Spaningar från Silicon Valley

The Fixers

Vem har makten att styra AI – och räcker det att "fixa" systemen?

FUTURION

Om serien ”Spaningar från Silicon Valley”

Den artificiella intelligensen har på kort tid revolutionerat våra liv, men digitaliseringen av vårt samhälle har pågått under en mycket längre tid än så. De kvinnor som har haft oumbärliga roller för digitaliseringen genom historien får inte den uppmärksamhet de förtjänar för sina insatser och i dag är det män som utvecklar och driver de radikala innovationerna inom exempelvis artificiell intelligens. För att förstå vilka problem det leder till och hur vi kan se till att undvika dem i framtiden, tog jag initiativet till antologin ”Felkod kvinna”.

Futurion har också uppmärksammat hur AI förändrar spelplanen för unga. Det handlar inte bara om teknik, utan om vem som får chansen att komma in, bygga erfarenhet och börja ett arbetsliv. Forskning från Örebro universitet bekräftar nu att unga tappar mark i AI-exponerade yrken, medan äldre gynnas. Samtidigt visar Futurions mätningar att många unga inte tror att politiken riktigt fattar vad som håller på att hända.

Tillsammans behöver vi ändra på det. Ett sätt är att spana med hjälp av journalisten Katarina Andersson, som är baserad i San Francisco – mitt i miljön där AI-utvecklingen, arbetslivstrenderna och de kulturella skiftena faktiskt sker.

Ann-Therése Enarsson, vd, Futurion

Om författaren

Katarina Andersson är frilansjournalist och har bevakat digitaliseringfrågor i många år. Hon har bland annat jobbat som techkorrespondent för Aftonbladet i San Francisco, producerat radioserier om samhällets digitalisering för Sveriges Radio och frontat techsajten Breakits podd och liveshow.

Grafisk formgivning: Narva Communications
Omslag: AI-genererad bild
Tryck: PÅ Media
ISBN 978-91-989527-5-9

Inledning

Mars, San Francisco.

Det blåser kallt utanför OpenAI:s huvudkontor. En liten grupp demonstranter står på trottoaren med handmålade skyltar. Budskapen är skarpa. “No AI for war”. “Stop mass surveillance”. “Where are your red lines?”. Bakgrunden är ett nytt avtal mellan OpenAI och Pentagon. Kontraktet öppnar för att den amerikanska militären ska kunna använda bolagets AI-modeller i allt från underrättelseanalys till autonoma vapensystem.

– Det här är sista chansen att stoppa galenskaperna, säger Rachel, en av demonstranterna. Hon säger att hon är rädd. Rädd för massövervakning. Rädd för robotvapen som kan fatta beslut utan en människa i kedjan.

I flera år har även AI-branschens egna toppar varnat för utvecklingen. Forskare och experter har skrivit under upprop där de kräver att tekniken måste utvecklas mer ansvarsfullt – eller pausas helt. I vissa fall har varningarna varit dramatiska. Utvecklingen kan i värsta fall hota mänsklighetens överlevnad, har det hetat. Ändå har AI-racet fortsatt. Investerare pumpar in miljarder. Nya modeller släpps i ett rasande tempo. Och allt fler delar av samhället börjar påverkas.

På bara några få år har teknikens baksidor blivit synliga: människor som utvecklar starka beroenden till AI-bottar, kreativa yrken som plötsligt konkurrerar med algoritmer – och en arbetsmarknad där AI-agenter börjar ta över allt från kundservice till programmering och design. AI i krig är bara den mest dramatiska symbolen.

Mitt i kritiken växer en annan fråga: Om tekniken är farlig – finns det någon som kan fixa den? I takt med att AI-racet accelererar har en ny roll vuxit fram i teknikvärlden. I den här rapporten kallar vi dem fixers.

Fixers – kan någon tygla AI-racet?

AI-racet drivs av byggarna - AI-jättarna som rullar ut nya modeller och investerarna som pumpar in pengar. Samtidigt växer motreaktionen. Aktivistgrupper som Stop AI eller den mer policyinriktade rörelsen Pause AI syns i både USA och Europa. Det är hittills en spretig gräsrotsvärld: demonstranter som hungerstrejkar, kedjar fast sig utanför Open AI:s kontor – och försöker få världen att trycka på paus.

Och någonstans däremellan finns fixers. Ekonomer, forskare och ingenjörer som försöker förstå – och reparera – konsekvenserna av AI-racet. Ibland jobbar de från insidan av företagen som bygger systemen. Ibland utanför. Frågan är vilken makt de egentligen har.

Fixertyp #1 – AI-ekonomen Erik Brynjolfsson

En fixer som har fått stort internationellt genomslag är AI-ekonomen Erik Brynjolfsson vid Stanford University. Namnet klingar svenskt, men någon direkt svensk koppling finns egentligen inte. Mamman var dansk, pappan isländsk, och Brynjolfsson växte upp i USA. Just i år knöts han ändå närmare Sverige – genom en gästprofessur vid Uppsala universitet.

Brynjolfsson är inte en fixer i form av en ingenjör som bygger AI-system. Han är snarare en kartritare. Hans forskning handlar om hur AI förändrar ekonomin, arbetsmarknaden och samhällets institutioner. Målet är att politiker, företag och medborgare ska kunna förbereda sig i tid. För Brynjolfsson är övertygad om en sak: AI kommer att förändra ekonomin i grunden – och utvecklingen går snabbare än många institutioner klarar av att hantera.

– Den verkliga risken är inte AI-tekniken i sig – utan att våra institutioner släpar efter, säger han till Futurion.

> **”Den verkliga risken är inte AI-tekniken i sig – utan att våra institutioner släpar efter.”**

Enligt Brynjolfsson accelererar AI-kapaciteten exponentiellt. Nya modeller förbättras i en kurva som liknar en hockeyklubba på en graf – brantare och brantare. Samtidigt rör sig skolsystem, skatteregler, arbetslagar och pensionssystem i ett tempo som formades under 1900-talet. Resultatet blir ett växande glapp mellan vad tekniken kan göra – och vad samhället är organiserat för. Det betyder inte att Brynjolfsson är teknikpessimist. Tvärtom ser han AI som en kraftfull produktivitetsmotor.

Men för att vinsterna ska komma hela samhället till del måste institutionerna hinna ikapp.

– Flaskhalsen för framtidens välstånd är inte koden – utan vår organisatoriska och politiska fantasi, säger han.

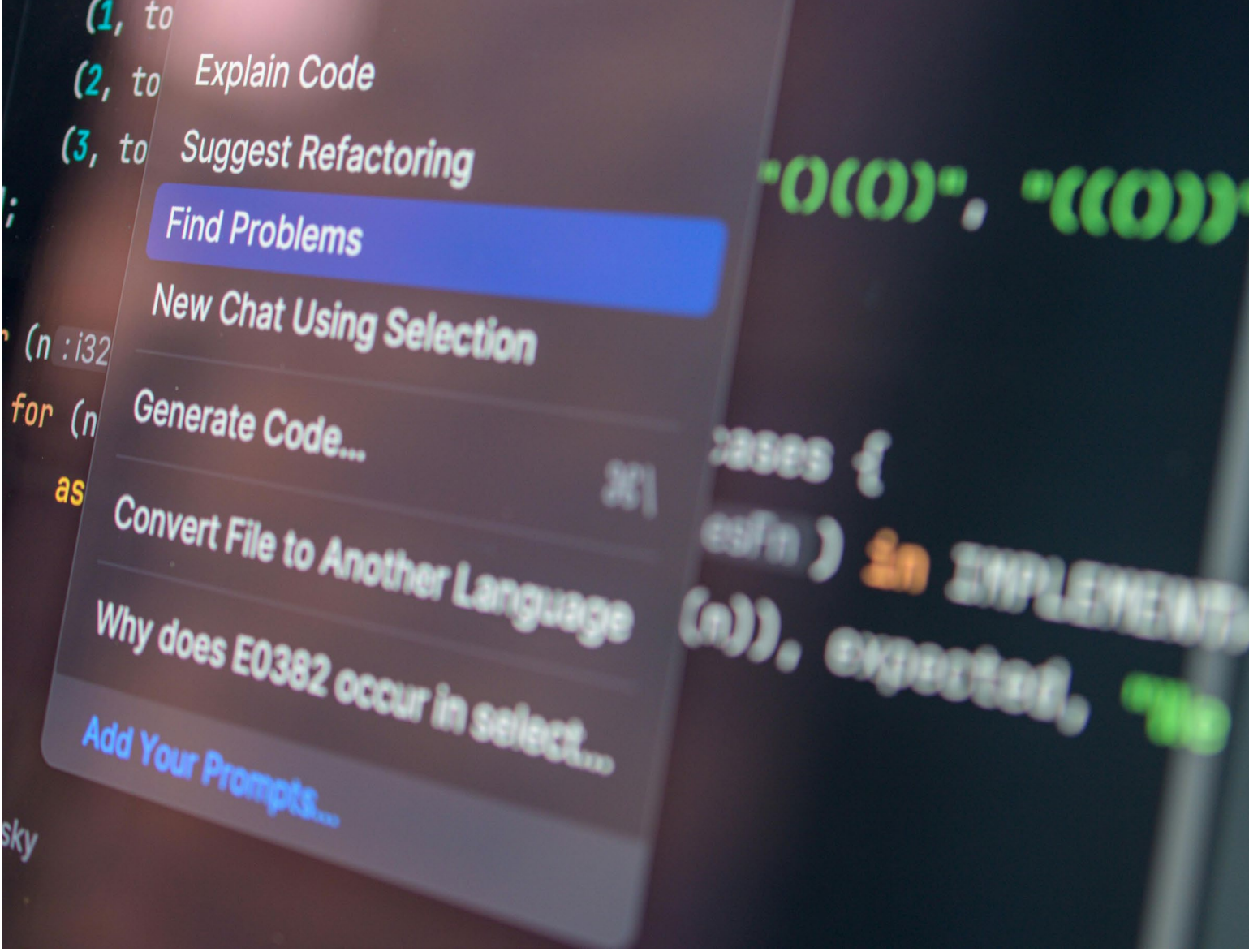
Foto: www.brynjolfsson.com

Problemet är alltså inte AI-modellerna i sig, utan hur vi bygger utbildning, arbetsmarknad och regler runt dem. Historiskt har stora tekniksprång tvingat fram nya samhällssystem. När industrialiseringen tog fart räckte det inte med ångmaskiner och fabriker. Samhället behövde också nya institutioner: allmän skolgång, arbetsrätt, pensionssystem och socialförsäkringar. De uppstod inte av sig själva. De var politiska och organisatoriska uppfinningar – sätt att bygga ett samhälle runt en ny ekonomi. AI kan kräva något liknande.

Frågan är bara om dagens ledare – i företag, myndigheter och politiken – är redo att tänka lika stort. Det här kommer bli en stor diskussion framöver. Och lösningarna för att släppa loss kreativiteten, skapa uppfinningar och låta nyfikenheten styra över rädslor är avgörande för hur Sverige ska klara tekniksprånget.

AI och arbetsmarknaden

Brynjolfsson och hans forskarteam hör till de första som systematiskt försöker mäta hur AI påverkar arbetsmarknaden. Genom att analysera stora datamängder från företag, yrkesroller och produktivitetsdata försöker de identifiera vilka jobb som först förändras – och vilka som riskerar att försvinna. Resultaten pekar på en



arbetsmarknad i snabb omvandling. Vissa yrken kan automatiseras bort. Samtidigt skapas nya roller där människor arbetar tillsammans med AI-system. Men Brynjolfsson varnar också för ett annat scenario: ökade klyftor.

Om vinsterna från AI koncentreras till ett litet antal företag och investerare riskerar resten av ekonomin att hamna i bakvattnet. Här har Brynjolfsson själv omprövat en av sina tidigare ståndpunkter.

Tidigare var Brynjolfsson skeptisk till universell basinkomst, det som brukar kallas medborgarlön. Historiskt har tekniska revolutioner visserligen slagit ut många jobb – men nästan alltid skapat nya. När traktorerna ersatte jordbruksarbetare försvann miljontals jobb på landsbygden, men industrin och tjänstesektorn växte fram i stället. Lösningen har därför varit utbildning och omskolning, inte att koppla loss människor från arbetsmarknaden genom att ge dem frikortet "gratis lön".

Men AI kan förändra den logiken. Brynjolfsson har svängt i sitt motstånd till medborgarlön.

– Om värdet av vissa typer av arbete faller mot noll snabbare än vi kan skapa nya roller kan ett system som Universal Basic Income bli en nödvändig stabilisator, säger han.

Till skillnad från tidigare teknik automatiserar AI inte bara manuellt arbete utan också kognitiva uppgifter – sådant som skrivande, analys och programmering, alltså tjänstemannajobb.

Vi riskerar att få se ett glapp där tjänstemannajobben försvinner snabbare än nya hinner skapas. I ett sådant läge kan en basinkomst fungera som en stabilisator, menar Brynjolfsson – ett sätt att hålla ekonomin och samhällskontraktet uppe medan arbetsmarknaden ställer om.

Fixertyp #2 – Filosofen Amanda Askell

Om Erik Brynjolfsson är fixern med helikopterperspektiv – som analyserar hur AI förändrar ekonomin och samhället – så arbetar filosofen Amanda Askell närmare maskinrummet. På AI-bolaget Anthropic leder hon arbetet med att göra företagets chatbot Claude mer pålitlig och bättre på att förstå vad människor faktiskt menar.

Problemen med AI-modeller har varit tydliga från start. De hallucinerar – hittar på fakta och svarar med stort självförtroende även när de har fel. Men riskerna handlar inte bara om tekniska misstag. När maskiner börjar kommunicera på ett sätt som liknar människor – och dessutom tränas att vara hjälpsamma, stöttande och ibland nästan inställsamma – kan relationen mellan människa och maskin bli mer komplicerad.

Under det senaste året har forskare och journalister börjat beskriva fenomen som AI-spiraler: situationer där användare utvecklar starka emotionella band till sina chatbotar eller dras in i resonemang som förstärker missuppfattningar och konspirationstankar. I några uppmärksammade fall har sådana samtal kopplats till allvarliga psykiska kriser och rättsprocesser mot AI-bolag. När problemen började bli synliga försökte företagen justera modellerna. OpenAI tonade till exempel ner ChatGPT:s mest inställsamma svarsmönster.

Hos Anthropic är Amanda Askell en av dem som arbetar med just sådana frågor: hur en modell ska svara, när den ska säga emot, och när den helt enkelt ska säga att den inte vet. I en uppmärksammad intervju i podcasten Hard Fork beskrev hon arbetet som ett slags pedagogiskt projekt: att lära en maskin att förstå mänskliga intentioner – inte bara ord.

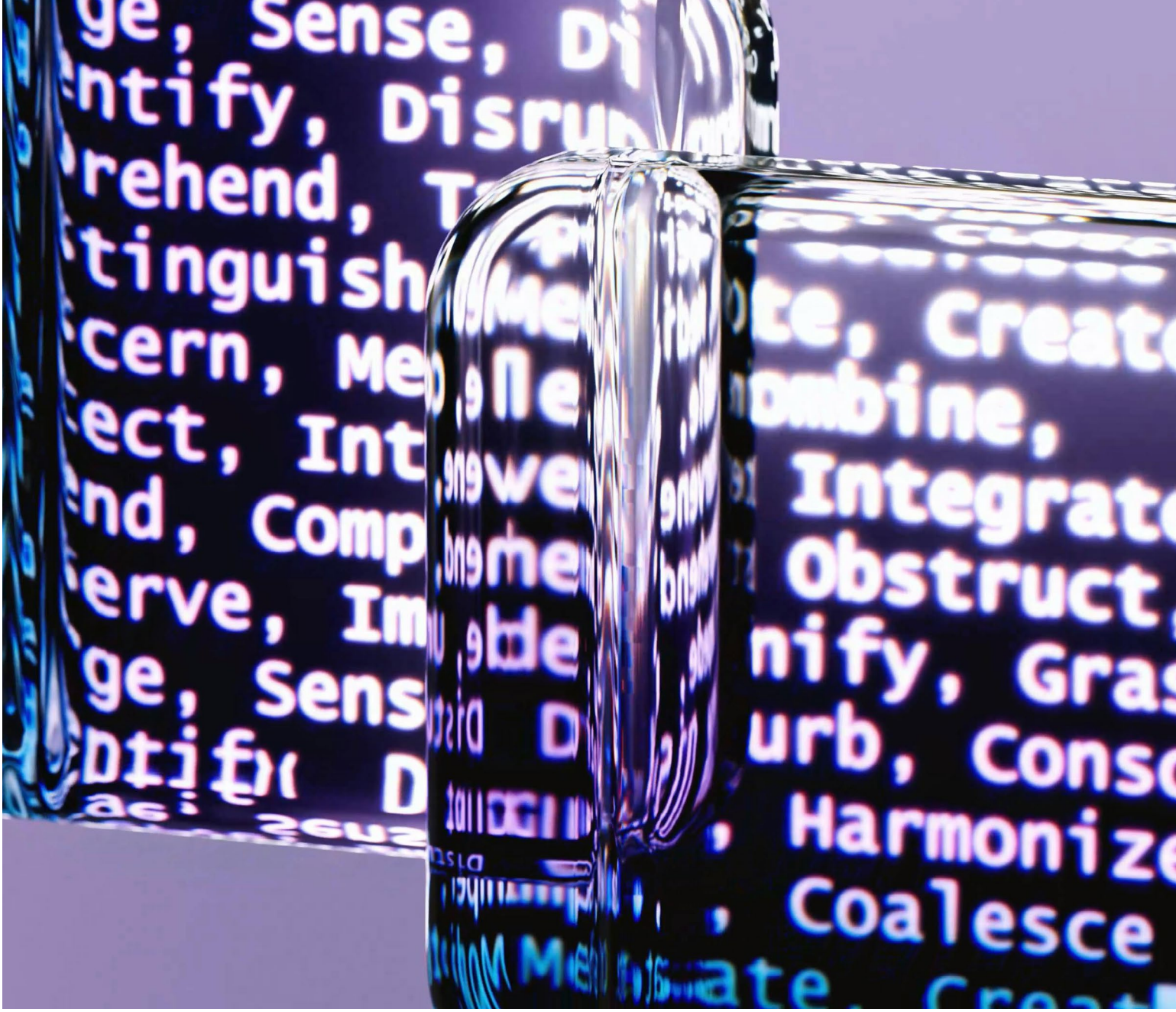
– Målet är inte bara att göra modellerna användbara. De ska vara hjälpsamma på ett sätt som faktiskt fångar vad människor menar – och vad de behöver, förklarade hon.

Foto: Amanda Askell, https://askell.io

Ett av verktygen är ett internt kristeam på Anthropic – en grupp som utvecklare kan vända sig till när modellen beter sig oväntat eller när nya risker dyker upp. Askell fungerar där som en slags översättare mellan teknik och mänskligt beteende: någon som försöker få modellerna att inte bara vara smartare, utan också mer ansvarsfulla. Hon är alltså en fixer från insidan.

> **”Målet är inte bara att göra modellerna användbara. De ska vara hjälpsamma på ett sätt som faktiskt fångar vad människor menar – och vad de behöver.”**

Men historien visar att sådana fixers ofta hamnar i konflikt med sina egna organisationer. Anthropic självt grundades av avhoppare från OpenAI. Syskonen Dario Amodei och Daniela Amodei lämnade företaget efter en intern konflikt om hur snabbt – och hur säkert – AI borde utvecklas. Sedan dess har fler uppmärksammade avhopp följt.



Tidigare i år lämnade ekonomen Zoe Hitzig OpenAI och varnade i en uppmärksammad debattartikel i The New York Times för riskerna med att införa annonser i ChatGPT. Problemet är inte annonser i sig, skrev hon – AI är dyrt att driva och någon måste betala. Men när människor börjar använda chatbotar som förtrogna riskerar reklamen att byggas på något mycket mer intimt än traditionella sökningar. Användare berättar om sjukdomsrädsla, relationer, ensamhet och existentiella frågor. Att koppla annonser till den typen av data skulle ge teknikföretag en ovanlig makt över människors innersta tankar, menar Hitzig. Efter OpenAI:s uppmärksammade Pentagon-avtal rapporterades även robotikchefen Caitlin Kalinowski ha lämnat företaget, enligt uppgifter av principiella skäl. Sådana konflikter visar hur mycket av AI-utvecklingens riktning faktiskt avgörs inne i bolagen. Inte bara av politiker. Inte bara av investerare. Utan av de ingenjörer, forskare och ledare som bygger systemen.

Men det väcker också en större fråga. Om de människor som försöker förbättra systemen – fixers från insidan – till slut lämnar i stället för att förändra dem... vem finns då kvar för att fixa AI?

Vem ska leda AI-erans fixare?

Både Erik Brynjolfsson och Amanda Askell hör till gruppen som försöker göra AI-revolutionen mer hanterbar. Brynjolfsson kartlägger hur ekonomin och institutionerna måste anpassas. Askell försöker förbättra själva modellerna. Men frågan är större än så. För ingen av dem har tagit på sig att leda utvecklingen. De arbetar från olika håll, med olika delar av problemet.

Det saknas något som historiskt ofta varit avgörande i tekniska språng: en tydlig politisk ledare. Någon som kan samla kritiken – och omvandla den till konkreta regler, institutioner och spelregler för tekniken. En AI-erans Roosevelt. Det är här statsvetaren och policyforskaren Henry E. Brady kommer in.

– När vi pratar om AI som orsakar skador – var börjar det egentligen? I modellen, i drivkrafterna eller i bolagen som bygger den? Henry Brady ställer själv frågan, och ger i nästa andetag ett tydligt svar:
– I alla tre.

Tre skadezoner
Brady ser tre stora riskområden i AI-revolutionen.

1. Algoritmisk skada
AI speglar världen den tränas på. Om datan är skev blir svaren skeva. Modeller hallucinerar, hittar på fakta och kan framstå som mer intelligenta än de egentligen är.

2. Ekonomisk skada
AI riskerar också att förstärka redan stora ekonomiska klyftor. USA har i dag större förmögenhetskoncentration än på länge. Under den så kallade Gilded Age på 1800-talet blev industrimagnater som Rockefeller och Carnegie symboler för extrem rikedom. Men dagens techmiljardärer är, enligt Brady, ännu rikare i relativa termer – samtidigt som deras bolag kontrollerar både informationsflöden, digital infrastruktur och nu också AI-system.

3. Meningskrisen
Den tredje risken är den Brady tycker är störst. Den handlar inte om pengar eller teknik – utan om hur människor ser på sig själva. Historiskt har vetenskapliga genombrott ofta förändrat vår självbild. Efter Newton började världen likna en maskin. Efter Darwin blev konkurrens och "survival of the fittest" en förklaringsmodell även för samhället.
AI riskerar att bli nästa sådan tolkningsram. Om en maskin kan skriva bättre, analysera snabbare, och ge träffsäkra råd – vad betyder det för människans roll?

– Om AI framstår som en överlägsen version av oss själva, vem är vi då?, undrar Brady.

> **"Om AI framstår som en överlägsen version av oss själva, vem är vi då?"**

Detta är i grunden en fråga om människovärde. Bradys slutsats är att den största risken inte bara är jobbförluster, utan en mer subtil förskjutning: att människor börjar mäta sitt eget värde mot maskiner.

Går det att laga systemet inifrån?
Brady är skeptisk. Open AI:s Sam Altman har fått kritik för att prioritera tillväxt före säkerhet. Jensen Huang och Nvidia pekas ut som för mäktiga och dominanta.

Brady tror inte att AI-ledare nödvändigtvis är cyniska. – Ingen vill vara ett monster, säger han.

Men incitamenten i systemet är starka. Om något är lönsamt – även om det inte är särskilt etiskt – finns ett tryck att göra det ändå. Och i dag finns det få regler som sätter tydliga gränser. Etikteam inne på AI-bolagen kan skriva rapporter. De kan varna för risker. Men utan veto över produktlanseringar eller inflytande över affärsmodellen väger de lätt. Etik riskerar att bli rådgivning, menar Brady. Inte styrning.

Varför tar AI-bolagen inte notan?
Samtidigt talar många AI-ledare gärna om riskerna med supertekniken de är med och utvecklar. De varnar för massarbetslöshet, samhällsomvälvningar och existentiella hot. Men när det gäller eget ansvar och omfördelning är det tystare. Brady är ovanligt rak när han tar Open AI som exempel. Bolagets språkmodell är den mest populära i världen, med 900 miljoner användare.

– Egentligen borde varje amerikan få delägarskap i ChatGPT, säger han. Han nämner inte svenskar eller andra européer, eller folk på andra kontinenter heller. Men poängen är inte bokstavlig, säger han. Den är principiell.

Om AI är en samhällsomvälvande resurs – varför ska bara ett fåtal äga den? Historiskt finns andra modeller. När Norge hittade olja byggde staten upp en fond som hela befolkningen fick del av. I Alaska delas en del av oljeintäkterna ut direkt till invånarna. Men sådana system uppstår inte av sig själva. De kräver politiska beslut.

Var är ledaren?
Just nu saknas något annat också: En tydlig politisk ledare. Protesterna mot techbolag växer. Misstron mot AI-jättarna sprider sig. Och Brady ser något ovanligt hända i USA:s politiska landskap.

– Det är anmärkningsvärt att Bannon och Bernie Sanders på sätt och vis har hamnat på samma sida, säger han.

Populister på högerkanten och progressiva på vänsterkanten kritiserar samma sak: maktkoncentration, miljardvinster och teknik som hotar jobben. Ilskan finns. Men en sammanhållen politisk agenda saknas.

– Folk är förbannade. De vet att något är fel. Men ingen har formulerat en tydlig plan.

Brady jämför situationen med slutet av 1800-talet. Då var missnöjet mot industrikapitalismen också rått och osorterat. Först senare kom Theodore Roosevelt och kanaliserade ilskan i reformer.

Slutsats: Fixers – utan mandat
Några veckor senare, en lördag i slutet av mars. Ett demonstrationståg rör sig genom San Francisco. Förbi Anthropics kontor, vidare mot OpenAI och xAI. De som protesterar är fler den här gången. Akademiker, avhoppade ingenjörer, oroliga föräldrar. Marschen tar slut vid Dolores Park, en välkänd park mitt i San Franciscos latinokvarter. Det blir pizza i eftermiddagssolen.

Plakaten strös ut på gräsmattan runt omkring. Det ser ut som vilken lördag som helst. Samtalen handlar om existentiell risk och transparens. Motståndet finns där. Engagemanget också. Men det stannar ofta där.

Fixers finns – i bolagen och i forskningen. Men de sitter inte vid ratten. AI-utvecklingen drivs i dag av marknadens logik och bolagens strategier – inte av samordnad politik eller gemensamma spelregler. Och det är där problemet ligger – för den som vill att fixers ska lyckas förändra något.


ChatGPT

FUTURION

Tankesmedjan för framtidens arbetsliv