mycket mer direkt”.

Enligt en bedömning har den mängd pengar som går till nya och icke vinstdrivande företag inom AI-säkerhet fyrdubblats mellan 2020 och 2022, och uppgår till 144 miljoner dollar. Det är svårt att hitta en motsvarande siffra för AI-etikgruppen. Men civilsamhället från båda lägren ställs i skuggan av det industrin lägger ut. Open Secrets rapporterade att det bara under första kvartalet 2023 spenderades 94 miljoner dollar på lobbyarbete för AI i USA. LobbyControl beräknade att teknologiföretagen samma år lade ut 113 miljoner euro på lobbyarbete i EU, och vi påminner om att hundratals miljarder dollar just nu investeras i AI-industrin.

En sak som kan driva på fientligheten ännu mer än upplevda skillnader i makt och pengar är trendkurvan. Efter vida hyllade böcker som dataforskaren Cathy O’Neils *Weapons of Mass Destruction* [Massförstörelsevapen] från 2016, och chockerande upptäckter av fördomar i algoritmer, som artikeln ”Genusskuggor” av Buolamwini och Gebru, hade perspektivet på AI-etik fångat allmänhetens uppmärksamhet och stöd.

2014 fick de som varnar för existentiella risker med AI sin egen överraskande bästsäljare, filosofen Nick Bostroms *Superintelligence*, som hävdade att AI över mänsklig nivå kan leda till utrotning och lovprisades av figurer som Elon Musk och Bill Gates. Men Yudkowsky berättade för mig att folk före ChatGPT hade tittat konstigt på en om man drev bokens tes utanför vissa kretsar i Silicon Valley. Tidiga förespråkare av AI-säkerhet, som Yudkowsky, har befunnit sig den märkliga ställningen att ha haft nära band till rikedom och makt via datorspecialister i Bay Area samtidigt som de var marginaliserade i den bredare diskussionen.

I världen efter ChatGPT kommer Turing- och Nobelpristagare ut ur AI-säkerhetsgarderoben och anammar argument som har populariserats av Yudkowsky, vars mest kända publikation är ett verk om Harry Potter-fanfiction på totalt mer än 660.000 ord.

Det kanske mest chockande förebudet om denna nya värld sändes i november, när värdarna för New

48 ”On the Dangers of Stochastic Parrots”.

York Times' teknologipodd *Hard Fork* frågade Konkurrensmyndighetens ordförande: ”Vad är din dom, Lina Khan? Hur sannolikt tror du att det är att AI kommer att döda oss alla?” EA-snack i kön till kaffeautomaten har alltså blivit normalt. (Khan sa att hon är ”optimist” och gav en ”låg” uppskattning på *15%*.)

Det är lätt att se alla öppna brev och mediaserier och tro att de flesta AI-forskare mobiliserar mot existentiella risker. Men när jag frågade Bengio om hur detta upplevs idag inom maskininlärningsgruppen, sa han: ”Åh, det har ändrat sig en hel del. Det brukade vara sådär 0,1% av alla som uppmärksammade frågan. Och nu är det kanske 5%.)

Sannolikheter

Liksom många andra som bekymrar sig för utrotningsrisker med AI använde filosofen David Chalmers ett sannolikhetsargument under vårt samtal: ”I en sådan här situation måste man inte vara 100% säker på att vi måste oroa oss för en AI på mänsklig nivå. Om det är 5% så är det något vi måste oroa oss för.”

Den här sortens statistiskt tänkande är populärt inom EA-gruppen och det var till stor del det som fick dess medlemmar att rikta in sig på AI från början. Om man accepterar expertargument kan man bli mer förvirrad. Men om man försöker ta genomsnittet av experters oro i den handfull undersökningar som finns, så kanske man tror att det åtminstone är några procents chans att utrotning av AI kan äga rum, vilket borde räcka för att göra det till den viktigaste saken i världen. Och om man sätter något värde på alla de framtida generationer som skulle kunna existera, så är utrotning av mänskligheten definitivt värre än katastrofer som går att överleva.

Men det finns ett överflöd av anklagelser om förmätenhet i AI-debatten. Skeptiker som Melanie Mitchell och Oren Etzioni sa till mig att det inte finns några bevis som stöder risken för utrotning, medan de som tro det, som Bengio och Leahy, pekar på de överraskande kapacitetsökningarna och frågar: Och om inte utvecklingen stannar? En vän till mig som är AI-forskare och akademiker liknade uppkomsten av AGI med att kasta den globala ekonomin och politiken i en mixer.

Även om AGI av någon anledning bara kan matcha och inte överträffa människans intelligens, så är framtidsutsikten att dela världen med ett nästan godtyckligt stort antal digitala hjälpredor på mänsklig nivå skrämmande, speciellt när de troligen kommer att försöka tjäna pengar åt någon.

Det finns alldeles för många politiska idéer om hur den existentiella risken med AI ska minskas för att diskutera det här. Men ett av de tydligaste budskapen från AI-säkerhetsgruppen är att vi borde ”sakta in”. Förespråkare av en sådan inbromsning hoppas att det ska ge beslutsfattare och samhället i stort en chans att hinna ikapp och aktivt besluta hur en möjligen omvandlande teknolog ska utvecklas och användas.

Internationellt samarbete

Ett av de vanligaste svaren på varje försök att reglera AI är invändningen ”men Kina då!” Exempelvis sa Altman till en senatskommitté i maj att ”vi vill att Amerika ska leda” och medgav att en risk med att sakta in är att ”Kina eller någon annan gör snabbare framsteg”.[49]

49 TechPolicy, 17 maj 2023.

Anderljung skrev till mig att det ”inte är ett tillräckligt starkt skäl att inte reglera AI”.

I en artikel i *Foreign Affairs* i juni rapporterade Helen Toner och två politiska vetenskapsmän att kinesiska AI-forskare som de intervjuade trodde att kinesiska LLM är minst två år efter de främsta amerikanska modellerna. Dessutom hävdar författarna, att eftersom utvecklingen av kinesisk AI ”till stor del litar till att reproducera och anpassa forskning som publiceras utomlands”, så skulle en ensidig inbromsning ”troligen bromsa” även kinesiska framsteg.[50] Som Anthropics policychef Jack Clark har observerat har Kina också agerat snabbare än något annat större land för att reglera AI på ett meningsfullt sätt.

Yudkowsky säger: ”Det är faktiskt inte i Kinas intresse att begå självmord tillsammans med resten av mänskligheten.”

Om en utvecklad AI verkligen hotar hela världen så räcker det inte med inhemsk reglering. Men hårda nationella restriktioner kan skicka trovärdiga signaler till andra länder hur allvarligt man tar på riskerna. Den framstående AI-etikern Rumman Chowdhury har uppmanat till global övervakning. Bengio säger att vi ”måste göra båda”.

Inte överraskande har Yudkowsky intagit en maximalistisk ståndpunkt, och säger till mig att ”den riktiga linjen handlar mer om att sätta all AI-hårdvara i ett begränsat antal datacentraler under internationell övervakning av organ med ett symmetriskt avtal där ingen – inklusive militären, regeringarna, Kina eller CIA – kan göra några verkligt ohyggliga saker, inklusive bygga superintelligenser.”[51]

I en kontroversiell debattartikel i *Time* i mars argumenterade Yudkowsky för att ”stänga ner allting” genom att upprätta ett internationell moratorium på ”nya stora inlärningsrundor”, med stöd av militärt hot.[52] Med tanke på Yudkowskys starka tro att en utvecklad AI vore mycket farligare än några kärnvapen eller biologiska vapen följer denna radikala ståndpunkt naturligt.

Alla 28 länderna vid toppmötet om AI-säkerhet nyligen, inklusive USA och Kina, skrev under Bletchleydeklarationen, som medger existerande skador av AI och det faktum att ”det kan uppstå betydande risker av möjligen avsiktligt missbruk eller oavsiktliga kontrollproblem relaterade till inriktning med mänskligt uppsåt.”[53]

Under toppmötet utsåg värden, den brittiska regeringen, Bengio till att leda produktionen av den första rapporten med ”senaste vetenskapliga rön” om ”förmågor och risker med ledande AI”, i ett viktigt steg mot en permanent expertgrupp likt FN:s klimatpanel.

Ett samarbete mellan USA och Kina kommer absolut att vara nödvändigt för en meningsfull internationell samordning av utvecklingen av AI. Och när det kommer till AI är de två länderna inte direkt på talefot. Med 2022 års CHIPS Act och exportkontroll försökte USA omintetgöra Kinas AI-kapacitet, något som en industrianalytiker tidigare skulle ha ansett vara en ”krigshandling”. Som *Jacobin* rapporterade i maj spelade en del forskare om existentiella risker troligen en roll för att få igenom de hårda kontrollerna. I oktober skärpte USA restriktionerna i CHIPS Act för att stänga igenom en del kryphål.

50 Foreign Affairs, 2 juni 2023.
51 Twitter, 1 november 2023.
52 Time, 29 mars 2023.
53 Bletchley Declaration, brittiska regeringen november 2023.

Men ett uppmuntrande tecken är att Biden och Xi Jinping i november diskuterade AI-säkerhet och förbud mot AI i dödliga vapensystem. Ett pressutlåtande från Vita huset sa: ”Ledarna bekräftade behovet att ta itu med riskerna med avancerade AI-system och förbättra AI-säkerheten genom regeringssamtal mellan USA och Kina.”[54]

Dödliga självstyrande vapen (LAW) är också också ett område där man är relativt överens i AI-debatten. I sin nya bok *Avslöja AI*,[55] talar Joy Buolamwini för kampanjen Stoppa mördarrobotarna, och upprepar farhågor som många förespråkare av AI-säkerhet har uttryckt länge. X-risk-organisationen Future of Life Institute samlade 2016 ideologiska motståndare för att skriva under ett öppet brev för att förbjuda LAW, inklusive Bengio, Hinton, Sutton, Etzioni, LeCun, Musk, Hawking och Noam Chomsky.

En plats vid bordet

Efter åratal av passivitet riktar världens regeringar slutligen uppmärksamheten mot AI. Men genom att inte på allvar ta itu med vad framtida system skulle kunna göra, lämnar socialister sin plats vid bordet.

Till stor del på grund av den sorts människor som drogs till AI, beslutade sig många av de tidigaste anhängarna till tanken på risk för utrotning att antingen engagera sig i extremt teoretisk forskning om hur avancerad AI skulle kontrolleras, eller så startade de AI-företag. Men för andra människor är svaret på tron att AI kan göra slut på världen att man ska försöka få *folk att sluta bygga den*.

Boostern upprepar att AI-utvecklingen är oundviklig – och om tillräckligt många människor tror det så blir det sanning. Men ”det finns ingenting med konstgjord intelligens som är oundvikligt”, skriver AI Now Institute. Verkställande direktör Myers West upprepade detta, och nämnde att teknologi för ansiktsigenkänning verkade oundviklig 2018, men har sedan dess förbjudits på många ställen. Och som forskaren i existentiella risker Katja Grace poängterar borde vi inte känna behov att bygga alla teknologier bara för att vi kan.[56]

Dessutom har många av de beslutsfattare som har tittat på AI:s senaste utveckling *flippat ut*. Senator Mitt Romney är ”mer rädd för AI” än optimistisk, och hans kollega Chris Murphy säger: ”Konsekvenserna av att så många mänskliga funktioner läggs ut på AI kan bli katastrofalt.” Kongressledamöterna Ted Lieu och Mike Johnson blir bokstavligen ”utflippade” av AI. Om vissa datorexperter är de enda som villiga att medge att AI:s kapacitet har ökat dramatiskt, och i framtiden skulle kunna utgöra ett hot mot arten, så är det på dem som beslutsfattare kommer att lyssna extra mycket. I maj twittrade professorn och AI-etikern Kristian Lum: ”Det finns en existentiell risk som jag är säker på att LLM utgör om vi fortsätter att kalla dem patentlösningar, och det är för tilltron till området FAccT [Fakta, Ansvarighet, Transparens] / Etiskt AI.”[57]

Även om tanken på en av AI pådriven utrotning slår en som mer saga än vetenskap, så kan den ändå ha ett enormt inflytande på hur en omvandlande teknologi utvecklas och vilka värden den representerar. Att anta att vi kan få en hypotetisk AGI att göra det vi vill ställer kanske den viktigaste fråga

54 White House, 15 november 2023.
55 Joy Buolamwini, *Unmasking AI : my mission to protect what is human in a world of machines*, New York: Random House 2023.
56 AI Now Institute, 11 april 2023; ACLU, 17 januari 2020; World Spirit Sock Stack, 20 december 2022.
57 Mitt Romney, 19 september 2023; Twitter, (Murphy) 27 mars 2023; Twitter, (Lum) 3 maj 2023.

mänskligheten någonsin kommer att ställas inför: vad ska vi *vilja* att den vill?

När jag frågade Chalmers om det sa han: ”Vid någon tidpunkt sammanfattar vi den politiska filosofins alla frågor: vilken sorts samhälle vill vi egentligen ha och värderar vi faktiskt?”

Ett sätt att tänka på om uppkomsten av ett AI på mänsklig nivå är att det skulle vara som att skapa en konstitution för ett nytt land (Anthropics ”konstitutionella AI” tar denna tanke bokstavligen, och företaget experimenterade nyligen med att införliva demokratisk inläsning i sin modells grundningsdokument). Regeringar är komplicerade system som utövar en enorm makt. Den grund på vilken de upprättas kan påverka miljontals människors liv nu och i framtiden. Amerikaner lever under oket från döda män som var så rädda för allmänheten att de byggde in antidemokratiska åtgärder som fortfarande plågar vårt politiska system mer än två sekel senare.

AI kanske är mer revolutionärt än någon tidigare uppfinning. Det är också en unikt normgivande teknologi, med tanke på hur mycket vi bygger den för att återspegla det vi föredrar. Som Jeff Clark nyligen funderade på i *Vox*: ”Det är verkligen skumt att det inte är ett regeringsprojekt.”[58] Chalmers sa till mig: ”När vi plötsligt har teknologiföretag som försöker bygga in dessa mål i AI-systemen, så litar vi på att teknologiföretagen lyckas hantera dessa mycket djupa sociala och politiska frågor rätt. Jag är inte så säker på det.” Han betonade: ”Det är inte bara tekniska eftertankar i detta utan också sociala och politiska.”

Falska val

Vi kanske inte behöver vänta på att hitta superintelligenta system som inte prioriterar mänskligheten. Övermänskliga ombud som skoningslöst optimerar belöning på bekostnad av allt annat vi kan bry oss om. Ju skickligare ombud och ju mer skoningslös optimerare desto extremare resultat.

Låter det bekant? I så fall du inte ensam. AI Objectives Institute (AOI) betraktar både kapitalismen och AI som exempel på feljusterade optimerare. Detta forskningslaboratorium har grundats gemensamt av tidigare radiovärden Brittney Gallagher och ”privatlivshjälten” Peter Eckersley strax innan hans oväntade död, och analyserar området mellan utplåning och utopi, ”en fortsättning av de befintliga trenderna att makten samlas i allt färre händer – superladdad av avancerad AI – snarare än en skarp brytning med det nuvarande.”[59] AOI:s ordförande Deger Turan sa till mig: ”Existentiell risk är underlåtelse att samordna sig inför en risk.” Han säger att ”vi behöver skapa broar mellan” AI-säkerhet och AI-etik.

En av de mer inflytelserika idéerna i x-risk-kretsar är ensidighetsförbannelsen, ett begrepp för situationer där en enda aktör kan förstöra saker för hela gruppen. Om exempelvis en grupp biologer hittar ett sätt att göra en sjukdom dödligare, behövs det bara en för att publicera det. Under de senaste decennierna har många personer blivit övertygade om att AI skulle kunna radera ut mänskligheten, men bara de mest ambitiösa och risktoleranta av dem har startat företag som nu flyttar fram gränsen för AI:s kapacitet, eller som Sam Altman nyligen uttryckte det, pressar ”tillbaka okunskapens slöja”. Som VD:n antyder har vi inget sätt att verkligen veta vad som finns bortom den teknologiska gränsen.

Vissa förstår riskerna fullständigt men ångar på framåt ändå. Med hjälp av toppforskare hade

58 Vox, 25 september 2023.
59 AI Objectives Institute Whitepaper, februari 2023.

ExxonMobil redan 1977 definitivt upptäckt att deras produkt orsakade global uppvärmning. Sedan ljög de om det för allmänheten samtidigt som de byggde allt högre oljeplattformar.

Tanken att brinnande kol skulle kunna värma upp klimatet framkastades första gången i slutet av 1800-talet, men det tog nästan hundra år innan en vetenskaplig samsyn på klimatförändringarna tog form. Tanken att vi skulle kunna förlora kontrollen över maskinerna för gott är äldre än datorerna, men det är fortfarande långtifrån vetenskaplig enighet om det. Och om den nuvarande utvecklingen av AI fortsätter i samma takt, så kanske vi inte har decennier på oss att bli eniga innan vi agerar på ett meningsfullt sätt.

Den debatt som utspelar sig i offentligheten kan få en att tro att vi måste välja mellan att ta itu med AI:s omedelbara skador och dess inneboende möjliga existentiella risker. Och det finns förvisso val som måste övervägas noga.

Men när man tittar på de materiella krafter som är i rörelse framträder en annan bild: i ena hörnet står biljondollarföretag som försöker göra AI-modellerna kraftfullare och lönsamma, i det andra finns civilsamhällets grupper som försöker få AI att återspegla värderingar som i allmänhet krockar med vinstmaximering.

Kort sagt är det kapitalismen mot mänskligheten.