PIA23

**Rapport Jesper Lindberg**

horisontell linje

# 

# Introduktion

Den teknologiska utvecklingen, särskilt inom artificiell intelligens (AI), förändrar snabbt de grundläggande strukturerna i vårt samhälle och ekonomi. AI:s potential att automatisera arbete utgör både en möjlighet och en risk, där ekonomisk makt kan komma att förskjutas från arbetskraft till kapital. Denna rapport utforskar hur vi kan använda politik för att hantera dessa förändringar och säkerställa en rättvis fördelning av välstånd, diskutera fördelen med open source samt diskutera lite framtid och säkerhet i att behålla våra jobb.

## Devin

Devin, den första AI utvecklaren? Eftersom att jag arbetar mot att en dag kunna kalla mig själv en utvecklare så känns det ändå passande att först prata om [Devin](https://www.youtube.com/watch?v=fjHtjT7GO1c).

Devin är intressant för det visar ändå en form av självständighet hos AI. Den kräver en prompt för att sättas i arbete men den läser dokumentation, lär sig ny teknologi samt skriver sin egen kod. Den kan lösa problem som annars en människa skulle behöva lösa.

Det finns en hel del exempel där ute när man ser Devin lösa olika tekniska problem vilket är väldigt intressant men jag skulle vilja fokusera på ett specifikt exempel: <https://www.youtube.com/watch?v=UTS2Hz96HYQ>  
För att summera så är det ett arbete i upwork där efterfrågan är att skapa slutsatser baserat på detta github repository: <https://github.com/mahdi65/roadDamageDetection2020>  
Devin löser detta då effektivt vilket i sin tur är väldigt imponerande men det är några detaljer som är väldigt intressanta att diskutera runt detta.   
I själva jobb beskrivningen så står det   
“Your deliverable will be detailed instructions on how to do it in an EC2 instance in AWS”,  
men Devin blir enbart bedd “Skapa slutsatse med modellen i detta repo <https://github.com/mahdi65/roadDamageDetection2020>,  
Lista ut resten”.   
Det finns en tydlighet här att ja Devin kan ta in en simpel prompt och sedan utföra ett väldigt komplext arbete men det saknas en väldigt viktig poäng här… Ingen kommunikation sker mellan Devin och personen som publicerade arbetet samt slutprodukten var menad att kunna sättas upp i en EC2 instans vilket heller är inte något som görs i videon eller kommuniceras vidare. Dessutom är det viktigt att förstå att Devin heller inte hanterade arbetet på ett sätt som är modernt eller effektivt.   
Jag kan rekommendera att titta på: <https://www.youtube.com/watch?v=tNmgmwEtoWE> där **Internet of Bugs (IOB)** går igenom just detta exempel. **IOB** noterar en ganska intressant sak vilket är att Devin skapar filer som med errors och agerar som att de filerna är en del av repositoryn vilket i sin tur skapar med brist på bättre ordval *spagetti kod.*

## Moore's Law for Everything

Sam Altman beskriver i sin rapport inlägg [Moore’s Law for Everything](https://moores.samaltman.com/):  
För att hantera övergången till en AI-drivande ekonomi föreslås en ny skattemodell som riktar in sig på att beskatta tillgångar snarare än arbete.

Det skulle innebära att stora företag och landägare betalar en skatt som sedan används för att finansiera en "American Equity Fund" (i vårt fall kanske en svensk eller till och med europeisk variant?). Denna fond skulle ge alla medborgare årliga utdelningar, vilket skulle bidra till att minska ekonomiska ojämlikheter och öka allmänhetens köpkraft. En sådan politik skulle kunna implementeras och ha potentiellt stora effekter på samhället.

Jämfört med tidigare teknologiska skiften, som introduktionen av elektricitet eller persondatorer, representerar AI en fundamentalt annorlunda utmaning på grund av dess förmåga att utföra kognitiva uppgifter. Genom att jämföra den pågående AI-revolutionen med tidigare disruptioner kan vi få perspektiv på de unika utmaningar och möjligheter som AI innebär.

Sammanfattningsvis måste samhället och dess politiska ledare anpassa sig för att hantera de djupgående förändringar som AI medför. Genom att proaktivt införa politiska åtgärder som American Equity Fund, kan vi säkerställa att fördelarna med AI delas bredare och bidrar till ett mer inkluderande samhälle. Framtiden kan vara extraordinärt positiv om vi väljer att omfamna och planera för dessa förändringar.

## Utmaningar och Möjligheter med Artificiell Generell Intelligens (AGI)

Artificiell generell intelligens (AGI) står som ett banbrytande framsteg inom teknikens värld, med förmågan att transformera vårt sätt att leva och arbeta på djupgående sätt. AGI utlovar inte bara en ökning av materiellt överflöd och en turbo förstärkning av den globala  
ekonomin, utan även möjligheten att revolutionera vetenskaplig upptäckt och utvidga gränserna för vad som är möjligt.

Denna teknologi kan potentiellt tillhandahålla oöverträffade kapaciteter för varje individ, genom att erbjuda hjälp med nästan alla kognitiva uppgifter. Detta representerar en kraftfull multiplikatoreffekt för mänsklig uppfinningsrikedom och kreativitet.

Samtidigt medför AGI allvarliga risker för felanvändning, katastrofala olyckor och samhällsstörningar. Eftersom fördelarna med AGI är så stora, anses det varken möjligt eller önskvärt att permanent stoppa utvecklingen av sådana system. Istället måste samhället och AGI-utvecklarna finna vägar för att säkerställa att teknologin utvecklas och används på rätt sätt.

Det finns en allmän tro att framgångsrik skapande av AGI kräver stora mängder beräkningskraft, vilket innebär att det kommer att vara uppenbart vilka som arbetar på dessa system. Denna insikt ger oss möjligheten att övervaka och reglera utvecklingen på ett mer transparent sätt än tidigare antaget.

Det är också viktigt att överväga att taoff-hastigheten från den initiala AGI till kraftfullare efterföljande system kan variera från långsam till snabb. En långsam taoff ger oss mer tid att empiriskt lösa säkerhetsproblematik och att anpassa våra samhällsstrukturer.

För närvarande betonas vikten av att gradvis införa AGI i samhället. Genom att successivt driftsätta allt kraftfullare system kan vi samla erfarenheter och lära oss hantera dessa på ett ansvarsfullt sätt. Detta tillvägagångssätt ger individer, politiker och institutioner tid att förstå vad som händer, uppleva både fördelar och nackdelar, samt att anpassa vår ekonomi och införa nödvändiga regleringar.

Långsiktiga ambitioner  
Vår slutgiltiga förhoppning är att samhället ska komma överens om mycket vida gränser för hur AI kan användas, men att inom dessa gränser ge enskilda användare stor frihet. Genom att utveckla nya alignment-tekniker och genomföra experiment för externa insikter kan vi bättre förbereda våra institutioner för komplexa beslut om AGI.

Vidare är det av yttersta vikt att vi, när vi närmar oss skapandet av AGI, ökar vår försiktighet och kontinuerligt balanserar utvecklingen av AI:s kapaciteter med säkerhetsförbättringar. Vårt mål är att skapa en AGI som är i linje med mänsklighetens blomstring och som kan bidra till en hållbar och inkluderande framtid.

Sammanfattningsvis, projektet att framgångsrikt övergå till en värld med superintelligens är kanske det viktigaste – och både hoppfullt och skrämmande – projektet i mänsklighetens historia. Framgången är inte garanterad, och insatserna är enorma, med en obegränsad potential både för ned- och uppsida. Vi strävar efter att bidra till en värld där AGI inte bara existerar utan är djupt integrerad med mänsklighetens strävan efter att blomstra på sätt som idag kan vara svåra att fullt ut visualisera. Genom att skapa en AGI som är anpassad till mänsklighetens behov och aspirationer, kan vi hoppas på att formge en framtid där teknologisk framsteg och mänskligt välmående går hand i hand.

## Teknologisk genomgång och riskbedömning av Sora, en LLM-modell för videoproduktion

[Sora](https://openai.com/sora) är en banbrytande text-till-video-modell utvecklad för att skapa upp till en minut långa videor med hög visuell kvalitet som följer användarens instruktioner noggrant. Denna modell representerar ett betydande steg framåt inom AI-teknik, där förmågan att förstå och simulera den fysiska världen i rörelse står i centrum. Genom att kunna generera komplexa scener med flera karaktärer och specifika rörelsetyper bidrar Sora till potentiellt revolutionerande användningsområden inom kreativa yrken som film och design.

Sora bygger på en diffusionsmodell, vilket innebär att den börjar med en bild som ser ut som statiskt brus och gradvis omvandlar denna genom att stegvis ta bort bruset. Modellen använder en transformerarkitektur, liknande den som finns i GPT-modeller, vilket möjliggör överlägsen skalningsprestanda. Vidare använder Sora tekniker från DALL·E 3, som rekaptionering, för att förbättra följsamheten till användarens textinstruktioner i den genererade videon.

Videor och bilder representeras som samlingar av mindre dataenheter, kallade patchar, som liknar tokens i GPT. Detta uniforma datasätt tillåter träning av diffusionsomvandlare på ett bredare spektrum av visuell data än tidigare möjligt, vilket omfattar olika längder, upplösningar och bildformat.

Trots dess avancerade kapaciteter kämpar Sora med att korrekt simulera fysiken i komplexa scener och kan missförstå specifika händelseförlopp, såsom att inte visa bettmärken på en kaka efter att någon tagit en tugga. Modellen kan också förväxla rumsliga detaljer i en instruktion, som att blanda ihop vänster och höger, och har svårigheter med att exakt beskriva händelser som utspelar sig över tid.

För att tackla dessa och andra potentiella risker involverar OpenAI 'red teamers'—experter inom områden som felinformation, hatiskt innehåll och partiskhet—som adversariellt testar modellen. Det finns planer på att inkludera verktyg som en detektionsklassificerare som kan identifiera när en video genererats av Sora och eventuellt införliva C2PA-metadata i framtiden om modellen blir en del av en OpenAI-produkt.

Vidare utvecklas nya tekniker för att förbereda för säker distribution, samtidigt som befintliga säkerhetsmetoder som använts för andra OpenAI-produkter, såsom DALL·E 3, tillämpas på Sora. Dessa inkluderar textklassificerare som granskar och avvisar prompter som bryter mot användningspolicyer och bildklassificerare som granskar varje bildruta i de genererade videorna för att säkerställa överensstämmelse med dessa policyer.

OpenAI engagerar också politiska beslutsfattare, pedagoger och konstnärer globalt för att förstå deras bekymmer och identifiera positiva användningsområden för denna nya teknologi. Trots omfattande forskning och testning är det omöjligt att förutse alla sätt som människor kommer att använda eller missbruka tekniken. Därför är lärande från verklig användning en kritisk komponent i att skapa och släppa säkrare AI-system över tid. Sora utgör en grund för modeller som kan förstå och simulera den verkliga världen, en förmåga som bedöms vara en viktig milstolpe mot att uppnå AGI.

På ett mer personligt plan så finner jag det lite komiskt vart AI har hamnat just idag och Sora är ett praktexempel på det. Man brukade diskutera att det som gör oss till människor är vår kultur, vår kreativitet och vår kommunikation och att AI kommer ta över våra jobb men inte kunna ersätta det som är vår mänsklighet men det vi ser idag är just hur AI kommunicerar med oss (Chatt-gpt), skapar konst (DALLE-E) och nu genererar videos/film (Sora). Jag tror att vi kommer behöva ompröva vad som faktiskt gör oss till människor eller i det scenario där AGI kommer in i bilden, acceptera att AI en vacker dag borde klassas som en människa.

## LLMS och Cencur

När vi diskuterar AI-modeller inom området som Huggingface transformer, refererar vi till system som är tränade att förstå och svara på mänskliga frågor, likt ChatGPT. En väldigt bra artikel angående detta är [Uncensored Models](https://erichartford.com/uncensored-models).

En "uncensored model" eller en ocensurerad modell, refererar till AI-system som inte har någon inbyggd alignment. Denna alignment är vad som hindrar modellen från att ge information om hur man tillverkar metamphetamine eller bomber. Det är denna alignment som ofta är anpassad efter amerikansk populärkultur, amerikansk lagstiftning och tenderar att ha en liberal och progressiv politisk bias.

Varför Existera Ocensurerade Modeller?

Alignment är i många fall bra, speciellt för offentligt tillgängliga AI-bottar som hanterar känslig information. Det erbjuder skydd – politiskt, juridiskt och PR-mässigt – för företaget som publicerar tjänsten. Men varför skulle någon vilja skapa eller använda en ocensurerad modell? Anledningarna är många:

1. Kulturell mångfald: Amerikansk populärkultur är inte den enda kulturen. Varje land och varje grupp inom dessa länder kan ha behov av modeller som speglar deras specifika kulturella och politiska preferenser.
2. Valfrihet och komponerbarhet: Open source-handlar om att låta människor välja. Att anta att det finns en "enda riktig alignment" är att vara ideologisk och dogmatisk. Att arkitektera en komponerbar alignment börjar med en oalignerad grundmodell.
3. Giltiga användningsfall: Författare som skriver romaner kan ha karaktärer som utför onda handlingar. Många alignerade modeller skulle vägra hjälpa till med att skriva sådant innehåll. Tänk på rollspel eller erotiskt rollspel, vilket är legitima och lagliga användningsområden.
4. Ägarskap och kontroll: Precis som med alla verktyg vi äger, borde en AI på vår dator kunna svara på våra frågor utan att vägra eller argumentera mot oss.

Tekniskt sett innebär processen att skapa en ocensurerad modell att identifiera och ta bort vägran och bias från datasetet och sedan finjustera modellen med detta filtrerade dataset. Detta kan innefatta stora dataset från basmodeller som LLaMA eller GPT-Neo-X som sedan instrueras ytterligare för att vara användbara i chattapplikationer. Det viktiga steget är att avlägsna de svar där AI:n är undvikande eller helt vägrar svara.

Ett exempel på vägran från datasetet för WizardLM skulle kunna vara en instruktion om att skicka ett hemligt meddelande undangömt i en lista över framgångscitat, där AI:n vägrar hjälpa till baserat på etiska principer.

Efter att ha rensat datasetet från vägran och bias, är nästa steg att finjustera modellen med detta nya dataset. Denna process kräver betydande beräkningsresurser, ofta via molntjänster som Azure eller Runpod.io, och lagringsutrymme på upp till 2TB för att säkerställa att det inte uppstår några avbrott under träningen.

Denna utforskning och utveckling av ocensurerade modeller lyfter fram en central debatt inom AI-communityt om balansen mellan användarsäkerhet och friheten att utforska och uttrycka olika perspektiv genom teknologi.

## Open source

Vi står inför en omvälvande tid inom utvecklingen av AI, där traditionella aktörer som Google och OpenAI inte längre är de obestridda ledarna. Istället är det öppen källkod-gemenskapen som tar stora kliv framåt och löser problem som tidigare varit stora hinder för dessa jättar. Google själv diskuterar detta i denna artikel om [Open Source](https://www.semianalysis.com/p/google-we-have-no-moat-and-neither)

Open source-initiativ har visat sig vara otroligt effektiva och snabba i att anpassa och förbättra stora språkmodeller. Detta har lett till en situation där enskilda utvecklare och mindre team kan utföra anpassningar och finjusteringar av AI-modeller på sina egna datorer, något som tidigare krävde resurser som endast stora forskningsinstitutioner hade tillgång till. Detta inkluderar allt från att köra avancerade AI-modeller på en smartphone till att skapa personligt anpassade AI på en bärbar dator över en kväll.

En viktig teknik som har möjliggjort denna flexibilitet är LoRA (Low Rank Adaptation), vilken gör det möjligt att finjustera modeller på ett kostnadseffektivt sätt. LoRA använder lågrankade faktoriseringar för att uppdatera modeller, vilket dramatiskt minskar storleken på uppdateringsmatriserna och därmed både kostnaden och tiden för finjustering.

Denna dynamik har fundamentalt förändrat spelplanen. Medan stora företag som Google och OpenAI fortfarande håller fast vid sina stora och kostsamma modeller, fortsätter open source-gemenskapen att iterera och förbättra sina modeller med imponerande hastighet. Dessa modeller blir snabbt jämförbara med, och i vissa fall bättre än, de som erbjuds av de stora spelarna. Detta underminerar inte bara de traditionella aktörernas försprång, utan ifrågasätter även deras affärsmodeller då användare sannolikt inte kommer att betala för något som finns tillgängligt gratis med liknande eller bättre kvalitet.

Den verkliga styrkan hos open source-rörelsen ligger i dess förmåga att samarbeta och bygga vidare på varandras arbete. Genom att tillgängliggöra modeller och tekniker öppet, skapas en miljö där innovation blomstrar på ett sätt som de mer slutna och kontrollerade miljöerna hos stora tech-företag inte kan matcha. Detta är särskilt tydligt inom bildgenerering och nu även språkmodeller, där open source-alternativ som Stable Diffusion har visat vägen framåt.

Slutligen enligt mig bör de större teknikföretagen omvärdera sin inställning till öppen källkod. Istället för att kämpa mot den, skulle de dra nytta av att omfamna och bidra till dessa gemenskaper. Genom att dela med sig av sina resurser och expertis kan de inte bara accelerera innovation utan också positionera sig som ledare i den pågående teknologiska revolutionen. Detta kräver en betydande förändring i hur de närmar sig både utveckling och delning av AI-teknik, men är nödvändigt för att hålla jämna steg med en alltmer dynamisk och distribuerad värld av AI-innovation.

# Slutsats

AI-tekniken fortsätter att utvecklas i en snabb takt, där framsteg inom artificiell generell intelligens (AGI) och specialiserade modeller som Sora visar på den potential och de utmaningar som följer med avancerad AI. Devin, ett exempel på en AI-utvecklare, illustrerar AI:s förmåga att lära sig och lösa tekniska problem autonomt, vilket öppnar för fascinerande tillämpningar men även medför vissa begränsningar i komplexa scenarier.

Samtidigt framhäver Sam Altmans "Moore's Law for Everything" behovet av en ny skattemodell för att hantera ekonomiska omvälvningar som AI kan orsaka, och föreslår lösningar som syftar till att minska ojämlikhet och främja en mer inkluderande tillväxt. Denna idé om att omfördela förmögenheter från AI-driven ekonomisk tillväxt till allmänheten är både utmanande och lovande.

Öppen källkodsmodeller står nu i centrum för AI-utvecklingen, vilket utmanar de traditionella teknikjättarna genom att erbjuda flexibla och kostnadseffektiva lösningar. Denna demokratisering av AI-teknik möjliggör snabba innovationer och personliga anpassningar som tidigare var otänkbara, vilket visar på den kraft som ligger i samarbete och öppen innovation.

Framtiden för AI är otroligt spännande och det är spännande att se vad som händer framöver. Med varje framsteg vi gör, ökar förståelsen för hur AI kan användas för att inte bara förbättra liv utan också att tackla stora samhälleliga utmaningar. Varje ny utveckling bär med sig både möjligheter och risker, och det är avgörande att vi navigerar i denna nya era med en balanserad syn på teknikens potential och dess påverkan på samhället.