Etiske utfordringer ved bruk av GPT-2

Det eksisterer ingen uenighet om at språkverktøy som GPT-2 er verdige objekter for interesse og fascinasjon. Som rent teknologiske fremskritt krever de sin respekt. De vil utvilsomt være ressursfrigjørende, men etikken stopper ikke der. Vi vil forsøke å peke på dysfunksjonelle aspekter ved språkteknologien, samt understreke hvor de ikke strekker til.

1.
Hva nyhetsbyråer ønsker å prioritere, er beklageligvis ikke alltid sammenfallende med hva som er det etisk ideelle. Det er for eksempel ikke nødvendigvis ønskelig, fra et brukerperspektiv, at en nyhetssak publiseres for tidlig, da dette går på bekostning av kvaliteten og etterrettelighetен i nevnte sak. Sett fra eierens økonomiske øyne, kan en slik reduksjonistisk maksimering trolig være gunstig, men den økonomisk gevinsten går ikke nødvendigvis til frigjøring av ressurser og dermed dydige formål. Det argumentet står stødig uavhengig av algoritmens ferdighetsnivå og effektivitet, altså er det et deskriptivt og amoralsk spørsmål. Imidlertid hviler det forutgående på en antagelse om at algoritmer ikke vil overta for journalister, men være et assisterende verktøy, noe journalister kan gå over og verifisere, samt stå ansvarlig for.

Hva algoritmisk genererte tekster angår, kan dette være etisk forsvarlig gitt at en rekke kriterier tilfredsstilles. En unyansert utilitaristisk tanke som utelukkende ser på tids- og ressursoptimalisering er utilstrekkelig. Skal genererte artikler brukes, vil dette under ingen omstendigheter fraskrive ansvar fra redaksjonen eller journalisten som brukte algoritmen som et behjelpelig verktøy. Språkmodeller som GPT-2 er ikke velegnet til å utføre journalistiske oppgaver, særlig hva kvalitet angår. Utvilsomt kan de genererte tekstene fange folks oppmerksomhet, men de kan ikke per nå generere reelle kilder og andre faktorer som konstituerer en artikkels troverdighet og etterrettelighet. Algoritmen ekstraherer ut fra datasettet, ord og setninger som på en eller annen uforklarlig måte er nært assosiert med setningen som mates inn – dette er ingen artikkel – selv om alle *lesere* ikke nødvendigvis merker forskjell.

Om redaksjoner ikke tar noe av dette til seg, men velger å publisere algoritmisk genererte tekster likevel, kommer de ikke unna kravene som stilles til dem som redaksjon. Algoritmen er ingen selvstendig aktør, dermed faller alle feil på redaksjonens skuldre. Algoritmen evner ikke forsvare sine valg, dermed mister redaksjonen sin rolle som medium for offentlig diskurs – da dialog umuliggjøres om den ene parten ikke kan svare for seg. Dette er et viktig poeng for redaksjonen må stå ansvarlig slik at ikke ansvaret faller på “ingen”. For at leserne, som epistemiske subjekter, skal være best stilt i møtet med informasjon og skrevne artikler, pliktes redaksjonen til å informere om hvem og hva som står bak teksten, samt hvor informasjonen kommer fra. Dette er essensen av journalistikk, som vil

helhetlig forsvinne med genererte tekster. Ulike institusjoner må gjerne generere og publisere slike tekster, men da mister de sin autoritet og etos som *journalister*. Du må gjerne utarbeide en nettside som genererer tekster om ulike temaer, men journalistikk kan det ikke kalles. Fiffig likevel.

2.
Ettersom den kommunikative forståelsen mellom to personer fra ulike språkkulturer ofte er tilnærmet lik null, vil introduksjonen av et hvilket som helst verktøy kraftig øke kommunikasjonsevnene mellom dem. Jo lavere terskel og jo mer uformelt, jo bedre er Google sin simultanoversettelse. Det samme gjelder generering av tekster til alt av allerede-eksisterende film. Jo mer formell og avgjørende oversettelsen blir, jo mindre relevant blir Google sitt verktøy, det må forstås som et lavterskel tilbud.

For de fleste multikulturelle diskurser ville denne teknologien forbedret forståelsen, men det impliserer ikke at dette gjelder for *alle* diskurser. Det heter seg at enhver oversettelse samtidig er en tolkning, det er fordi det ikke finnes et 1:1 forhold mellom et språk og et annet. Samt er ikke oversettelse noe som skjer simpelthen ord for ord, det er også setning for setning, paragraf for paragraf. Her oppstår den vanlige blackbox- og forklaringsproblematikken, i og med algoritmen ikke kan forklare hvorfor den valgte som den gjorde – da det ikke finnes én objektivt korrekt oversettelse og algoritmen nødvendigvis måtte ta valg. Uunngåelig sniker det seg inn en bias, som gjør seg gjeldende i særlig sensitive saker. I slike – og andre – tilfeller, ville en profesjonell (og tradisjonell) tolk som oversetter, ha forståelse for omkringliggende kontekst, og gjøre en bedre jobb.

3.
Hvorvidt Google har ansvar for misforståelser og feil som oppstår ved bruken av deres produkt, avhenger i stor grad av hva slags tjeneste Google tilbyr, altså er det et juridisk spørsmål i langt større grad enn det er et etisk spørsmål. En holdes ansvarlig i kraft av hvilke løfter man inngår. Dersom Google mener at denne teknologien gjør profesjonelle, og utdannede oversettere overflødig, så kan dette testes og verifiseres empirisk. Det de plikter til å gjøre, i størst mulig grad, er å informere og dele informasjon om hvordan deres algoritme er trent opp og internt sammensatt. Som nevnt er det rart å snakke om "feil" hva oversettelse angår, ettersom det sjeldent finnes noe som ene og alene er "rett". En oversettelse bør heller forstås som et spektrum, mye kan være delvis rett, avhengig av kontekst og rollen til mottaker.

Dersom a) Google ikke eksplisitt nevner hva deres verktøy er istand til, eller b) fraskriver seg ansvar i en klausul, vil ansvaret lenes på brukernes skuldre. De velger tross alt å bruke tjenesten. Det kan her dras paralleller til Tesla og deres selvkjørende biler. De vil heller aldri (frivillig) ta på seg ansvaret for ulykker i trafikken, og viser til at sjåførene tross alt sitter med ansvaret til slutt og hendene på rattet, bokstavelig talt. Imidlertid kan det gå så langt at oversettelsene via Google blir noe alle bruker, i så

fall skal deres tjenester etterprøves og kritiseres så langt det lar seg gjøre, slik at vi over tid får en tjeneste med mest mulig bredde og minst mulig feil. Utover Google sitt ansvar som produsent, vil brukere trolig innse at denne teknologien ikke helhetlig erstatter eksisterende rutiner, men vil være et nyttig verktøy til visse instanser og situasjoner. Kulturelt sett vil kollektivet konvergere på en felles forståelse om verktøyets begrensning, slik vi i dag gjør med Google Translate sitt skriftlige tilbud.

4.
Personlig mener vi at det rette å gjøre, dersom man skal tilby tjenester som GPT-3 omtrentlig gratis, er å samtidig tilby kildekoden. OpenAI er en ideell-organisasjon uten kapitalistiske formål, som hevder de forsker for menneskehetens skyld. Dette impliserer at en bør tillate mest mulig innsyn, da dette vil muliggjøre mest mulig tilbakemelding og forbedringsforslag av den eksisterende tjenesten. Spørsmålet er, hvor mye forklaringspotensial ligger det i kildekoden? Ikke nødvendigvis så mye.

Selv om vi i oppgave 1) antydet at den genererte teksten var uegnet som journalistikk, er det helt tydelig at slik språkteknologi, dersom det får operere uhemmet, representerer en enorm risiko for demokratiet. Dette er fordi kvaliteten er *tilstrekkelig høy*, slik at vi ikke evner å differensiere mellom algoritmisk- og menneskeprodusert tekst – de fleste mennesker er, tross alt, om ikke dårlige skribenter, ofte ukritiske til hva vi leser. Frem til de siste årene, har vi hatt en gyldig epistemologisk antagelse: når jeg leser en tekst, så vet jeg at den er produsert av samme vesen som meg selv. Derfra følger en rekke assosiative idéer som er gyldig i kraft av dette. For å forhindre at denne epistemologiske tilliten brytes, mener vi at en bør etterstrebe å tydelig kommunisere at en algoritme står bak teksten man lever, da dette er nyttig informasjon for leseren. Men, og dette er et men vi har diskutert en del, for hvilke tekster gjelder dette. Jo lenger inn i diskusjonen vi gikk oppdaget vi at det kanskje bør trekkes et skille. Deskriptivt/normativt er nærliggende, men nok for unøyaktig. For hva med bruksanvisninger, oppramsende og rent deskriptive faktabobler om kjøleskap eller tektoniske platers bevegelse?

Uansett, hvis man ikke unngår det epistemologiske tillitsbruddet, risikerer man en enorm forsøpling av våre viktigste sosiale forum. Særlig for medium der terskelen for en tekst er lav – Twitter – kan det allerede i dag være umulig å differensiere mellom menneske og maskin. Dette vil bidra til å ekstremifisere allerede eksisterende ekkokamre og radikalisere uvitende brukere. Antallet mennesker som støtter en tanke eller ideé er med på å påvirke, bevisst og ubevisst, hvilket syn man selv inntar. Det er derfor essensielt å vite om majoriteten av disse "menneskene" er generert av algoritmer eller ikke. Masseprodusering av et ståsted mener vi er en av de største farene med bruken av en slik algoritme. Samtidig vil du ha en selvspeilende effekt, en slags innsnurping, der proporsjonaliteten av maskinprodusert tekst som konstituerer treningsdata til algoritmen vil øke, over tid vil den dermed trenes på seg selv, heller enn å utledes fra originalt menneskelig tekst.

Avhengig av hvordan man vekter denne risikoen – vi vekter den ekstremt høyt, særlig i covid-tilværelsen der man kommuniserer heldigitalt – bør myndigheter kanskje allerede nå vekkes og gripe inn. Dette gjør de trolig allerede, samt er OpenAI klar over det radikalt omveltende potensialet i deres teknologi, da de valgte å ikke slippe løs GPT-3 helt fritt, velvitende om problematikken det kunne innebære.

Språkverktøy som GPT-2 er verdige objekter for interesse og fascinasjon, og som rent teknologiske fremskritt krever de sin respekt. GPT-2 vil utvilsomt være ressursfrigjørende, og et praktisk hjelpemidler for journalister. Samtidig mener vi det fortsatt er en del uløste etiske og juridiske utfordringer ved anvendelse av et slikt verktøy. Det bør i større grad operere som et hjelpemiddel og verktøy for journalister, fremfor å opptre som en autonom aktør.