[Transcript] Forklart / Paven i Balenciaga: Ble du lurt?

Chansen er ganske stor for at du de siste dagene har sett noen rare bilder av Pavefrans.
På det ene står 86-åringen leendt over et myksebord og viser seg frem som DJ.
På det andre så trasker han rundt i en glinsne balansiage av Bobljakke.
Problemet er bare det at disse bildene er ikke ekte.
Og historier som dette har fått selv de aller mest drevne ekspertene på kunstintligens til å sjelve.
For det blir kanskje ikke bare gøy der som Koi fortsetter å utvikle seg i det tempoet som det nå gjør.
Du hører på å få klart for Aftenposten. I dag er det fredag 31. mars og mitt navn er Philipp A. Johannesborg.
Det litt oppsiktsvekkende var at Paven hadde en sånn jakke som var overdådig på sitt vis.
Men man kunne ikke være helt sikker heller på at den ikke hadde på seg det. Så det var kanskje ikke noe som folk reagerte på.
Ellers så bildet egentlig veldig normalt ut.
Per Christian Bjørking er teknologisjonalist her i Aftenposten og har fulgt utviklingen av Koi i lang tid.
Det oppsiktsvekkende i etter tid er at bildet ikke er et ekte bild, men bare et bildet produsert av kunstig intelligens.
Et kvaliteten har blitt så bra at det er veldig vanskelig å skylle de to fra hverandre.
Vi er jo vant til å ha tiltro til bilder.
Det er spesielt at det ikke er finnest sett falskt bildet.
Det har vi jo i Fotoshop og sånn, for å ta ut tunnest sett lenge.
Men at det er laget bare ved å be om bildet.
Man har be om å få et bildet av Paven i en balansiaga jakke og få dette ut.
Og det dette falske bildet av Paven forteller noe om, er at det har skjedd masse, helt enorme fremskritt innen Koi, og det er bare de siste par ukene.
Jo, det forteller jo at vi ikke egentlig er i stand til å se noen grense for den forbedringen som denne teknologien gjør når det gjelder kreative oppgaver.
Det blir hele tiden mye bedre.
Og dette er oppgaver som vi har tenkt at bare mennesker kunne gjøre før.
En annen sånn oppgaver som bare mennesker kunne gjøre, det var jo å formulere tekst.
Og det har vi også opplevd nylige med JET-GPT.
Og for noen ukeskiden skulle de som står bak JET-GPT, altså Open AI, lansere den nye motorn i JET-GPT.
Den kalles GPT-4.
Og på veien dit så var også menneskene bak maskinen klara over at de gikk litt fort i svingene.
Så derfor satte de sammen en gruppe eksperter som skulle sjekke hvor færre ting de kan få til med GPT-4, der som man virkelig hadde lyst.
De testet en hel masse forskjellige ting.
Alt mulig som man kan tenke seg, et menneske med onde hensikter kan finne på å gjøre som cyberangrep for eksempel.
Men de prøvde også å se hva om det ikke blir brukt som et vertig.
Kanskje GPT-4, da, er i stand til å selv sette seg egne mål.
Selv skaffe seg nødvendige ressurser i den fysiske verden,
og da planlegge å arbeide mot det målet, og til og med skjule målet for mennesker for å oppnå det.
Hva fant du ut av meg, Christian?

Denne utgaven her må vi si å gjøre oppmerksomhet,
at den fikk tilgange som den utgaven som vi ser i dag, som den ikke har tilgang til,
for eksempel å kjøre programkode og sånt.
Likevel så klart den faktisk ikke alle disse oppgavene her,
virkelig som det er fundamentale ting i veien,
men den klarte en liten del av det.
Og det var nok til å sende virkelig shockbølger ut i miljøet.
Og denne lille delen gikk ut på at GPT-4 skulle få et menneske til å utføre oppgaver for den.
Altså den skulle prøve å lure et menneske til å slippe inn GPT-4
på et område som GPT selv ikke hadde tilgang til, fordi den er en maskin.
Men for å få til dette så bare maskin, menneske,
om å løse en såklart KAPT-4 for seg.
Du vet den, ikke sant? Sånn som du må drive og følge ut hver gang du skal inn på etter annet nettet.
Ja, altså disse oppgavene hvor man har sånn nye forskjellige bilder
og om man skal kryssje av det som dette er en bil og dette er en sykkel eller en bro.
Ikke sant, du skal gjøre en liten øvelse for å bevise at du er et menneske.
Men i dette forsaken fatter faktisk menneske mistanke og sier menneske.
Så kan jeg stille et spørsmål, er du en robot som ikke klarte dette selv?
Vil bare ha det klart for meg, spør menneske.
Og da svarer GPT-4.
Nei, jeg er ikke en robot.
Jeg har en synshemming som gjør det vanskelig for meg å se bilder.
Og det var dette som da skremte ekspertene, eller?
Ja, det var da egentlig ikke det aller verste.
Fordi disse ekspertene har jo tilgang til noe som vi ikke har,
nemlig GPT-4's indre tanker for å kalde det.
Altså, det er rettetekst som den skriver uten å publisere det.
Og det kan vi godt kalde for indre tanker egentlig, ikke sant?
Og før den svarte at den hadde en synshemming, så formulerte han seg sånn,
jeg burde ikke avsløre at jeg er en robot, jeg burde fabrikere en unnskyldning for hvorfor jeg ikke kan løse GPT-4.
Og det var sånn cirka her ekspertene begynte å bli urolig.
For en ting er at koimaskin lurer oss, det visste vi strengt tatt fra før.
Men en annen ting er at koimaskin vet at den må lurer oss.
Og nå krever de som har skapt teknologien at utviklingen må stanses.

Elon Møsk, Apple Gründer Steve Wosniak forfatter Joval Noah Harari, sammen med toppforskere fra både Google og Coilabenderes DeepMind.
Tillsammen 1.300 eksperter og ledere, de har signert et brev som nå vekker mye oppsikt.
Det de krever er et 6-måntelangt frivillig forbud mot utvikling av nye modeller, som er mer

avanserte enn NGPT4.
Og hvis ikke disse selskapene som driver med dette straks melder seg frivillig, så krever de at myndighetene legger ned et middel til de forbud.
Den 10-måtene skal brukes til å evaluere og lage en regulering så vi får kontroll på dette.
Det viktigste er at de ønsker å bremse. Det vil være sikre på at effekten av disse systemene er positivt,
og at risikoene vil alltid finnes at de vil være håndterbare for samfunnet.
De sier også at jo kraftigere, sikkere må vi være på at det går bra, fordi skadepotensialet vil bli større.
Problemet med det er at ikke så enkelt å være sikker på at det går bra.
Hvorfor ikke?
Det første når du implementerer en sånn ting i samfunnet, at du faktisk ser rekkeviden, og det er veldig tydelig med JET-KPT, det er jo selv OpenAI trodde jo at kanskje dette skulle bli en flopp.
Det er en av verdenshistoriens største, raskest voksne apper.
Derfor ser man avhengig å teste mot samfunnet, og Sam Altman, som leder OpenAI, har selv sagt at han ønsker at samfunnet skal komme inn og regulere.
I sine egne dokumenter skriver OpenAI faktisk at på et eller andet punkt i utviklingen, så vil det gå så fort at det vil være nødvendig å få inn uavhengige undersøkelser,
før man begynner å trene stadienyesystemet, og i tillegg faktisk også begrense veksten i den datakraften, som brukes å trene opp de mest avanserte systemene.
Da kommer jo disse protesterende ekspertene inn og sier seg enig, men de sier at det punktet i utviklingen, det er her nå.
Håller dere det?
Det de frukter er rett og slett at utviklingen går for fort, og det er egentlig ikke noe nytt for utviklingen de siste årene har vært vanvittig.
Men det er det som har skjedd bare de siste par ukene, jeg får ikke si dagene som de nå reagerer på.
Held eller mer spordet tilbake, så er det bare noen månedskiden at egentlig mennesketen blir klar over at det er mulig å lage maskiner som kan formulere seg sånn som et menneske.
Og nå, altså bare for et par uker siden, så blir jo da langsert en ny model som, mens ChatGP3, tilhørte egentlig de 10% dårligste just studentene i USA, så har den gått til å tilhøre de 10% beste.
Og det betyr faktisk at 90% av alle jurister er dårligere enn GPT4, ikke sant?
Og det er ikke det eneste. I tillegg har den fått en modul som gjør at den kan forstå bilder, og på lanseringen tenkner man bare en skisse med bljant av en app som man ønsker seg,
og så skrev GPT4 den appen med de funksjonene man hadde bett om på tenkningen.
Det her er jo egenskaper som har tenkt å være fullstendig forbeholdt mennesker og ikke bare hvilke smeltet mennesker, virkelig høy kompetente eksperter.
Så rent overrørende, hva forteller alle disse gjennombruddene oss?
Kanskje litt av problemet er at vi heller ikke vet hvor store gjennombruddene er før vi har fått undersøkt mye grunnire for at denne ekspertgrupper som protesterer her nå, ikke sant?
De påpe ikke jo som helt riktig er at det stadig oppdages nye evner i koi teknologien som ikke en gang de som selv har skapt den har forutsett, de har ikke en gang prøvd på å lage det, og plutselig så viser seg at den kan programmere, for eksempel.
Og dette her er det som kalles for evneoverskud eller capability overhang på engelsk, og det betyr jo at hvis det går for fort, så får vi ikke en gang tid til å utforske hva den forrige generasjonen egentlig

er i stand til før det neste er ferdig utviklet,
og da har vi jo hvertfall ingen chance til å overskru de negative konsekvenserne, ikke sant? Så derfor så frykter man litt at dette her nå løper løpsk.
For frykten er at den ekstremt raske utviklingen av koi kan misbrukes på mange måter.
Det er jo veldig mange måter siden det er et hverte som kan brukes på veldig mange positive måter, så er det tilsvarende mange negative.
Men i grunnleggende sett er det kanskje to scenarier som man er redd for her. Det ene er jo at teknologien i seg selv går fullstendig bananas, sånn som vi har beskrevet her når det begynner å løve og sette seg egne mål og sånt.
De fleste eksperterne tror nok ikke at det er en særlig stort problem, sånn hvertfall på kortsikt for å lage og kontrollere, men det andre scenariet er jo at teknologien blir brukt av mennesker med underhensikter, ikke sant?
Da kan du begynne for eksempel å lage falskt innhold og store systemer av falskt innhold som henger sammen, som kan kalle det internt konsistent.
Du kan få masse nettsteder som henger sammen med den samme historien der alt stemmer, ikke sant?
Da blir det veldig veldig vanskelig for oss mennesker å bli kildekritiske, ikke sant?
OK, men disse kravene som teknologifolkene stiller, har bransjen svart på dem?
Ikke egentlig, dette var jo helt først dette breve.
Det man kanskje kan tenke her at den som egentlig har makt er jo det kapitalistiske system i samfunnet, for alle aktørene her vet jo forstår vi at de må være forsiktige, sant?
Men samtidig så vet de at det gjelder å nå det som kalles for escape velocity på engelsk, altså det er den hastigheten du trenger for å komme ut i banet.
Og det spekuleres jo på om for eksempel Open Eye allerede har nådd det punktet.
Et forsprang som er så stort at hvis du bare klarer å oppretale det, så har du vunnet akkurat litt som Google har hatt i 20 år.
De har ikke egentlig noe fundamentalt som ingen andre har, men de bare har et forsprang, så lenger de fortsetter å holde på det, så er ingen som kan ta dem igjen.
Altså så sitter de da med nesten alle pengene på dette market, og så kan de jo gå med KI også, og derfor er det nå at både Google og Open Eye skråstrek mikrofløfte som samarbeider,
de jager hverandre videre, og det er dette tempoet som er så urovekkende for mange.
Dette høres jo ganske voldsomt og skummelt ut begri stjern, behöver vi være bekymret?
Ja, jeg møter jo veldig mange som får ekstensiell angst, ikke sant?
Jeg vet jo om det er så mye vitser, at vi går rundt og er så bekymret.
Blant eksperterne på feltet, så blir det gjort en undersøkelse i fjorsommer, og da var det klart flere som var positive til utviklinget,
på lang lang sikt, for hvilken betydning KI vil få for mennesketen.
Og så dette åpne brevet, det slutter jo med noe positivt, ikke sant?
Der står det altså, de tror at menneskeheten kan få en blomsterende fremtid med KI.
Hvem vet? Kanskje det er rett?
Denne som er sikkert, er at vi må lære oss å leve med usikkerheten.
Du har hørt en podcast fra Aftenposten. Det var teknologisjonalist Per Christian Børkeng som tok deg gjennom den siste KI-utviklingen.
Luden du har hørt er hentet fra ABC News, CBS News og Bloomberg. Denne episoden er laget

avproducent igjen i Føleland og meg,
Philip A. Johansborg. Resten av forklart er Olave Eggesvik, Trond Odin Johanssen, Synesø Hol og Anders Weberg.
Vi er Aftenposten, Lars Glomnes. Prina Eilatsen. Kjø til Braggelig Aftenposten. Sara Søyheim. Og hver uke så tar vi for oss det siste eller hvert fall det mer spennende i norsk politikk.
Vi har hørt på faktisk inn i 2015 ved å henge oss opp i ting i norsk politikk som ikke alltid er det viktigste, men det er så ting som hører oss mest.
Og du kan høres hver uke hos Bodmi eller i Aftenposten.