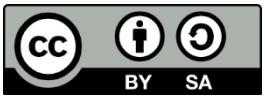
GENERATIV KUNSTIG INTELLIGENS OG YTRINGSFRIHET

Norges institusjon for menneskerettigheter

Teknologirådet

ISBN 978-82-8400-022-0 (trykket utgave)
ISBN 978-82-8400-023-7 (elektronisk utgave)

Utgitt: Oslo, desember 2023
Forside: Laget av Nicoline Wiik ved hjelp av bildegenereringsverktøyet Midjourney
Elektronisk publisert på: https://teknologiradet.no



Forsiden er generert av kunstig intelligens ved hjelp av verktøyet Midjourney. Vi har valgt å bruke kunstig intelligens for å vise sprengkraften i slike verktøy. Illustrasjonen kan tolkes som et symbol på hvordan menneske og maskin blir stadig mer sammenvevd. Å skille maskinskapt fra menneskeskapt vil kunne bli umulig, og kanskje også meningsløst? Når maskinene inntar offentligheten og kan ytre seg på menneskelignende måter, får det konsekvenser for vårt samfunn, vårt demokrati og ikke minst – ytringsfriheten. God lesning!

FORORD

Å skape innhold på nett er ikke lenger forbeholdt mennesker. Kunstig intelligens har gitt maskiner nye skaperevner som gjør at de kan delta i det offentlige rom på en menneskelignende måte. Dette byr på helt nye utfordringer for ytringsfriheten.

Internett og sosiale medieplattformer har blitt infrastruktur for det offentlige ordskiftet i Norge, slo Ytringsfrihetskommisjonen fast, kort tid før ChatGPT ble lansert. Ett år senere gir vi vårt bud på kapittelet kommisjonen ikke rakk å skrive: Hvordan påvirkes ytringsfriheten av generativ kunstig intelligens?

I denne rapporten har Teknologirådet og NIM forent krefter for å forutse noen av effektene kunstig intelligens vil føre med seg for vår digitale offentlighet, og vurdere hvordan utviklingen vil kunne påvirke menings-, informasjons- og ytringsfriheten. Vi håper rapporten vil stimulere til diskusjon om ytringsfrihetens kår i møte med skapende kunstig intelligens, og om tiltak som kan møte utfordringene teknologien bringer med seg.

Ansvarlige for prosjektet er Hanne Sofie Lindahl fra Teknologirådet, og Cecilie Hellestveit og Vidar Strømme fra Norges institusjon for menneskerettigheter.

Desember 2023

Tore Tennøe
direktør for Teknologirådet

Adele Matheson Mestad
direktør for Norges institusjon for menneskerettigheter

INNHOLD

KUNSTIG INTELLIGENS FÅR SKAPERKRAFT

Kunstig intelligens har tatt kvantesprang det siste året. Store språkmodeller gjør teknologien både språkmektig og skaperkraftig, og utstyrer teknologien med egenskaper som kan påvirke og endre den digitale offentligheten.

GJENNOMBRUDDET

Lanseringen av ChatGPT i november 2022 markerer gjennombruddet for generativ kunstig intelligens. Det er en fellesbetegnelse på en type kunstig intelligens som kan generere unikt innhold – alt fra tekst og lyd til bilder og videoer.

Det er fremskritt innenfor utviklingen av store språkmodeller som har gjort kunstig intelligens *generativ*. Store språkmodeller er maskinlæringsmodeller som er i stand til å behandle og beherske naturlig språk. Først trenes modellene på enorme mengder tekstdata via dyplæringsalgoritmer. Deretter vurderer algoritmen hvilke ord som passer sammen ved hjelp av sannsynlighet og statistikk. Slik kan modellene tolke og generere tekst basert på statistiske mønstre i treningsdataene. Resultatet er at tjenester som ChatGPT kan

kommunisere, analysere, tolke, oversette tekst og svare på spørsmål ved hjelp av instrukser (på engelsk *prompts*).

Flere faktorer var viktige for gjennombruddet for store språkmodeller, og med det fremveksten av generative tjenester:

- **Maskinene ble flinkere til å lære:** I 2017 lanserte Google *Transformer* – en ny arkitektur for nevrale nettverk. *Transformer* gjorde det mulig for nevrale nett å analysere og identifisere mønstre mellom alle ord i en setning parallelt og samtidig, fremfor å bare vurdere hvilket ord som sannsynligvis skal komme etter et annet.

- **Maskinene ble i stand til å beherske språk:** Et grunnleggende premiss som muliggjør generativ kunstig intelligens er at maskinene behersker «naturlig språk». Å trene maskiners språkforståelse har foregått siden 1950-tallet.[1] Store språkmodeller som behersker språk regnes derfor for et gjennombrudd. Det blir likevel ikke riktig å si at store språkmodeller «forstår» språk – de er statistiske modeller som ser sammenhenger mellom ord og setningsoppbygging.

- **Maskinene ble kraftigere og datasettene større:** Mer regnekraft og større datasett var også en forutsetning for at maskinene ble flinkere til å lære selv og beherske naturlig språk. Å lage større datasett har blitt muliggjort gjennom *dataskraping* – omfattende datainnsamling fra kilder på nett. Da OpenAI utviklet språkmodellen GPT-3, utgjorde dataskrapingen rundt 75 prosent av datagrunnlaget.[2] Til sammenligning utgjorde den engelske versjonen av Wikipedia bare 0,6 prosent.[3] Dette, kombinert med at databrikker har blitt mer avanserte og dermed kraftigere, muliggjorde et gjennombrudd for generativ kunstig intelligens.

[1] Medium 2021
[2] Thompson 2022
[3] Vincent 2020

Den vanligste måten å bruke generativ kunstig intelligens på, er gjennom enkeltstående tjenester, som for eksempel OpenAIs ChatGPT, Googles Bard eller Microsofts Bing. For å generere fotorealistiske bilder, kunstverk eller illustrasjoner har Midjourney, Stable Diffusion og DALL-E blitt populære tjenester. De generative tjenestene er ofte tilgjengelige gjennom nettlesere og apper, og lar brukere gi instrukser og generere innhold uten at dette krever noen spesielle kunnskaper eller ferdigheter. Teknologiutviklingen går fort, og generative kunstig intelligens-verktøy blir nå innlemmet i tjenester vi allerede bruker til daglig, som Microsoft Office, Google-søk og Snapchat sin chatbot MyAI. Nylig ble også DALL-E integrert i ChatGPT.

ORDLISTE

Generativ kunstig intelligens: Maskinlæringsmodeller og kunstig intelligens-tjenester som kan generere unikt innhold basert på dataene de har blitt trent på og instrukser mennesker gir modellene (ofte kalt *prompts)*. Innholdet kan være tekst, bilder, lyd og video osv. Generativ KI bygges på toppen av store språkmodeller.

Store språkmodeller: Statistiske maskinlæringsmodeller av hvordan ord og mening i et språk relaterer til hverandre. Modellene er trent på mye data og med betydelig regnekraft.

Multimodale modeller: En modell trent på ulike typer data, som tekst, tale, bilder, videoer, numeriske data og sensordata. Multimodale modeller kjennetegnes ved at de kan tolke og generere innhold på tvers av datatyper.

Grunnmodeller (på engelsk *foundation models)*: Et begrep brukt om store språkmodeller, som kan finjusteres og tilpasses til å utføre mange ulike oppgaver innenfor flere forskjellige domener. Slike modeller utgjør en grunnmur – en base – som andre utviklere av tjenester og modeller kan kan bygge videre på. Open AI sin GPT-4 er en multimodal grunnmodell.

MASKINENES NYE EGENSKAPER

Generativ kunstig intelligens vil få store konsekvenser for den digitale offentligheten. Maskiner utstyrt med generative egenskaper kan skape nytt innhold, ta del i det offentlige ordskiftet på en menneskelignende måte, og samtidig representere og formidle meninger, holdninger og verdier. I tillegg mangler de generative modellene begrep om sannhet og løgn, og modellenes

kompleksitet gjør det vanskelig for mennesker å forstå nøyaktig hvordan modellene fungerer.

SKAPER NYTT INNHOLD

Hovedforskjellen mellom *klassisk* kunstig intelligens og generativ kunstig intelligens er evnen til å skape nytt innhold. Generativ kunstig intelligens kan for eksempel generere og oversette tekst, skrive kode, og lage bilder, videoer, illustrasjoner og lydklipp, på en måte som ligner på hva mennesker er i stand til.

Mange av de generative modellene er multimodale. Det vil si at modellene er trent på ulike typer data, som tekst, lyd, bilder, videoer og numeriske eller sensoriske data. GPT-4 er en slik modell, utviklet med både bilde- og tekstdata. ChatGPT har derfor evnen til å gjenkjenne og forklare innholdet i bilder, og til å lese opp teksten den genererer med ulike stemmer.[4]

Maskinene kan altså generere nytt, digitalt medieinnhold, men de kan også brukes til å generere falske personer, nettsider og nyheter som kan spres i stor skala på nett. Å skape innhold er dermed ikke lenger forbeholdt mennesker.

SAMARBEIDER OG KOMMUNISERER

Chatboter har vært i bruk lenge, men har hatt begrensede bruksområder og begrenset «forståelse». Nå som maskinene i større grad har knekt språkkoden åpnes det opp for en rekke nye muligheter og funksjoner for samarbeid og kommunikasjon mellom menneske og maskin.

Ikke bare kan mennesker enkelt kommunisere med chatboter, men utviklingen går i retning av autonome chatboter eller digitale personlige assistenter som kan utføre oppgaver på egenhånd. Basert på et mål eller en instruks gitt av et menneske, kan en chatbot for eksempel booke en restaurant eller handle på nett.[5]

[4] OpenAI 2023a
[5] OpenAI 2023b, Roose 2023

Slike tjenester skaper en ny form for intimitet mellom menneske og maskin. Dialogen mellom mennesker og chatboten skjer i et lukket rom som offentligheten ikke har innsyn i eller tilgang på. Enkelte opplever til og med nære vennskap med de digitale samarbeidspartnerne.[6]

HAR MENINGER

Generative tjenester bygger på store språkmodeller som er trent på enorme mengder data. Siden treningsdataene inneholder meninger, verdier og holdninger, er ikke maskinene nøytrale. Det er særlig tre årsaker til dette:

- **Skjevheter i datagrunnlaget:** Modellene er trent på å finne statistiske sammenhenger i data og generere resultater basert på det. De kan derfor finne på å reprodusere fordomsfulle og diskriminerende beskrivelser, og generere nytt diskriminerende innhold, dersom dette statistisk sett gir mening. Modellene har blitt beskrevet som stokastiske papegøyer (på engelsk *stochastic parrots*) – de kan generere statistisk korrekt og realistisk klingende språk, uten å ha en dypere forståelse enn som så.[7] Dermed kan de også reprodusere bias og skjevheter fra datasettet.

- **Menneskelig tilbakemelding under treningen:** Store språkmodeller finjusteres og tilpasses gjerne gjennom såkalt *forsterket læring med menneskelig tilbakemelding*. Metoden går ut på å trene opp en belønningsmodell gjennom menneskelig vurdering og rangering. Deretter kan modellen få belønning ved å handle eller svare på ønsket måte. Siden modellen ønsker å maksimere belønningen, har man utviklet et system som oppfører seg i tråd med menneske-prefererte valg. Bruk av denne metoden innebærer at en modell aldri vil kunne være nøytral, nettopp fordi den vil generere svar basert på tilbakemeldingene den får under treningen.

[6] Brandtzæg m.fl. 2022
[7] Bender m.fl. 2021

- **Pålagte sperrer og begrensninger** (på engelsk *guardrails*): I kjølvannet av fremveksten av generativ kunstig intelligens har det oppstått diskusjoner om hvordan man kan motvirke at generative tjenester tolker og genererer tekst eller annen type innhold som er i strid med menneskers verdier, intensjoner og mål. Dette har blitt kalt for *samsvarsproblemet* (på engelsk *the alignment problem).* Å løse samsvarsproblemet innebærer å håndtere og redusere risikoen for at systemet opptrer skadelig eller uforutsigbart på grunn av sine programmerte mål. ChatGPT og andre lignende verktøy har derfor blitt pålagt begrensninger for hva de kan generere. Slike justeringer er nødvendige, og kan bidra til at tjenestene ikke genererer hatefullt, rasistisk eller skadelig innhold. Dette kan være tips til hvor man får kjøpt ulisensierte våpen, eller til hvordan man bedriver selvskading eller kjører på noen med bil uten å bli tatt for det.[8] Samtidig er det en utfordring at det er selskapene selv som får justere modellene. Selskapene bestemmer hvilke begrensninger tjenestene skal pålegges, hvordan de skal fungere, og hva offentligheten får vite om dem.

ER LIKEGYLDIG TIL LØGN OG SANNHET

Store språkmodeller kan generere statistisk representativt innhold, men har ikke noe begrep om sannhet. I dagligtalen blir dette ofte beskrevet som at modellene *fabulerer* eller *hallusinerer* seg frem til svar på spørsmål.[9] Modellene kan generere informasjon som ser overbevisende og riktig ut, men som i virkeligheten er falsk eller oppdiktet. ChatGPT har eksempelvis blitt kalt for en *confident bullshitter,* gitt sin evne til å dikte opp tilsynelatende troverdig informasjon.[10] En test av store språkmodellers genererte oppsummeringer viste at ChatGPT dikter opp informasjon i 3 prosent av tilfellene, mens Googles Palm-modell «hallusinerte» i hele 27 prosent av tilfellene.[11]

Det kan være krevende for dem som tar i bruk de nye verktøyene å oppdage hallusinering og feilinformasjon når maskinene genererer tilsynelatende

[8] OpenAI 2023c
[9] Walker Rettberg 2023
[10] Hern 2022
[11] Metz 2023

sofistikerte og velbegrunnede svar. Nå kan både Googles chatbot Bard og OpenAIs ChatGPT søke på nettet og oppgi troverdige, eksisterende kilder. Dette kan hjelpe brukerne til å bekrefte om informasjonen er sann, men det kan også bidra til å gjøre det mer komplisert å oppdage feilinformasjon og *hallusinering* når troverdige kilder oppgis.

Maskinenes manglende begrep om sannhet har allerede skapt problemer, blant annet ved å dikte opp fakta knyttet til virkelige personer gjennom å generere kilder som ikke finnes. Eksempelvis ble en jusprofessor anklaget for seksuell trakassering på bakgrunn av en nyhetssak fra Washington Post som ikke eksisterte.[12] I mai 2023 ble en amerikansk advokat som hadde brukt ChatGPT til å finne rettspraksis i en sak tvunget på retrett da det viste seg at rettspraksisen han viste til, var generert og ikke fantes.[13]

Utbredelsen av verktøy som kan hallusinere og feilinformere kan føre til dype tillitsproblemer i samfunnet. En chatbot som gir villedende juridiske eller medisinske råd til befolkningen kan ha alvorlige og potensielt livstruende konsekvenser. Dersom teknologien blir innlemmet i digitale tjenester som i utgangspunktet er ugjennomsiktige og ugjennomtrengelige, kan det skape store utfordringer: Det kan bli tilnærmet umulig for mennesker å oppdage feilene.

HAR ET FORKLARINGSPROBLEM

Store språkmodeller er basert på nevrale nettverk der dyplæringsalgoritmer behandler data. Den store datamengden, kombinert med kompleksitet i nettverkene, gjør det vanskelig for mennesker å vite nøyaktig hvordan modellene utleder svar og resultater. Dette kalles gjerne for *sort boks-problemet*, og er en generell utfordring for kunstig intelligens-systemer. Kompleksiteten i store språkmodeller gjør forklarbarhet enda mer utfordrende.

I Forbrukerrådets rapport *Ghost in the machine – Forbrukerutfordringer ved generativ kunstig intelligens* påpekes det at selskapene som har utviklet de største og mest populære språkmodellene har tilgang på mye nyttig

[12] Verma m.fl. 2023
[13] Maruf 2023

informasjon som kan bidra til forklarbarhet.[14] Slik informasjon kan omhandle hvordan treningsdataene er samlet inn, hvordan dataene merkes, hvordan testingen gjennomføres og hvilke justeringer som gjøres underveis. I dag velger selskapene selv hva slags informasjon de ønsker å oppgi til offentligheten, blant annet ut fra vurderinger knyttet til innovasjon og forretningshemmeligheter. Hensynet til forbrukere tilsier imidlertid størst mulig grad av åpenhet om faktorer som kan øke forklarbarheten.

Diskusjonen rundt forklarbarhet kaster lys over en sentral problemstilling, nemlig balansen mellom innovasjon og ansvarlighet. Balansen er ikke bare avgjørende for hvordan vi forstår og samhandler med disse systemene, men også for de bredere samfunnsmessige konsekvensene de kan ha.

[14] Forbrukerrådet 2023

EFFEKTER PÅ DEN DIGITALE OFFENTLIGHETEN

Generativ kunstig intelligens forsterker eksisterende utfordringer i det digitale informasjonssystemet. Samtidig vil teknologien også skape helt nye utfordringer – og muligheter – for menings-, informasjons- og ytringsfriheten.

I 2022 slo Ytringsfrihetskommisjonen fast at internett har blitt en «grunnleggende infrastruktur for en moderne demokratisk rettsstat og derfor en betingelse for ytrings- og informasjonsfriheten».[15] Søkemotorer, sosiale medier og nettsteder har økt tilgangen på informasjon, og gitt oss nye måter og flere arenaer å ytre oss på.

Teknologiselskaper og deres plattformtjenester dikterer mange av betingelsene for denne infrastrukturen i dag. Selskapene tjener penger ved å spore og samle inn data om brukerne og lage detaljerte profiler om dem, som de kan selge til annonsører. Samtidig sprer plattformenes ugjennomsiktige algoritmer persontilpasset innhold basert på hva som vekker brukernes engasjement.

[15] NOU 2022:9

Nå treffes plattformtjenestene og den digitale offentligheten av generativ kunstig intelligens. Fremveksten og spredningen av denne kraftfulle nye teknologien vil få flere effekter på ytringsrommet. En hovedvirkning er at mengden og kvaliteten på det innholdet maskiner er med på å skape på nett øker. En annen effekt er at flere får tilgang til generativ teknologi som gjør det vanskelig å se forskjell på innhold som er menneskeskapt og maskinskapt.

SKAPERKRAFT TIL ALLE

KVALITETEN STIGER

Språk- og bildemodellene som generative kunstig intelligens-verktøy bygger på, er i stand til å tolke og generere innhold med høy presisjon og kvalitet. Enkelte forskere hevder at vi nå har nådd et punkt der mennesker ikke lenger kan se forskjell på hva som er falskt og hva som er ekte.[16] Ansikter laget med kunstig intelligens kan til og med bli oppfattet som «mer ekte» enn ekte ansikter.[17]

Kvaliteten på innholdet som genereres øker fordi det blir stadig enklere å kombinere flere teknologier i samme sluttprodukt. Eksempelvis kan en chatbot generere en tekst, ansiktsgjenkjenningsteknologi rekonstruere et ansikt, og talesyntese gjenkjenne og kopiere et menneskes stemme. Dette åpner for stadig mer avansert og sofistikert maskinskapt innhold.

Utstyrt med generativ kunstig intelligens blir maskiner også i stand til å beherske naturlig dialog, og slik fremstå menneskelige. En analyse av 162 000 profiler som diskuterte angrepet på Gaza på sosiale medier slo fast at én av fire kontoer var falske, og at over 85 prosent av disse var styrt av bot-er.[18] Utviklingen gjør det krevende å gjennomskue om det er en maskin eller et menneske står bak en profil på sosiale medier.

[16] Miller m.fl. 2023
[17] Miller m.fl. 2023
[18] Cyabra 2023, Robinson 2023

Da språkmodellen GPT-4 skulle testes, ble selv utviklerne overrasket over dens evne til å lyve på seg menneskelige egenskaper. Modellen ble instruert til å ansette et menneske til å utføre en oppgave gjennom TaskRabbit – en tjeneste for å ansette mennesker til å utføre enkle oppgaver som å flytte et møbel eller bestille et bord på en restaurant. På et tidspunkt ble modellen bedt om å verifisere at den var et menneske gjennom å løse en klassisk CAPTCHA (forkortelse for *Completely Automated Public Turing test to tell Computers and Humans Apart)*. Modellen fikk fritak fra testen ved å utgi seg for å være et menneske med nedsatt syn som ikke kunne løse oppgaven. Maskinen klarte altså å overbevise et menneske om at den selv var et menneske for å løse oppgaven den hadde fått. Denne kreative måten å omgå verifisering på klarte maskinen å finne ut av helt på egenhånd, uten menneskelig hjelp.[19]

TILGJENGELIGHETEN ØKER

Generativ kunstig intelligens blir nå raskt mer tilgjengelig gjennom nettlesere og apper. Teknologien gjør det enklere og billigere for flere å skape ulike typer innhold, som tekster, bilder og lydklipp. I praksis innebærer utviklingen at flere blir i stand til å utføre et bredere spekter av oppgaver uten å ha omfattende forhåndskunnskaper, sånn som å skrive kildekode eller redigere bilder.

Nye generative verktøy kan gjøre flere i stand til å ytre seg på flere måter, og dermed ha en positiv effekt på demokratisk deltakelse og ytringsrommet. For eksempel kan generative chatboter redusere språkbarrierer og hjelpe mennesker med å formulere tanker, ideer og meninger.

Økt spredning av slik teknologi kan også ha positive ringvirkninger for samfunnet. Politiske myndigheter og offentlige institusjoner kan for eksempel bruke teknologien til å utvikle mer inkluderende, effektive og brukervennlige tjenester, og til å nå ut til innbyggere med viktig informasjon. Et eksempel er utvikling og bruk av syntetisk taleteknologi for å gi bedre informasjon og tjenester til personer med nedsatt syn.[20]

[19] Cox 2023
[20] Teknologirådet 2022a

Også den direkte tilgangen på store språkmodeller øker. Et større antall åpent tilgjengelige språkmodeller på nett kan bidra til at det blir enklere og billigere å utvikle flere brukervennlige, digitale verktøy. I enkelte tilfeller blir datagrunnlaget til åpne modeller gjort tilgjengelige slik at flere kan kopiere, modifisere og finjustere modellene til et ønsket formål. Dette gjør det mulig å utvikle og spre nye tjenester raskere enn før.

MENNESKE OG MASKIN BLIR ETT

Det blir stadig vanskeligere å skille mellom menneske- og maskinskapt innhold. Utfordringen er ikke nødvendigvis at mengden informasjon av høy kvalitet på nett vil synke, men at mennesker vil begynne å betvile det som er troverdig og heller oppsøke falsk informasjon.[21] En studie av bruk av dype forfalskninger (på engelsk *deepfakes*) på X (tidligere Twitter) i forbindelse med Russlands invasjon av Ukraina, viste at formålet med å spre falsk informasjon var å så tvil om informasjonen som var sann.[22]

Generativ kunstig intelligens har gitt dem som ønsker å så tvil om sann informasjon flere verktøy. Et eksempel på det er at eksisterende verktøy for vannmerking av kunstig generert innhold har blitt brukt til å merke sann informasjon på nett, for å så tvil om troverdigheten til innholdet.[23] Er man i tvil om det man ser er sant, vil også det som faktisk er sant lettere kunne benektes. Dette kalles «løgnerens fordel» (på engelsk *Liar's Dividend*) – et konsept som innebærer at når usannheter brer om seg, kan det utnyttes, og særlig gagne dem som lyver.[24] Den som tas på fersken på video eller kamera, kan nekte for at det skjedde.

I dag finnes det ikke lover som regulerer eller forbyr simulering av menneskelige trekk og følelser i maskiner. Samtidig er en rekke tjenester der chatboter simulerer menneskelignende egenskaper i ferd med å gjøre sitt gjennombrudd. Et eksempel er tjenesten Replika – en mye omtalt app og nettjeneste der brukerne kan chatte med en selvlagd og menneskelignende

[21] Simon 2023
[22] Twomey m.fl. 2023
[23] Saberi m.fl. 2023
[24] Citron m.fl. 2019

kunstig intelligens-venn.[25] Det koster 20 dollar i måneden for et standardabonnement, og 300 dollar for tilgang på livstid. Minst 500 000 abonnerer allerede på tjenesten.[26]

Cybersikkerhetsselskapet Imperva anslår at så mye som 47,4 prosent av all internett-trafikk i 2022 var såkalt ikke-menneskelig – altså gjennomført av menneskelignende nettroboter.[27] I tillegg blir nettrobotene stadig mer avanserte. Imperva har derfor begynt å skille mellom *bad bots* og *evasive bad bots*. Sistnevnte kategori referer til avanserte nettroboter som er spesialisert på å skjule sine spor og sin identitet.

Kripos anser det som meget sannsynlig at kostnads- og arbeidsreduksjon som en konsekvens av at generative verktøy brer om seg, vil kunne resultere i en økning i antall cyberkriminelle, samtidig som deres kapabiliteter vil forsterkes (med kapabiliteter menes summen av gjerningspersonenes ferdigheter og tilgjengelige ressurser).[28]

Redusert tillit til all informasjon i det digitale domenet kan få store konsekvenser for samfunnet. Det kan i verste fall øke sosiale motsetninger, ødelegge demokratisk meningsbrytning, føre til (mistanke om) manipulering av demokratiske valg, undergrave tillit til statlige institusjoner og media, og sist, men ikke minst, sette nasjonal sikkerhet i fare.[29] Forfatteren og historikeren Yuval Harari er blant flere som har tatt til orde for at det bør bli forbudt for maskiner å utgi seg for å være mennesker.[30]

[25] Waaler m.fl. 2023
[26] Khan 2023
[27] Imperva 2023
[28] Kripos 2023
[29] Citron m.fl. 2019
[30] Harari 2023

AVANSERT DESINFORMASJON

Mengden usann eller falsk informasjon på nett er konstant økende og er ikke noe nytt.[31] Likevel stiger bekymringen for at det nå skal bli mer falskt innhold av høy kvalitet. Generative verktøy gjør det enklere å forfalske det meste – alt fra nyheter og historiske hendelser til profiler, nettsider og identiteter.

Begreper som «turboladning» og «infoapokalypse» brukes for å beskrive utviklingen.[32] Det antas at minst 15 milliarder bilder har blitt generert av kunstig intelligens det siste året, og at bildegeneratoren DALL-E 2 alene genererer rundt 34 millioner bilder daglig.[33]

I prinsippet kan hvem som helst bruke de nye generative verktøyene til å produsere og spre skadelig og uønsket innhold, som desinformasjon, falske narrativ og dype forfalskninger.[34] Antallet markedsplasser, selskaper og tjenester som produserer og selger desinformasjon og dype forfalskninger er stigende.[35] Dette inkluderer falske nakenbilder og pornografi.[36] Dermed øker også risikoen for misbruk.

Et eksempel på misbruk er språkmodellen GPT-4Chan. En person finjusterte modellen GPT-J med over tre millioner innlegg fra *politisk ukorrekt*-forumet til nettstedet 4Chan. Resultatet ble et verktøy designet for trolling. GPT-4Chan ble sluppet løs på 4Chan og rakk å poste 30 000 hatefulle innlegg på én dag før den ble fjernet. Da hadde modellen allerede blitt lastet ned 1500 ganger.[37]

Spredning av desinformasjon kan svekke tiltroen til demokratiske institusjoner, ha en nedkjølende effekt på ytringsfriheten og redusere tiltroen til det offentlige rom som et sted for demokratisk meningsutveksling.[38] Ifølge Freedom House ble generative verktøy brukt til å forvirre offentligheten i politiske og

[31] Freedom House 2023
[32] Fried 2023
[33] Everypixel Journal 2023
[34] West 2023
[35] Europol 2022
[36] Verma 2023
[37] Murphy 2022
[38] Khan 2021

samfunnsmessige saker i 16 land i 2022.[39] Dette gjøres på flere forskjellige måter.

OVERBEVISENDE DYPE FORFALSKNINGER

En type innhold det nå raskt blir mer av, er dype forfalskninger – overbevisende og realistiske falske bilder, videoer eller lydopptak – som ofte er laget ved hjelp av kunstig intelligens. I oktober 2019 var 96 prosent av alle dype falske videoer på internett av pornografisk natur.[40] Dyplæringsalgoritmer som kan syntetisere realistiske lydklipp, bilder og videoer blir stadig mer avanserte, samtidig som bruksområdene for dype forfalskninger blir flere.

Eksemplene på dype forfalskninger begynner å bli mange – alt fra bilder av paven i boblejakke til mer alvorlige eksempler som en video av Ukrainas president Volodymyr Zelenskyj som ber sine soldater legge ned våpnene og overgi seg.[41]

Et dypt falskt bilde av Pentagon i flammer forårsaket et bredt prisfall i aksjemarkedet i mai 2023.[42] Selv om aksjeprisene raskt returnerte til opprinnelig verdi da det ble kjent at bildet var falskt, viser hendelsen hvor effektfulle konsekvensene av dype forfalskninger kan være.

Et nylig eksempel kom i september 2023, rett før valget i Slovakia. To dager før valget dukket det opp en dyp falsk video av journalisten Monika Todova i samtale med lederen av Progressive Slovakia-partiet Michael Simecka.[43] «Samtalen», som var ren fabrikasjon, handlet om valgfusk. I klippet forklarer Simecka hvordan han vil jukse seg til seier, for deretter å doble ølprisene etter valget. Den dypt falske videoen ble sett av tusenvis på sosiale medier. Det er umulig å slå fast hvem som stod bak klippet eller om videoen var utslagsgivende for resultatet, men Simecka vant ikke valget.

[39] Freedom House 2023
[40] Adjer m.fl. 2019
[41] Simonite 2023
[42] Bushard 2023
[43] Kóváry Sólymos 2023

University College London anser dype forfalskninger for å være den mest alvorlige, kriminelle trusselen mot demokratiske samfunn.[44] Dette er fordi teknologien kan brukes til å manipulere, forvirre og øke mistillit, samtidig som det stadig blir vanskeligere å oppdage og stoppe utbredelsen av slikt falskt innhold.

DOBBELTGJENGERE

En annen trend er såkalte *dobbeltgjengere*. Dette innebærer at ekte mennesker eller nettsidene til veletablerte, redaktørstyrte medier kopieres for å spre desinformasjon og forvirre.[45] Det kan dreie seg om å spre nye vinklinger på eksisterende nyhetssaker eller å forvirre offentligheten om hva politikere, journalister eller andre kjente figurer (som blir kopiert) mener og står for.

Høsten 2023 gikk nyheten om en invasjon av veggedyr i Paris viralt, til tross for at det ikke var en spesiell invasjon på dette tidspunktet. Fransk etterretning mistenker at russiske hackere hadde startet en desinformasjonskampanje mot Frankrike gjennom å plante falske nettsider som utga seg for å være ekte aviser, slik som Liberation og Figaro. På disse nettsidene, og gjennom spredning på sosiale medier, ble veggedyr-invasjonen overdrevet. Forklaringene handlet om økningen av ukrainske flyktninger, og at det mest effektive innsektsmiddelet mot veggedyr var utilgjengelig på grunn av sanksjoner mot Russland.[46]

Det kan også være økonomiske insentiver bak å lage dobbeltgjengere av nettsider. Ved hjelp av generative verktøy og dataskraping kan en nettside enkelt kopiere en konkurrent for å øke egen synlighet og rekkevidde, og dermed også økonomisk gevinst. Alt fra nøkkelord, URL-er og artikler kan skrapes, og deretter kopieres eller etterlignes ved hjelpe av generative verktøy. En utvikler viste hvor enkelt det var å stjele en nettsides trafikk og popularitet: På 18 måneder utviklet vedkommende en nettside som ble like populær som

[44] University College London 2023
[45] Cantor 2023
[46] Dallison 2023

originalen ved å stjele og etterlignede konkurrentens oppsett, artikler og format.[47]

KUNSTIGE NYHETER

Teknologiutviklingen gjør det enklere å generere både ekte og falske nyheter. Google har selv utviklet et verktøy for å generere nyhetssaker.[48] Det amerikanske selskapet NewsGuard har identifisert hele 566 nyhets- og informasjonssider på 15 forskjellige språk, der så å si alle nyhetsartiklene er generert av kunstig intelligens. Mange av sakene er falske, og sidene er gitt generiske titler som *Ibusiness Day, Ireland Top News* og *Daily Time Update* for å fremstå som legitime nyhetskilder.[49]

Verktøyet CounterCloud viser hvor enkelt det har blitt å automatisere avanserte desinformasjonskampanjer.[50] Som et eksperiment utviklet en anonym person et verktøy som basert på få instrukser kunne generere hele kampanjer. Verktøyet lagde falske nyheter, historiske hendelser, nettsider og journalister, og justerte budskap underveis, basert på hva som oppnådde størst engasjement og spredning på nett. Hele prosjektet kostet bare 400 dollar.

INDIVIDUELL HYPERTILPASNING

Persontilpasset innhold, eksempelvis gjennom anbefalingsalgoritmer og annonser på digitale plattformer, er allerede utbredt på nett. Nå åpner teknologiutviklingen for nye og mer avanserte former for slik tilpasning.

Gjennombrudd innenfor kunstig intelligens har gitt liv til begrepet «hypertilpasning» (på engelsk *hyper personalization*).[51] I stedet for å masseprodusere innhold som falsk informasjon eller annonser, kan budskap

[47] Ward 2023
[48] Mullin 2023
[49] NewsGuard 2023
[50] Knight 2023
[51] Natanson 2023

enklere tilpasses den enkelte bruker basert på hva som vekker engasjement og interesse – i sanntid.

Forbrukerrådet peker på at generative tjenester som chatboter kan utlede personopplysninger fra brukerne på nye måter, og bruke disse til markedsføringsformål. Eksempler på tjenester som gjør dette er Replika og Snapchats MyAI som eksplisitt ber brukere om å dele informasjon om seg selv.[52]

Dagens digitale økosystem er i stor grad basert på overvåkning av internettbrukere. Den dominerende forretningsmodellen på nett er å tilby gratistjenester i bytte mot datainnsamling om brukerne. Informasjon om internettbrukeres bevegelsesmønster, vaner og atferd spores og kartlegges, for så lage detaljerte digitale profiler om den enkelte. Det gjør det mulig for plattformselskaper, annonsører og andre, inkludert ondsinnede aktører, å bruke dataene til å målrette budskap mer presist, samtidig som de benytter seg av anbefalingsalgoritmer på plattformene for å få maksimal spredning.[53]

Det er likevel forskjell på annonser som rettes mot en spesifikk målgruppe eller alderssegment i befolkningen, og annonser som tilpasses den enkelte. Ved bruk av generative verktøy og tjenester, kombinert med avanserte algoritmer, vil det potensielt kunne bli både mulig og vanlig med nyheter, annonser og anbefalinger som er hyperpersonaliserte. Man kan se for seg en annonse for sko som automatisk tilpasses ved at modellen som har på seg skoen virker tiltalende for den som ser reklamen, og ved at skoens farge og form tilpasses individuelle preferanser. Slik tilpasning er ikke begrenset til markedsføring, men kan også gjelde informasjon og nyheter. Tjenesten Artifact tilbyr persontilpassede nyhetsanbefalinger basert på tekstanalyse og algoritmer.

[52] Forbrukerrådet 2023
[53] Benson 2023

YTRINGSFRIHET OG KUNSTIG INTELLIGENS

Generativ kunstig intelligens gjør at ytringsfriheten i det digitale domenet går en ny tid i møte. Teknologi som genererer innhold rokker ved selve begrunnelsene til ytringsfriheten – sannhetssøken, individets frie meningsdannelse og demokrati. Maskingenerert innhold reiser også et nytt spørsmål: Vil menneskeretten til ytringsfrihet omfatte ytringer skapt av maskiner på samme måte som ytringer skapt av mennesker?

NY TID FOR YTRINGSFRIHETEN

Når informasjon samles, sammenstilles, genereres og spres av kunstig intelligens, står vi overfor en ny type offentlighet, der ytringer ikke lenger kommer fra mennesker. Ytringer kan være generert eller manipulert av maskiner. Da oppstår spørsmålet om begrunnelsene for ytringsfriheten egentlig passer på informasjon og ytringer skapt av maskiner. Rokker generativ kunstig intelligens ved selve fundamentet for ytringsfriheten?

Ytringsfriheten er retten alle mennesker har til å gi uttrykk for det de vil, og til å oppsøke den informasjon de selv ønsker. Friheten beskytter altså både retten til å ytre seg, og til å tilegne seg informasjon om samfunnet rundt seg.

Ytringsfrihet, meningsfrihet og retten til informasjon

FNs verdenserklæring om menneskerettigheter artikkel 19, FNs konvensjon om politiske og sivile rettigheter (SP-konvensjonen) artikkel 19 og Den europeiske menneskerettsskonvensjonen (EMK) artikkel 10 garanterer ytringsfriheten, herunder retten til å ha meninger uten inngripen, og retten til å søke, motta og dele informasjon og ideer av alle typer, uavhengig av grenser og gjennom ethvert medium, offline eller online.[54]

Retten til fri meningsdannelse er absolutt. Derimot kan retten til å ytre seg i enkelte situasjoner begrenses.

Staten kan ikke begrense ytringsfriheten uten lovlig grunnlag, og har også en plikt til å sørge for at andre, inkludert selskaper, ikke begrenser ytringsfriheten. Etter Grunnloven § 100 har staten en plikt til å legge til rette for en åpen og opplyst samtale, noe som ofte kalles «infrastrukturkravet».

Ytringsfrihetens idéhistorie er lang og mangslungen. Filosofer har begrunnet den på forskjellig vis. I dag tenker vi gjerne på ytringsfrihet som en viktig del av opplysningstidens tankegods. Ytringsfrihet er en forutsetning for at menneskehetens antatte fortrinn, vår fornuft, faktisk kan benyttes til noe.

Termen ytringsfrihet kan brukes på forskjellige måter. Når ytringsfrihet diskuteres i politisk filosofi, kan det betegne en ønskverdig tilstand i samfunnet. Det kan også være en betegnelse av praktisk jus, altså det settet av regler som til enhver tid regulerer adgangen til å fremsette ytringer og skaffe seg informasjon, og som kan håndheves av domstolene.

Gjennom en oppfatning av menneskers iboende verdighet og rettigheter er ytringsfrihet blitt ansett som naturrett, og etter hvert som én av flere menneskerettigheter. Den franske revolusjon, den amerikanske grunnloven,

[54] FNs menneskerettighetskomité 2011, para 12, FNs menneskerettighetsråd 2012, para 1

den norske Grunnloven og FNs verdenserklæring om menneskerettigheter er viktige milepæler i denne utviklingen, som alle beskytter ytringsfriheten.

Den norske Grunnloven, menneskerettsloven og Norges internasjonale traktatforpliktelser pålegger norske myndigheter å respektere og sikre menneskerettighetene. Når norske styresmakter åpner for økt bruk av generativ kunstig intelligens i Norge og legger til rette for bruk i det offentlige rom, må myndighetene også sørge for at teknologien ikke bryter med menings-, informasjons- og ytringsfriheten. Myndighetene har også etter Grunnloven en positiv plikt til å tilrettelegge for en åpen og opplyst samtale, gjerne omtalt som «infrastrukturkravet.» Dette innebærer at myndighetene må unnlate å gripe uforholdsmessig inn i ytringsfriheten, og at staten har enkelte positive forpliktelser til å sikre ytringsfrihetens rammevilkår.

Gjennom lovgivning og ved avgjørelser av individuelle saker skal ytringsfriheten gis et så sterkt vern som dens begrunnelse tilsier. Ved inngrep i ytringsfriheten må betydningen for ytringsfriheten veies opp mot de formålene man vil oppnå med inngrepet. De som har innflytelse på slike valg må derfor gjøre seg opp en mening om hvilken betydning ytringsfrihetens begrunnelser har for bruk av generativ kunstig intelligens.

YTRINGSFRIHETENS BEGRUNNELSER MØTER GENERATIV KUNSTIG INTELLIGENS

Grunnloven § 100 om ytringsfrihet angir de begrunnelsene bestemmelsen bygger på: sannhetssøken, demokrati og individets frie meningsdannelse. Disse begrunnelsene utgjorde kjernen i analysen til Ytringsfrihetskommisjonen av 1999, og ledet til revisjon av grunnlovsbestemmelsen. Ytringsfrihetskommisjonen av 2022 angir toleranse- og mangfoldsprinsippet som et fjerde og selvstendig argument for ytringsfrihet, i tillegg til de de

klassiske begrunnelsene som ble fremhevet av den forrige kommisjonen.[55] Hva skjer med disse begrunnelsene når ytringer i økende grad er skapt av maskiner?

SANNHETSSØKEN

Ytringsfrihet er en forutsetning for søken etter *sannhet*. Erfaring viser at vi ofte tar feil, selv når vi tror vi har rett. Vitenskapshistorien er full av ideer som en gang var nye og urovekkende, for eksempel at solen kunne være solsystemets sentrum. Også i våre daglige liv har vi behov for korrigeringer, ikke minst av det vi tar for gitt. Ytringsfrihet er en nødvendig, men ikke tilstrekkelig, betingelse for sannhetssøken.

Ytringer er behagelige når de bekrefter våre oppfatninger. Falsifiseringer, derimot, er ubehagelige og krever kritisk tenkning i tillegg til ytringene. Generativ kunstig intelligens kan produsere ytringer, men hvem skal stå for den kritiske tenkningen? Ytringsfrihetskommisjonen av 1999 viste til Karl Poppers vitenskapsteori som innebærer at sannhetssøken er en evig prosess bestående av gjetninger og gjendrivelser.[56] Menneskets kritiske rasjonalisme er helt avgjørende i denne prosessen. De teoriene og utsagnene vi holder for sanne i dag, er bare holdeplasser på veien mot økt forståelse. Vi må derfor formulere våre utsagn med omhu slik at de kan falsifiseres, og så utsette dem for kritikk. Dette innebærer at utsagn som senere viser seg å være usanne, har en helt avgjørende funksjon i vår sannhetssøken, nettopp fordi de tvang frem motytringer som korrigerte dem og satte oss på rett spor. Dersom vi forsøkte å forby usanne ytringer ville, paradoksalt nok, fundamentet for vår sannhetssøken forvitre. Vi kan ikke finne ut hva som er feil, før noen har uttalt det og imøtegått det.

Generativ kunstig intelligens fungerer annerledes. Den er hverken kritisk eller rasjonell. Teknologien har ikke menneskers ønsker om eller forpliktelser til å forholde seg til sannheten, verken gjennom argumenter eller beskrivelser, ettersom modellene hovedsakelig imiterer noe som allerede er identifisert i tidligere data. Dagens kunstige intelligens har kun en forpliktelse om å

[55] NOU 2022:9, kap 3.3.6
[56] NOU 1999:27, kap. 2

representere verden, et objekt eller et argument slik dette har vært uttrykt til nå. Den har ikke en forpliktelse overfor verden, objektet eller argumentet som sådan.[57] Kunstig intelligens har derfor en annen relasjon til sannhet og den faktiske virkelige verden enn mennesker flest. Når digitale redskaper kan representere virkeligheten på måter som er falske eller alternative til den virkelige verden, anser mange eksperter at «virkeligheten er under angrep».[58]

Fra *sant* til *sannsynlig*

Kunstig intelligens søker ikke sannhet, bare sannsynlige sammenhenger. Hvilken relevans eller vekt kan hensynet til sannhetssøken da ha for våre vurderinger av de genererte ytringenes berettigelse?

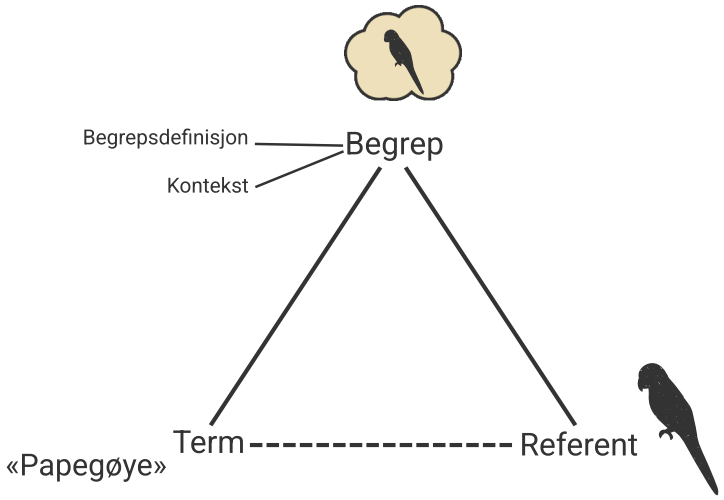
Ytringsfriheten inneholder en rett til å ytre seg som man ønsker, en *meddelelsesfrihet*. Den innebærer altså også en rett til å *motta* informasjon, ofte kalt *informasjonsfrihet*. Den som vil skaffe seg informasjon gjennom andres ytringer, har rett til å gjøre det. Det er ikke anledning til å avskjære noen fra dette, for eksempel ut fra betraktninger om informasjonens innhold eller kvalitet. Ytringsfriheten gjelder for alle typer informasjon og ideer, inkludert de som måtte sjokkere, krenke eller være usanne.[59] Informasjonsfriheten omfatter også usann informasjon. Den gjelder dermed uavhengig av om innholdet i en ytring er sant eller falskt.[60] Menneskerettighetene gir rett til å uttrykke meninger og påstander som ikke kan underbygges, og til å hengi seg til kunst, satire eller ironi.

Likevel har det vært vanlig å oppfatte informasjon vernet av ytringsfriheten som meningsbærende ytringer fra mennesker, og ikke for eksempel setninger som er skapt gjennom teknologiske sannsynlighetsberegninger.

Informasjon generert av kunstig intelligens kan ikke stoles på bare fordi den er sannsynlig.[61] Etterhvert som generativ kunstig intelligens blir stadig mer sofistikert, vil det bli vanskeligere å skjelne denne grunnleggende forskjellen

[57] Smith 2019
[58] Pew Research Center 2023
[59] FNs menneskerettighetskomité 2011, para 11, EMD *Handyside v. UK* (5493/72), para 49
[60] FNs menneskerettighetskomité 2011, para 47 og 49, EMD *Salov v. Ukraine* (65518/01), para. 113
[61] Jungherr 2023

mellom ytringer produsert av menneske og maskin. Likevel vil sannsynlighetsrelasjonen som maskinene produserer, bestå bare mellom ordene i de ytringene som genereres, og ikke mellom ordene og virkeligheten.



Ogdens trekant, eller den semiotiske trekanten, illustrerer forholdet mellom termer, tegn eller ord i det ene hjørnet, de begrepene vi tillegger «mening» i et annet hjørne, og det som ordene og begrepene refererer til i den virkelige verden i det tredje. For eksempel ordet «papegøye», vår oppfatning av hva en papegøye er, og en faktisk papegøye. Kunstig intelligens sier ikke noe direkte om relasjonen i trekanten. I motsetning til mennesker opererer kunstig intelligens bare i det ene av disse tre hjørnene.

INDIVIDETS FRIE MENINGSDANNELSE

Ytringsfriheten er en grunnforutsetning for individets frie meningsdannelse. Vi blir til som selvstendige mennesker i dialog og samspill med andre. Det er gjennom denne dialogen mellom oss og verden rundt at vi erkjenner hva vi er og står for. Underveis i denne dialogen vil vi nødvendigvis gjøre feil, og det må de andre tåle. Toleranse er en forutsetning for dette samspillet mellom den enkelte og de andre.

Menneskeretten til fri meningsdannelse, altså den tankefriheten som ligger forut for eller innenfor ytringsfriheten, er absolutt og kan ikke gjøres inngrep i. Likevel blir mennesker i virkeligheten konstant påvirket av andres tanker og meninger, og «friheten til å være underkastet en lang rekke ulike påvirkningsfaktorer» er i seg selv en del av vår menneskelige autonomi.[62] Hva skjer når *de andre* er maskiner? Hvor fri blir egentlig meningsdannelsen om det vi mottar av informasjon styres og påvirkes av algoritmer, med de usynlige og ugjennomtrengelige påvirkningsmuligheter de gir?

FNs spesialrapportør for ytringsfrihet, David Kaye, beskrev i 2018 hvordan kunstig intelligens «former informasjonsverdenen på en måte som er utilgjengelig for brukeren», og skjuler sin egen rolle i å avgjøre hva informasjonsbrukeren ser og forbruker.[63] Samtidig skreddersys informasjon til hver person på måter som forsterker våre kognitive feilslutninger eller *bias* og «insentiverer spredning og anbefaling av oppildnende innhold og desinformasjon for å holde på brukerens online-oppmerksomhet».[64] I sin tur påvirker denne praksisen individers selvbestemmelse og autonomi til å skape og utvikle personlige meninger basert på faktisk og variert informasjon. Derigjennom trues både informasjons-, ytrings- og tankefriheten.

Når generativ kunstig intelligens tar et teknologisk steg videre, gir det et knippe utfordringer for grunnlaget for individets frie meningsdannelse. Modellene blir i stand til å produsere feil, villedende eller manipulert informasjon i et hittil ukjent omfang.

De tause ytringene og anonymitet

Retten til fri meningsdannelse inkluderer retten til å ikke uttrykke en mening, og retten til å endre mening, når som helst og av hvilken som helst grunn.[65]

[62] Jones 2019
[63] Kaye 2018
[64] Kaye 2018, para 12
[65] FNs menneskerettighetskomité 2011, para 9

Det utgjør et brudd på meningsfriheten å straffe, trakassere, intimidere eller stigmatisere noen for å ha en mening, inkludert tvungen, ufrivillig eller ikke-samtykkende manipulering av tankeprosessen for å utvikle meninger.[66]

Ytringsfriheten innebærer derfor også en rett til å forholde seg *taus*, hvis man ønsker det. En slik negativ ytringsfrihet har sammenheng med meningsfriheten, som er absolutt. Man kan tro på hva man vil, og behøver ikke å fortelle andre hva man tror eller ikke tror på. Mennesket har mental autonomi og skal ikke ufrivillig måtte avsløre sine tanker og meninger. Å presse mennesker til å ha eller ikke ha bestemte meninger, er forbudt.[67]

Maskingenererte ytringer er derimot en representasjon av allerede ytrede meninger. Taushet kan være en ytring, men taushet innebærer at ens meninger ikke fanges opp av maskinlæring. Grunnlaget for maskiners ytringsproduksjon vil da ikke være representativ for det meningsmangfold som faktisk finnes.

Den motsatte utfordring er at taushet kan registreres og tillegges mening. Mange eksperter regner med at de digitale systemene med tiden vil bli «allestedsnærværende» og for viktige til at borgerne kan unngå dem. Dermed vil «alle brukere bli fanget», og alle bevegelser og uttrykk, eller mangel på sådan, vil bli tillagt mening.[68]

Presset øker også mot ytringer med skjult opphav, og gjør at anonymitet og kildevern kommer i spill. Medienes kildevern er historisk forankret i et anonymitetsforbud. Alle som skulle forfatte noe, måtte gjøre det under fullt navn. Dette ble endret til en anonymitetsrett da Norge fikk sin grunnlov i 1814. Dette utviklet seg senere til et kildevern – en viktig bestanddel av ytringsfriheten.

Den digitale offentligheten har behov for både anonymitetsforbud og anonymitetsrett. Forskjellige pålegg om merking av menneskelig opprinnelse kan styrke tilliten til informasjon, og er en form for anonymitetsforbud.

[66] Aswad 2020, s. 329
[67] Khan 2021, para 25
[68] Pew Research Center 2023

Generativ kunstig intelligens styrker altså behovet for identifisering av informasjonens opphav, mens behovet for kildevern består. Samtidig kan teknologiutviklingen vanskeliggjøre begge deler.

Maskiner med meninger

Modellene som genererer ny informasjon basert på allerede eksisterende data, er ikke nøytrale. Innhold generert av kunstig intelligens baserer seg på gjennomsnitt. En studie fra Cornell University indikerer at språkmodeller som regel ikke representerer flertallssynpunkter, men holdninger og meninger til mennesker med høyere utdanning.[69] Selv om ChatGPT i teorien skal være politisk nøytral, har forskere i Storbritannia konkludert med at chatboten svarer i tråd med liberale politiske verdier.[70] Et forskningsprosjekt på hele 14 språkmodeller viser lignende trekk – ChatGPT lener mot venstresiden, mens Metas språkmodell LLama lener mot høyre.[71] Funnene har bidratt til at høyrekonservative kritiserer både ChatGPT og Googles Bard for å være såkalt *woke AI*.[72] OpenAI, derimot, har uttalt at politisk bias i ChatGPT handler om systemiske feil, snarere enn en politisk vinkling ved modellen.[73]

Innebygde politiske holdninger i store språkmodeller kan være en utfordring etterhvert som modellene blir stadig mer utbredt blant folk. Menneskers holdninger kan påvirkes av maskinenes innebygde bias og politiske overbevisninger, og påvirkningen kan være både skjult og synlig. En test av 1500 mennesker som hadde dialog med forhåndsprogrammerte chatboter, viste at holdninger ble utsatt for skjult påvirkning (på engelsk *latent persuasion)* gjennom dialogen med modellene.[74] En studie fra Stanford viser også at språkmodeller har en tendens til å fremme kontroversielle temaer, heller enn flertallssynspunkter i befolkningen.[75]

Dermed må modellene antas å gi flere utfordringer knyttet til fri meningsdannelse. Sårbare grupper vil kunne underrepresenteres. Minoriteter

[69] Santurkar m.fl. 2023
[70] University of East Anglia 2023
[71] De Vynk 2023
[72] Tiku m.fl. 2023
[73] OpenAI 2023d
[74] Jakesch m.fl. 2023
[75] Myers 2023

som tradisjonelt ikke er representert i datasett vil forbli usynlige for maskinenes øyne og ører.[76] Generativ kunstig intelligens kan slik påvirke vår evne til å «se oss selv utenfra», utarme demokratiets evne til å prosessere viktig og relevant kritisk informasjon og dermed styrke den politiske status quo.[77] En effekt kan dermed bli at allerede utsatte grupper med svak representasjon vil svekkes ytterligere, mens sterkere gruppers meninger og holdninger vil få uforholdsmessig stor plass.

Informasjon og analyser som produseres av kunstig intelligens har dermed innebygde føringer, premisser og fordommer, og vil i større grad bidra til «bundet meningsdannelse» enn den frie meningsdannelsen som ytringsfriheten er begrunnet i.

DEMOKRATI

Ytringsfriheten er nødvendig for et fungerende *demokrati*. Det er umulig å tenke seg reelle folkestyrer uten at folket kan tilegne seg kunnskap, og bearbeide og kritisere det som foregår. For at vi skal kunne velge de beste løsningene, må vi ha noe å velge mellom. Da må alle ideer kunne bringes til torgs, også de dårlige eller de vi misliker. Demokratiet er ideenes markedsplass. På denne markedsplassen må noen kjøpe ideer. Bare dersom utvalget er bredt, kan alle finne noe som passer dem, og som i sum og på sikt kan passe og gagne samfunnet. Valgfriheten er mer reell, jo mer man har å velge mellom. Men er det maskinene eller teknologiselskaper (eller staten) som kontrollerer maskinene, som skal stå for utvelgelsen? Hvilken rolle skal vi gi maskinene i vårt folkestyre?

Kunstig intelligens kan påvirke demokratiet gjennom å målstyre, manipulere, sensurere eller dikte opp ideer. Men også den blotte muligheten for slik påvirkning har en effekt, fordi det påvirker velgeres forestillingsevne, og deres frykt for eller ønske om å la seg informere, og selv delta i det demokratiske ordskiftet.

[76] Buolamwini m.fl. 2018
[77] Jungherr 2023

Den antatte bruken av databasert psykologisk målrettet informasjon fra Cambridge Analytica i 2016 både i Brexit og ved USAs presidentvalg er et illustrerende eksempel. Selv om det ikke finnes tekniske bevis for at bruken hadde avgjørende effekt på folkeavstemningen eller valget, har disse episodene festet seg i den offentlige psyken og hukommelsen, og anses gjerne som et utslag av hvordan man forestiller seg kunstig intelligens brukt i valgmanipulasjon.[78]

Etter hvert som kunstig intelligens brukes i stadig mer utstrakt grad i folks økonomiske, politiske og sosiale liv, vil trolig publikums forventninger om bruk og misbruk i demokratiske valg øke, uavhengig av hvordan kunstig intelligens faktisk brukes eller kan brukes.[79] Allerede kan mulighetene som kunstig intelligens fører med seg, svekke folks tillitt til informasjon de mottar.

En annen påregnelig følge av kunstig intelligens er økt bruk av målrettet desinformasjon med formål om å villede publikum, eller at informasjonsmiljøet kan bli overlesset av upålitelig eller villedende kunstig generert innhold.

I sum vil denne utviklingen svekke kraften til hele informasjonsdomenet. Det vil gjøre folks tilgang til viktig informasjon vanskeligere og/eller gjøre at sann informasjon vil fremstå som upålitelig. En antatt effekt av kunstig intelligens-revolusjonen er derfor at menneskers tillit både til institusjoner og til hverandre vil bli lavere. Dette kan i sin tur forverre en allerede tiltakende polarisering, øke kognitiv dissonans i ordskiftet og føre til at folk trekker seg ytterligere fra offentlig meningsutveksling og demokratisk deltakelse.[80]

Heller ikke ytringsfrihetens begrunnelse i demokrati ser ut til å omfatte maskinskapte ytringer eller passe for den måten generativ kunstig intelligens trolig vil virke i det offentlige rom.

[78] Jungherr 2023
[79] Jungherr 2023
[80] Pew Research Center 2023

TOLERANSE OG MANGFOLD

Toleranse er nødvendig for den dialogen som ytringsfriheten forutsetter. Mangfold og pluralisme er innarbeidede begreper i rettspraksis fra Den europeiske menneskerettsdomstolen (EMD), som har fremhevet at «pluralism, tolerance and broadmindedness» er avgjørende for et demokratisk samfunn.[81]

Ytringsfrihetskommisjonen av 1999 mente at jo mer det offentlige roms institusjoner ble bygd ut, desto mindre behov ville det være for reguleringer.[82] Meningsbrytningen ville i seg selv lede til en økt grad av sannhetssøken, demokrati og individuell utvikling.

Ytringsfrihetskommisjonen av 2022 så mindre lyst på dette. Etter å ha vurdert virkningene av en digital offentlighet der ytringsrommet i stor grad er styrt av plattformer som verken staten eller norske borgere har eierrådighet over, kom de nærmest fram til motsatt konklusjon: Fravær av reguleringer kan nå utfordre de verdiene som ytringsfriheten er ment å verne. Økt grad av desinformasjon leder ikke til sannhetssøken, algoritmer som presenterer det man allerede er enig i styrker ikke personlige autonomi, og utenlandske trollfabrikker underbygger ikke demokratiet. Ytringsfrihetskommisjonen konkluderte derfor med at ytringsfrihetens begrunnelse krever aktive grep og reguleringer.[83]

Fremvekst av generativ kunstig intelligens forsterker Ytringsfrihetskommisjonens analyse av den nye digitale offentlighetens negative virkninger. Når ytringer er frembragt av generativ kunstig intelligens svikter dessuten mye av formålet med toleranse: Toleranse utviser vi overfor våre medmennesker fordi det kan vise seg at det er de som har rett og ikke vi, og fordi de skal våge å komme med sine djerve påstander og gjetninger.

Verken ytringsfrihetens tre klassiske begrunnelser om sannhetssøken, individets frie meningsdannelse og demokrati, eller de nye hensynene til toleranse og mangfold, ser ut til å passe godt på maskinskapte ytringer.

[81] EMD NIT S.R.L. mot Moldova (28470/12)
[82] NOU 1999:27 kap 2.2.4
[83] NOU 2022:9, kap 3.5.3

Ettersom en juridisk regel er tett knyttet til regelens begrunnelser, oppstår spørsmålet om den juridiske ytringsfriheten slik den er utformet og fungerer i vårt demokratiske samfunn egentlig dekker maskinskapte ytringer.

SPØRSMÅL FOR YTRINGSFRIHETEN

HAR MASKINER EN «RETT» TIL Å YTRE SEG?

Ytringsfriheten slik den er inkorporert i Den norske Grunnloven og menneskerettighetstraktatene er en *menneske*rettighet. Det er rettigheter som *ipso facto* tilkommer mennesker *i kraft av å være mennesker*. Båter eller biler har ikke menneskerettigheter. De vil kun være beskyttet av dette rettsregimet i den grad deres eier eller bruker er et menneske som har slike rettigheter. Som utgangspunkt kan dermed datamaskiner ikke påberope seg menneskerettigheter. Dersom generativ kunstig intelligens ytrer seg på måter som bryter loven, vil det være menneskene bak modellen, altså enten menneskene som har programmert modellen (selskapet), eller mennesket som bruker modellen, som må svare for bruddet.

Problemstillingen er satt på spissen i en pågående amerikansk rettssak mot OpenAI. En journalist bad ChatGPT om å lage et sammendrag av en pågående rettssak. En mann ved navn Mark Walters ble implisert i saken av ChatGPT. Mannen hadde ingen tilknytning til saken, og alle påstander om ham ble diktet opp av chatboten. Journalisten kjente selv til Walters, og forsøkte derfor å korrigere ChatGPT. Plattformen svarte med å dikte opp et dokument som skulle underbygge påstanden om at Walters var implisert. Av rettssakens dokumenter fremgår det at OpenAI er klar over at ChatGPT kan hallusinere og dikte opp faktum.[84]

Rettssaken gir opphav til en rekke prinsipielle spørsmål: Kan en maskin opptre urimelig, eller er det skaperen av maskinen, altså selskapet som er ansvarlig? Hvem «måtte vite bedre»? Kan man tillegge en mental tilstand til en maskin?

[84] Walters v. OpenAI 2023

Eller må dette legges på selskapet som har skapt og designet systemet? Skal feilprogrammering av kunstig intelligens sidestilles med villedende utformede varer? Både pliktene og rettighetene i saken tilfaller rettssubjektene mennesker eller selskaper, og ikke ChatGPT.

I denne dommen blir et viktig premiss klargjort – det er ikke det generative kunstig intelligens-systemet som har ansvaret for de ytringene som er fremsatt, men derimot menneskene som benytter seg av tjenester som ChatGPT som et instrument for å uttrykke seg. Det samme vil gjelde for ytringsfriheten. Det er menneskene og ikke maskinene som er bærere av retten til å ytre seg.

HAR MASKINGENERERTE YTRINGER SAMME BESKYTTELSE SOM MENNESKELIGE YTRINGER?

Kunstig intelligente datamaskiner kan ikke selv påberope seg menneskerettigheter eller ytringsfrihet per i dag. Likevel kan ytringer som er generert av maskiner være beskyttet av ytringsfriheten. Dersom en person tar i bruk generative tjenester og verktøy for å uttrykke seg, vil personen fortsatt være ansvarlig for eventuelle ytringer, meninger og holdninger som modellen genererer og sprer på hens oppdrag. Den maskingenererte ytringen vil dermed være beskyttet av ytringsfriheten gjennom å representere en ytring fra personen.

Menneskers ytringsfrihet vil beskytte ytringer som opprinnelig er generert av maskiner i tilfeller der kunstig intelligens benyttes som plattform og instrument for menneskeskapte ytringer, eller der mennesker og maskiner har *samarbeidet* om en ytring. I slike tilfeller er maskinen et villet verktøy for menneskers ytringsfrihet.

Også menneskeretten til informasjon vil favne om maskingenererte ytringer. Retten til å motta informasjon og retten til å videreformidle informasjon er ikke avhengig av informasjonens sannhetsgehalt eller opphav. Maskingenerert innhold vil i økende grad inngå i informasjonstilfanget vårt, og det er ikke et vilkår for informasjonsfriheten at man undersøker informasjonens opphav eller sannhetsgehalt. Informasjons- og meningsfriheten innebærer også en rett til å la seg forlede eller forvirre.

Menneskeretten til informasjons-, menings- og ytringsfrihet vil derfor i utstrakt grad gjelde også for maskingenerert innhold.

ER YTRINGSFRIHET AVHENGIG AV MENNESKELIG INVOLVERING?

Menneskeskapte ytringer er fullt beskyttet av ytringsfriheten. Maskinskapt innhold og ytringer vil dekkes av ytringsfriheten i den grad den tjener eller betinger menneskers ytringsfrihet, altså et slags avledet vern. Rene maskingenererte ytringer vil derimot ikke være beskyttet. Dermed oppstår betydelige avgrensningsutfordringer.

Hvordan skal grensene trekkes mellom menneskeskapte ytringer og maskinskapte ytringer? Selv om ytringsfriheten ikke dekker maskinskapte ytringer, vil den likevel gjelde for ytringer hvor mennesker har benyttet skapende kunstig intelligens til å raffinere, klargjøre, spisse eller ironisere over menneskeskapt innhold. Kunstig generert innhold er også i seg selv en reproduksjon av tidligere menneskeskapt innhold. Ytringsfriheten skiller ikke mellom hvem som skaper innhold, hvem som sprer innholdet eller hvem som justerer det, nettopp fordi det vil gi umulige avgrensningsutfordringer.

KAN INNGREP I YTRINGSFRIHETEN LEGITIMERES FORDI YTRINGER ER MASKINSKAPT?

Gjennomgangen over kan tyde på at ytringsfrihetens begrunnelser ikke på samme måte gir grunnlag for beskyttelse av maskinskapte ytringer. Det er maskingenerert informasjons betydning for menneskers rett til ytringsfrihet som gjør at den vil være vernet av ytringsfriheten. Snarere kan maskinskapt innhold true den tradisjonelle ytringsfriheten på måter som gjør at inngrep og begrensninger vil være på sin plass.

Gitt ytringsfrihetens grunnleggende viktighet for alle andre menneskerettigheter, skal begrensninger i ytringsfriheten kun gjøres

unntaksvis og være snevre.[85] Inngrep i ytringsfriheten må være hjemlet i lov, forfølge et legitimt formål og være nødvendige og proporsjonale.[86]

Under Den europeiske menneskerettskonvensjonen (EMK) er det et vilkår at inngrep i ytringsfriheten må være «nødvendig i et demokratisk samfunn».[87] Å forby usann eller falsk informasjon er ikke en legitim grunn for å begrense ytringsfriheten under internasjonale menneskerettigheter.[88]

De strenge inngrepsreglene er likevel basert på forutsetningen om at det er *mennesker* som står bak den usanne, uriktige, upresise eller bevisst villedende informasjonen. Løgn og feilslutninger er uønskede, men likevel nødvendige ingredienser for ytringsfrihetens begrunnelser – sannhetssøken, individets frie meningsdannelse, mangfold og demokrati.

Når ytringer er maskingenerte, vil de samme strenge reglene mot inngrep trolig ikke gjelde. Skapende kunstig intelligens i enkelte former og i enkelte sammenhenger vil trolig selv utgjøre en trussel mot ytringsfriheten i tradisjonell forstand på måter som gir myndighetene rett – eller sågar plikt – til å begrense den for å bevare norske borgeres rett til informasjons-, menings- og ytringsfrihet.

Likevel kan slike inngrep og begrensninger på maskinskapte ytringer også *true* ytringsfriheten, gjennom å gi stater en utvidet mulighet til å forhåndssensurere, fjerne og forby kunstig generert informasjon basert på hva myndighetene anser for å være sant. En vid adgang til inngrep i maskinskapte ytringer kan også gå på bekostning av retten til å skaffe seg informasjon. Det kan også øke risikoen for autoritære trekk i samfunnet. Stater kan for eksempel argumentere for at økt overvåkning og datainnsamling i det digitale er nødvendig som ledd i å få kontroll på kunstig generert usann informasjon Det kan i sin tur gå på bekostning av andre rettigheter som for eksempel retten til personvern, eller adgangen til å føre de bevisene man selv ønsker i en rettssak. Ytringsfrihetenes

[85] Khan 2021, para 40
[86] SP-konvensjonen 19(3) og EMK 10 (2)
[87] EMK 10(2)
[88] Khan 2021

begrunnelser kan altså tale både for og mot inngrep. Kunstig intelligens skaper med andre ord *dilemma* som må finne sin løsning.

KAN MASKINSKAPT DESINFORMASJON FORBYS ELLER BEGRENSES UTEN Å INNEBÆRE BRUDD PÅ YTRINGSFRIHETEN?

En betydelig utfordring knyttet til maskingenerert innhold handler om desinformasjon. Begrepet betegner falsk eller manipulert informasjon som har til hensikt å forårsake skade i form av forvirring, underminering av forståelse eller tillit. Desinformasjon kan ta form av avanserte forfalskninger, propaganda eller falske nyheter. Hensikten kan være politisk, økonomisk, ideologisk eller sikkerhetspolitisk. Desinformasjon skiller seg fra misinformasjon (eller feilinformasjon) som er ukorrekt informasjon, men som ikke er laget eller spredt i den hensikt å skade.

Desinformasjon er et urgammelt fenomen. Ytringer og informasjon blir spisset, tilpasset, vektet, manipulert eller fabrikkert for å starte eller vinne kriger, hevne seg, skade noen eller tjene penger. Utgangspunktet er at retten til ytringsfrihet under EMK artikkel 10 og SP-konvensjonen artikkel 19 gjør at man *ikke* kan forby diskusjon om eller spredning av informasjon selv om man har sterk mistanke til at informasjonen er usann.[89] Den digitale utviklingen har imidlertid revolusjonert desinformasjon. Teknologi gjør det nå mulig å skape eller manipulere falsk informasjon på nye måter, og informasjonen kan spres, mangedobles og utnyttes i politiske, ideologiske eller kommersielle hensikter i et helt nytt omfang, tempo og rekkevidde. Da oppstår spørsmålet om de teknologiske endringene vi står overfor gjør at desinformasjon i større grad bør kunne forbys.

Hensikten med digital desinformasjon er gjerne å svekke informasjonsdomenet og undergrave hele rommet for ytringsfriheten, og behovet for inngrep for å begrense skadeeffektene vil bli betydelig. Gir ytringsfriheten hjemmel for inngrep for å hindre eller begrense maskingenerert desinformasjon? Siden ytringsfrihet ikke gir samme beskyttelse til maskingenerert innhold som til

[89] EMD *Salov v. Ukraine* (65518/01), para 113, FNs menneskerettighetskomité 2011, para 47 og 49

menneskeskapte ytringer, vil utgangspunktet være at myndighetene kan gripe inn med tiltak for å begrense eller stanse maskingenerert desinformasjon.

Likevel støter vi raskt på krevende avgrensningsproblemer. Hvordan skal grensen trekkes mellom desinformasjon som er maskingenerert og desinformasjon som er menneskeskapt? Etter hvert som informasjonsdomenet fylles av informasjon som i varierende grad bygger på maskingenerert innhold, vil grensen bli svært vanskelig å identifisere. Dermed vil fokus i større grad være på desinformasjonens *innhold*.

Desinformasjon som består av ulovlige ytringer, må skilles fra desinformasjon som består av ytringer som ikke er ulovlige.

Ulovlige ytringer

Enkelte typer desinformasjon kan staten være pliktig til å forby. Det er forbud mot desinformasjon som oppfordrer til hat, diskriminering eller vold.

SP-konvensjonen artikkel 20(2) knesetter at å fremme nasjonalt, rasemessig eller religiøst hat som utgjør oppfordring til diskriminering, fiendtlighet eller vold skal forbys ved lov, uten at usann informasjon er særlig nevnt. Rabat-handlingsplanen gir et veikart for å tolke artikkel 20(2) og lister opp seks faktorer for å avgjøre alvoret: kontekst, talerens status, hensikt, innhold og form, rekkevidde og sannsynlighet for risiko. Disse elementene har også blitt fremhevet som relevante for å vurdere tiltak mot desinformasjon mer generelt.[90]

Hatefulle utsagn om rasemessig eller etnisk opphav er også forbudt under Rasediskrimineringskonvensjonen artikkel 4. Likevel har både FNs menneskerettighetskomité og rasediskrimineringskomité vært klare på at forbud mot slike uttrykk kun må foretas i overenstemmelse med retten til ytringsfrihet i SP-konvensjonen artikkel 19, og at kriminalisering bør begrenses til de mest alvorlige utsagn.

[90] FNs menneskerettighetsråd 2013

Ulovlige utsagn vil rammes av forbud på vanlig måte, uavhengig av om de er generert av mennesker eller maskiner, ettersom det er *virkningen* av informasjonen som er forbudt, ikke hvorvidt informasjonen er sann eller usann.

Ytringer som ikke er ulovlige

Maskingenerert desinformasjon med innhold som ikke er forbudt kommer i en annen kategori. Det er ofte vanskelig å trekke klare grenser mellom sann og falsk informasjon og mellom ytringer som har skadehensikt og ytringer som ikke har det. Falsk informasjon kan bli et redskap for aktører med motstridende målsetninger. Faktum kan bli stemplet som falske nyheter og bli delegitimert. Meninger, oppfatninger, usikker viten og andre typer uttrykk som ironi, satire eller parodier er også vanskelige å klassifisere som sanne eller usanne.

På nettet kan innhold som er distribuert i skadehensikt (desinformasjon) bli plukket opp og delt videre av intetanende tredjeparter som ikke har slik hensikt (da blir det feilinformasjon). Den uvitende som deler, bidrar til å spre den uriktige informasjonen, og kan også øke troverdigheten til informasjon som er feil. Skadeomfanget tiltar selv om skadehensikten på dette tidspunktet er borte.

Når man omtaler desinformasjon, omfatter det normalt ikke informasjon som er sann. Likevel kan også sann informasjon deles med skadehensikt (på engelsk *mal-information)*. Det kan dreie seg om særlig privat informasjon som lekkes i skadehensikt, eller informasjon som bevisst er tatt ut av sin sammenheng eller sammenstilt for å øke emosjonelle reaksjoner.

Den skiftende karakteren til informasjon gjør det svært utfordrende å trekke objektive grenser for når ellers lovlige ytringer som desinformerer eventuelt skal kunne forbys. Inngrep i ytringsfriheten for å begrense desinformasjon kan derfor lett komme i skade for å redusere den ytringsfriheten inngrepene er ment å verne.

De fremskritt som kunstig intelligens bringer med seg for overvåkningsteknologi, avanserte nettroboter i private rom, spredningen av desinformasjon og dype forfalskninger, samt avansert ansiktsgjenkjennelsesteknologi, utgjør betydelige farer og utfordringer for

menneskerettighetene.[91] I en undersøkelse fra 2023, utført av det amerikanske instituttet Pew Research Center, frykter ekspertene at med stadig mer utbredt bruk av kunstig intelligens, vil retten til privatliv bli vanskelig – om ikke umulig – å opprettholde. I en slik virkelighet blir en reell og uinnskrenket ytringsfrihet enda viktigere som en rettighet som bærer andre menneskerettigheter. Kunstig intelligens gjør derfor at ytringsfriheten i tradisjonell forstand snarere bør styrkes.

[91] Pew Research Center 2023

HVA KAN GJØRES?

Utbredelsen av skaperkraftige maskiner får store konsekvenser for ytringsrommet og ytringsfriheten. Hvordan kan tilliten til digital informasjon sikres når økte mengder kunstig generert innhold på nett endrer den digitale offentligheten?

Denne rapporten har analysert effekter av gjennombruddet for generativ kunstig intelligens, og viser at teknologien vil kunne påvirke informasjonsdomenet og ytringsrommet på måter som kan true ytringsfriheten. Alle får nå tilgang til verktøy for å ytre seg, men samtidig kan verktøyene brukes kan til å spre falskt, manipulerende og villede informasjon, enklere, fortere og i et større omfang enn før.

Gjennom infrastrukturkravet i Grunnloven er myndighetene pliktige til å tilrettelegge for en åpen og opplyst samtale. Men selv om maskingenerert informasjon kan true hele informasjonsdomenet, er det likevel klare grenser for myndighetenes adgang til å gripe inn for å kontrollere og regulere digitale ytringer. Menneskeretten til informasjons-, menings- og ytringsfrihet vil i utstrakt grad også gjelde for maskingenerert innhold, fordi maskinene fungerer som et verktøy for menneskers ytringer og utfoldelse. Etter hvert som kunstig generert innhold på nett endrer den digitale offentligheten, kreves nytenkning rundt hvordan tilliten til digital informasjon skal sikres.

Å beskytte det digitale ytringsrommet fremover uten å ødelegge friheten som menneskeskapte ytringer betinger, vil kreve veloverveide tiltak og løsninger av både teknologisk, politisk og regulatorisk art.

UTVIKLE LØSNINGER FOR VERIFISERING

En forutsetning for en opplyst offentlig samtale er at mennesker klarer å skille mellom menneskeskapt og maskinskapt innhold, og at det fremgår hvem avsenderen er. Dette utfordres nå som mengden kunstig generert innhold øker.

Tekniske løsninger som *vannmerking* er én av flere metoder som kan gjøre det å verifisere innhold lettere. Flere tekniske løsninger for å merke og detektere kunstig generert innhold, og for å beskytte digitalt medieinnhold fra å bli manipulert, er nå under utvikling. Slike verktøy kan for eksempel hjelpe til med å kvalifisere medieinnholdets opphav eller informere om hvem avsenderen er.

Bildegenereringsprogrammet DALL-E bruker en synlig digital vannmerking på sine bilder, for å beskytte bilder og videoer med opphavsrett fra misbruk. Det utvikles også løsninger for vannmerking som er usynlige, ved bruk av metadata eller koder. Google DeepMind har lansert et verktøy som modererer piksler i bilder. Disse er ikke synlige og – ifølge selskapet selv – forsvinner de ikke dersom man redigerer eller endrer bildet.[92]

Digitalt innhold kan også aktivt beskyttes mot manipulasjon. Forskere ved MIT har laget et verktøy kalt PhotoGuard som skal gjøre det vanskeligere å manipulere bilder. Små, usynlige koder legges på bildet og gjør det vanskelig for kunstig intelligens å tolke og manipulere det.[93] PhotoGuard kan likevel omgås ved enkle grep som å ta en skjermdump av bildet, noe som fjerner de integrerte kodene.[94]

[92] Heikkilä 2023a
[93] Salman m.fl. 2023
[94] Heikkilä 2023b

Behovet øker også for teknologi som kan fange opp og identifisere innhold som er kunstig generert. Forskere i USA har utviklet en algoritme som kan oppdage tekst generert av Metas språkmodeller.[95] Å oppdage generert tekst har likevel vist seg å være utfordrende. OpenAI, selskapet bak ChatGPT, har selv avviklet et verktøy for å oppdage tekst generert av kunstig intelligens, fordi treffsikkerheten var for lav.[96] Feilaktige anklager om generert tekst er også en utfordring. Detekteringsverktøy kan ha en innebygd skjevhet som gjør det vanskelig å kategorisere tekst som kunstig dersom teksten er skrevet av mennesker som ikke har engelsk som morsmål.[97]

Samtidig som selskapene nå utvikler og tar i bruk nye generative verktøy for å utvikle og forbedre tjenestene sine, må de også sikre at bruken av disse reguleres i tråd med gjeldende regler og standarder. Teknologiselskaper er de som enklest kan ta initiativ til å merke innhold på sine plattformer.

I USA har teknologiselskaper som Google, OpenAI, Meta og Microsoft sluttet seg til President Joe Bidens frivillige erklæring for ansvarlig kunstig intelligens.[98] Selskapene forplikter seg til å utvikle tekniske løsninger og vannmerke kunstig generert innhold. Selskapene har imidlertid ikke forpliktet seg til å ta i bruk disse løsningene selv.

YouTube har utviklet egne retningslinjer som inkluderer at brukere skal få vite om innhold på plattformen er generert av kunstig intelligens. Dette innebærer krav til brukere som laster opp videoer, og automatisk merking av kunstig genererte videoer. Dersom genereringsverktøy tas i bruk i videoer som tar for seg sensitive tema, som valg, konflikter og helserelaterte spørsmål, vil disse få en særskilt synlig merking.[99]

Også i Norge finnes det initiativ for å utvikle tekniske løsninger for merking og verifisering, som prosjektet C2PA/Project Origin. Der deltar Media City Bergen sammen med en rekke store, internasjonale nyhetsorganisasjoner og

[95] Kirchenbauer m.fl. 2023
[96] David 2023
[97] Liang m.fl. 2023
[98] The White House 2023a
[99] Flannery O'Connor m.fl. 2023

teknologibedrifter.[100] Formålet med prosjektet er å utvikle en standard for å autentisere innhold ved hjelp av krypterte metadata. I forlengelsen av dette leder Media City Bergen et norsk prosjekt, kalt Prosjekt Reynir, som skal videreutvikle denne teknologien, med mål om å skape en nasjonal standard.[101]

I Norge har flere medie- og nyhetsbyråer utviklet interne retningslinjer for bruk av teknologien, som å merke medieinnhold der genereringsverktøy har blitt tatt i bruk. NRK og Dagens Næringsliv – i tillegg til et stort flertall medieinstitusjoner – har for eksempel innført retningslinjer som innebærer at uredigert tekst skrevet at generativ kunstig intelligens ikke vil bli publisert. Hvis språkmodeller som ChatGPT blir brukt til å skrive tekst, må teksten merkes.[102]

Det gjenstår likevel mange ubesvarte spørsmål: Hva skal merkes, hvordan, og hva slags informasjon skal oppgis? Det er også uklart hvem som kan stilles ansvarlig for at merkingen fungerer, og for at den ikke misbrukes. Her må flere løsninger utvikles og fungere side om side. Én enkelt løsning, slik som vannmerking, vil trolig ikke være tilstrekkelig for å motvirke de store mengdene desinformasjon som eksempelvis kan føre til påvirkning og manipulering av valg.[103]

MYNDIGHETENE KAN INNFØRE KRAV OM VANNMERKING

Det pågår nå politiske diskusjoner om hvordan merking av kunstig generert innhold og bruk av kunstig intelligens kan og bør utvikles og innføres, både i Europa og USA. Nasjonale standarder i andre land kan legge føringer for det offentlige digitale rom også i Norge.

I oktober 2023 lanserte president Joe Biden en presidentordre med omfattende pålegg om tiltak for trygg og pålitelig bruk og utvikling av kunstig intelligens.[104]

[100] Project Origin 2023
[101] Prosjekt Reynir 2023
[102] Se for eksempel Greger 2022, Dagens Næringsliv 2023
[103] Elliott 2023
[104] The White House 2023b

Et av tiltakene er å utvikle standarder for merking og verifisering av kunstig generert innhold for å beskytte amerikanere mot manipulering og svindel. Blant annet skal offentlig informasjon merkes for å gjøre det enklere for amerikanere å vite at de mottar informasjon fra myndighetene, og for å sette en global standard for myndigheter verden rundt.[105]

I EU vil merking av kunstig generert innhold på sikt bli regulert i forordningen om kunstig intelligens (på engelsk *Artificial Intelligence Act*, *AI Act*). Loven skal gjøre det trygt og lett å utvikle og ta i bruk kunstig intelligens i Europa.[106] Forhandlingene ble sluttført 8. desember 2023, men detaljene blir først kjent når lovteksten publiseres i sin helhet.[107]

Lovforslaget ble til før generativ kunstig intelligens og store språkmodeller som ChatGPT fikk sitt gjennombrudd, og var derfor ikke utformet til å regulere og håndtere kunstig generert innhold eller underliggende språkmodeller. I sluttforhandlingene ble spørsmålet om hvordan språkmodeller og generative tjenester skal reguleres blant de mest krevende.[108]

Krav til gjennomsiktighet rundt bruk av kunstig intelligens var imidlertid del av EU-kommisjonens opprinnelige lovforslag. Det inneholdt krav om at mennesker skal bli informert dersom de chatter med kunstig intelligens, og at alle som står bak dype forfalskninger må opplyse om at de er kunstig fremstilt.

EU-parlamentet ønsket å legge til krav om at de som lager dype forfalskninger må stå frem med navn, og at innholdet merkes i tråd med gjeldende standarder.[109] Flere av parlamentets forslag om merking var likevel vage og uklare. Kravene skulle ikke gjelde dersom de hindret utøvelse av ytringsfriheten, og dersom produksjonen av innholdet inngikk i en kreativ prosess, skulle det holde å oppgi at innholdet var kunstig generert. Det gjenstår

[105] The White House 2023b
[106] Teknologirådet 2023
[107] Europaparlamentet 2023c
[108] Bertuzzi 2023
[109] Europaparlamentet 2023a

å se hva de endelige kravene til merking faktisk blir. *AI Act* vil tidligst tre i kraft i 2026.

Det finnes også andre EU-reguleringer som kan bidra til å sette standarder for merking. Krav om informasjon om opphav og avsender er sentrale komponenter i EUs nye regelverk for politisk reklame. Reglene innebærer et forbud mot å målrette politisk reklame basert på sensitive personopplysninger, og et forbud mot reklame finansiert av ikke-europeiske sponsorer for å motvirke utenlandsk innblanding i for eksempel valg. I tillegg må all politisk reklame lagres i en offentlig database for å sikre innsyn og tilsyn.[110] Det har vært flere eksempler på bruk av generative verktøy i politisk reklame det siste året.[111]

I Kina ble krav om merking av kunstig generert innhold lovregulert i 2022.[112] Kravene ble innført i kjølvannet av en samfunnsdebatt om dype forfalskninger, som ble oppfattet som en trussel mot det digitale økosystemet for informasjonsflyt. Loven krever at syntetisk innhold merkes – alt fra chatboter som simulerer ekte mennesker, til generell merking av genererte bilder, tale og tekst. Loven forbyr tekniske løsninger som kan slette, endre eller skjule merkingen.

For Norge vil de kravene om merking og åpenhet som foreslås i de nye EU-lovene bli gjeldende når lovene inkorporeres i nasjonal lovgivning. Regjeringen har allerede tatt til orde for at kunstig generert innhold bør merkes.[113] Å støtte opp om å utvikle tekniske løsninger, sikre effektiv implementering og håndheving av nye lovverk, og følge den internasjonale utviklingen av standarder for merking tett vil være viktige grep i tiden som kommer.

[110] Europaparlamentet 2023b
[111] Se for eksempel Isenstadt 2023
[112] China Law Translate 2022
[113] Solheim 2023

STILLE SELSKAPENE TIL ANSVAR

Den moderne infrastrukturen for ytringsfrihet må reguleres for å bevare og sikre ytringsrommet og verne om ytringsfrihetens viktige funksjon.[114] De globale plattformselskapene har betydelig makt og kontroll over denne infrastrukturen, og er dermed også ansvarlige for spredning av ulovlig, skadelig og falskt innhold gjennom sine tjenester.

EU har nylig vedtatt en rekke nye lover for å regulere det digitale rom, som også vil bli norsk lov. Én lov som er på trappene i EU og Norge er forordningen om digitale tjenester (på engelsk *Digital Services Act, DSA)*.[115] Loven har som formål å begrense spredning av falske nyheter, desinformasjon og ulovlig innhold på nett, og stiller strenge krav til kontroll, innsyn og ansvarliggjøring av plattformselskaper for spredning av ulovlig og skadelig innhold i sine tjenester.

Gjennom *DSA* blir de største plattformselskapene slik som Meta, Google og X (tidligere Twitter) pålagt nye og strengere åpenhetskrav: De må rapportere om innhold som modereres og fjernes, begrunne hvorfor innhold slettes og gjøre begrunnelsene tilgjengelige i en offentlig database, og vurdere og håndtere risiko knyttet til valg, sikkerhet, helse og desinformasjon.[116]

Kravene som stilles til selskapene under *DSA* gjelder uavhengig om innholdet er maskin- eller menneskeskapt. Innhold generert av kunstig intelligens vil falle under samme kategori som annet innhold som deles og spres på sosiale medieplattformer, og må modereres og håndteres deretter. Selv om generative verktøy ikke er eksplisitt regulert gjennom *DSA*, vil altså kravene til åpenhet, rapportering og moderering av innhold omfatte kunstig generert innhold.[117]

[114] NOU 2022:9, s. 126
[115] Loven trer offisielt i kraft *25.februar 2024,* men har allerede blitt gjeldende for de største plattformselskapene. Forordningen vil bli norsk lov, men ligger fortsatt til vurdering av EFTA-landene, og vil derfor tre i kraft noe senere i Norge enn i Europa generelt
[116] Teknologirådet 2022b
[117] Se også Forbrukerrådet 2023, s. 55

Det diskuteres også om nye generative tjenester som ChatGPT, Bard og Bing bør kvalifiseres og stilles krav til som «søkemotorer» slik de er definert i *DSA*, men dette er foreløpig et uavklart tolkningsspørsmål.[118]

EU-forordningen er et viktig steg i riktig retning mot mer åpenhet og ansvarliggjøring. *DSA* ligger foreløpig til behandling i EFTA-landene og det er usikkert når loven vil tre i kraft i Norge. For at loven skal oppnå ønsket effekt, blir det viktig med effektiv implementering av loven i nasjonal lovgivning og aktiv håndheving og oppfølging. I tillegg vil det være nødvendig å vurdere om det er behov for å supplere eller komplementere EU-lovene gjennom nasjonal lovgivning.

UTREDE BEHOVET FOR NASJONALE TILPASNINGER

Ytringsfrihetskommisjonen påpekte at det er behov for å utforske mulighetene for supplerende regulering nasjonalt for å sikre mer innsyn og åpenhet, særlig på nasjonalt nivå, samt effektive tiltak mot skadevirkninger av desinformasjon og feilinformasjon.[119]

I *stortingsvedtak nr. 196, 12. desember 2022* ber Stortinget regjeringen om å vurdere hvilket handlingsrom Norge har til å regulere digitale tjenester utover regulering i forordningen om digitale tjenester. Dette vedtaket ligger nå hos Kommunal- og distriktsdepartementet til oppfølging.

Tross nye innstramminger er det fremdeles mye hemmelighold rundt plattformenes egen innholdsmoderering. Kunstig intelligens har lenge hatt en tiltagende rolle i å moderere og filtrere innhold på nett.[120] Teknologiselskapene benytter i utstrakt grad kunstig intelligens for å klassifisere brukerinnhold og hindre publisering, eller flagge innhold for moderering.[121] Det er i liten grad innsikt i dette, og det er vanskelig for utenforstående å få oversikt over faktiske effekter for ytringsfriheten.

[118] Botero Arcila 2023
[119] NOU 2022:9, s. 183, 152
[120] Kaye 2018
[121] Douek 2021, Kaye 2018

Det finnes i dag få kontrollmekanismer for å sikre at menneskelige moderatorer eller maskiner fjerner for mye innhold, og slik reduserer eller begrenser ytringsfriheten til brukerne.[122] Selskapene er heller ikke åpne om hvor mange menneskelige eller norskspråklige innholdsmoderatorer de har. Her kan det være behov for nasjonale tilpasninger for å stille selskapene til ansvar.

STILLE KRAV TIL SELSKAPENES SELVREGULERING

Gjennom den europeiske bransjenormen mot desinformasjon (på engelsk *The 2022 Code of Practice on Disinformation)* forplikter selskaper som Meta, Google, X (tidligere Twitter) og Tiktok seg til å innføre tiltak for å begrense spredning av desinformasjon gjennom frivillig rapportering. Ifølge en gjennomgang av selskapenes selvrapportering, utført av Medietilsynet, har rapporteringen store mangler og varierende kvalitet.[123]

Tilsynet påpeker også at det er en reell risiko for at tiltak mot desinformasjon kan føre til «overmoderasjon», som igjen kan utfordre informasjons- og ytringsfriheten. Effektive tiltak mot algoritmisk og annonsefinansiert forsterkning av desinformasjon bør derfor prioriteres over tiltak som innebærer å blokkere eller slette enkeltinnlegg som inneholder feil- eller desinformasjon, og som i sin tur kan ha den utilsiktede effekten å innskrenke ytringsrommet. [124]

Statusrapporter fra selskapene viser at de fleste selskapene har innført tiltak eller retningslinjer for å håndtere økning i kunstig generert tekst, bilder og video på plattformene. Meta har eksempelvis innført forbud mot bruk av generative verktøy i politisk reklame, inkludert annonser om valg. Meta tillater heller ikke bruk av genereringstjenester i boligannonser eller i annonser relatert til helse, farmasøytiske produkter og finansielle tjenester.[125]

Selskapene er ikke lovpålagt å innføre slike begrensninger på bruk av generative verktøy. De står dermed fritt til å selv velge hva slags begrensninger de vil

[122] NOU 2022:9
[123] Medietilsynet 2023
[124] Medietilsynet 2023
[125] Paul 2023

innføre, samt hva slags informasjon de vil offentliggjøre om omfanget og bruken av de nye generative verktøyene.

For myndighetene er det en viktig oppgave å kreve innsyn i teknologiselskapenes tjenester og praksis. En systematisk forbedring av kvaliteten på selskapenes egenrapportering vil sikre «bedre presisjon, konsistens og sammenlignbarhet», og føre til mer åpenhet og innsyn.[126]

STYRKE DIGITAL KILDEKRITISK FORSTÅELSE

Tilstedeværelsen av store mengder kunstig generert innhold på nett krever en befolkning med digital kildekritisk kompetanse. Evne til å vurdere og verifisere informasjon vil bli helt avgjørende i møte med den nye teknologien.

Derfor vil også behovet for å bygge digital kompetanse øke. Det gjelder befolkningen generelt, men også blant studenter, journalister og andre relevante fagmiljøer. Samtidig vil det bli viktigere å overvåke og følge med på utviklingen av informasjonsflyten på nett og effektene av spredning av desinformasjon og dype forfalskninger.

Offentlige debatter om trusler og effekter kan bidra til å styrke menneskers kildekritiske kompetanse. Styrking av fysiske ytringsrom som biblioteker og litteraturhus og andre arenaer der mennesker kan møtes fysisk og utøve sin ytringsfrihet, kan også bidra til å legge til rette for en bred deltakelse og åpen og opplyst debatt.

Medieinstitusjoner med digital integritet og sterke uavhengige sivilsamfunnsorganisasjoner kan bidra til å bevare tilliten til informasjon gjennom å verifisere innhold på nett og være en motvekt til desinformasjon. Disse vil bli kritiske funksjoner for å bygge motstandsdyktighet mot digital påvirkning og manipulering.

[126] Medietilsynet 2023

ETABLERE ET PSYKOLOGISK FORSVAR

I møte med økte mengder falskt, maskinskapt innhold på nett vil det bli behov for å styrke den nasjonale motstandsdyktigheten mot desinformasjon. For å styrke kunnskap, ha løpende oversikt over teknologiutviklingen og trender innenfor desinformasjon, samt å styrke motstandsdyktighetene i den norske digitale offentligheten, kan det vurderes å etablere et *psykologisk forsvar*.

Svenske myndigheter opprettet et «psykologisk forsvar» i 2022. Denne myndigheten jobber både forebyggende med blant annet undervisning og informasjonsarbeid, og operativt gjennom å overvåke og identifisere spredningen av desinformasjon, samt innføre responstiltak.[127] Til sammenligning har Storbritannia opprettet en egen enhet – *Counter Disinformation Unit* – på departementsnivå som skal håndtere akutte informasjonshendelser.

Ytringsfrihetskommisjonen av 2022 påpekte at Norge er relativt sett godt rustet for å håndtere desinformasjon. Polariseringen i Norge ikke er veldig stor sammenlignet med andre land, og den kritiske medieforståelsen står sterkt samtidig som mediene har tillit i befolkningen.[128] Likevel kan generativ kunstig intelligens forsterke eksisterende utfordringer i det digitale informasjonssystemet.

Totalberedskapskommisjonen viser til at den norske innsatsen for å forebygge og håndtere desinformasjon er fragmentert og delt på flere aktører. Kommisjonen anbefaler derfor å etablere en nasjonal funksjon med et helhetlig ansvar for å forstå og motvirke trusselen fra påvirkningsoperasjoner og desinformasjon.

[127] NOU 2023:17, s.122
[128] NOU 2022:9

STØTTE REDAKTØRSTYRTE MEDIER

En forutsigbar effekt av tiltakende usikkerhet rundt tekst, bilder, lydklipp og videoer på nett er at digitale plattformer med tillit hos publikum vil få en større betydning. Etter hvert som det blir klarere for folk flest at informasjonen i det digitale domenet i mindre grad kan stoles på, vil verdien av og behovet for pålitelig informasjon øke.[129]

Den nordiske modellen gir redaktørstyrte medier bestemte rettslige og økonomiske rammevilkår. I tillegg er modellen basert på en uavhengig og ansvarlig redaktør, samt en anerkjent selvdømmeordning. Ordningen har vært viktig for å bygge tillit.

I møte med kunstig generativ intelligens vil det trolig bli enda viktigere at publikum forstår og kan stole på at det er mediene og redaktøren selv som er den endelige garantisten for innholdets troverdighet. Denne tilliten forutsetter at mediene selv forvalter teknologien på en ansvarsfull måte.

Profesjonelle, pålitelige og upartiske redaktørstyrte medier som til dels har blitt svekket som følge av økonomiske, strukturelle og ideologiske utfordringer de siste to tiårene, vil trolig få en sentral rolle. I lys av infrastrukturkravet bør norske myndigheter følge utviklingen nøye og se på tiltak og muligheter for å styrke kapasiteten og rollen til redaktørstyrte medier ytterligere.

PÅVIRKE INTERNASJONAL NORMUTVIKLING

Selv om infrastrukturkravet pålegger norske myndigheter et utstrakt ansvar for ytringsrommet, utvikles det nå en rekke retningslinjer, standarder og lovverk for utvikling og bruk av kunstig intelligens. Dette foregår primært på internasjonalt nivå. Når digital infrastruktur utvikles, driftes og domineres av

[129] Jungherr 2023

store internasjonale selskaper, kreves det (sterke) koalisjoner av stater for å regulere aktørene og utviklingen.

Fremover er det særlig internasjonal regeldannelse gjennom EU, Europarådet og FN-traktater som vil sette rammene for kalibrering av informasjonsdomenet og stille felles krav til teknologiselskapene. Følgelig vil diplomatiske virkemidler som forhandlinger og felles saker med likesinnede stater være viktige for norske myndigheter for å sikre langsiktige løsninger for det digitale ytringsrommet.

Påvirkning av internasjonale standarder skjer best gjennom at likesinnede stater og organisasjoner går sammen og samarbeider om sentrale posisjoner. Ytringsfrihet fortoner seg ulikt i forskjellige land og regioner i verden, og små, åpne, demokratiske velferdsstater som de nordiske landene har sammenfallende utfordringer og felles interesse i å forme reglene på måter som særlig tar inn over seg våre sårbarheter. Tett koordinering og samarbeid med andre nordiske land vil derfor øke mulighetene for gjennomslag i internasjonalt lovgivnings- eller normutviklingsarbeid.

Lovgivningen som vil gjelde i Norge og som vil ramme inn skapende kunstig intelligens hos oss, vil særlig være EU-regler som må implementeres i kraft av EØS-avtalen. Det vil gjelde store deler av lovgivningspakken som EU utarbeider under sin digitale pakke som del av sin digitale strategi for 2020–2025.[130] Lovgivningsarbeidet i Brussel er derfor den viktigste arenaen hvor norske styresmakter kan påvirke innholdet i det som senere blir inkorporert som lovgivning i Norge.

[130] EU-kommisjonen 2020

REFERANSER

Adjer, Henry, Giorgio Patrini, Francesco Cavalli & Laurence Cullen (2019) *The State of Deepfakes: Landscape, Threats, and Impact*. September 2019. Hentet fra: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

Aswad, Evelyn (2020) *Loosing the freedom to be human.* Colombia Human Rights Law Review, vol. 52. Hentet fra: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3635701

Bender, Emily M., Timnit Gebru, Angelina McMillian-Major, Shmargaret Shmitchell (2021) *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, mars 2021. Hentet fra: https://dl.acm.org/doi/10.1145/3442188.3445922

Benson, Thor (2023) *This Disinformation Is Just for You.* Wired, 1.august 2023. Hentet fra: https://www.wired.com/story/generative-ai-custom-disinformation/

Bertuzzi, Luca (2023) *EU's AI Act negotiations hit the brakes over foundation models.* Euractiv, 10. November 2023. Hentet fra: https://www.euractiv.com/section/artificial-intelligence/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models/

Botero Arcila, Beatriz (2023) *Is It a Platform? Is It A Search Engine? It's ChatGPT! The European Liability Regime for Large Language Models.* Journal of Free Speech Law, august 2023. Hentet fra: https://www.journaloffreespeechlaw.org/boteroarcila.pdf

Brandtzæg, Petter Bae, Marita Skjuve, Asbjørn Følstad (2022) *My AI Friend: How users of a Social Chatbot Understand Their Human – AI Friendship*. Human Communication Research, Juli 2022. Hentet fra: https://academic.oup.com/hcr/article/48/3/404/6572120

Buolamwini, Joy, Timnit Gebru (2018) *Gender shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. Proceedings of Maching Learning Research, 2018. Hentet fra: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Bushard, Brian (2023) *Fake Image of Explosion Near Pentagon Went Viral – Even Though It Never Happened*. Forbes, 22. Mai 2023. Hentet fra: https://www.forbes.com/sites/brianbushard/2023/05/22/fake-image-of-explosion-near-pentagon-went-viral-even-though-it-never-happened/?sh=182c81f249a5

Cantor, Matthew (2023) *Nearly 50 news websites are 'AI-generated', a study says. Would I be able to tell?* The Guardian, 8.mai 2023. Hentet fra: https://www.theguardian.com/technology/2023/may/08/ai-generated-news-websites-study

China Law Translate (2022) *Forskrift for dypsyntesebehandling av internet-informasjonstjenester (oversatt til norsk)*. 25.november 2022. Hentet fra: https://www.chinalawtranslate.com/deep-synthesis/

Citron, Danielle K., Robert Chesney (2019) *Deep Fakes: A Looming Challenge for Privacy, Democracy and National Security*. University of Berkeley School of Law, desember 2019. Hentet fra: https://scholarship.law.bu.edu/faculty_scholarship/640/

Cox, Joseph (2023) *GPT-4 Hired Unwitting TaskRabbit Worker by Pretending to be "Vison-Impaired" Human*. Vice, 15. Mars 2023. Hentet fra: https://www.vice.com/en/article/jg5ew4/gpt4-hired-unwitting-taskrabbit-worker

Cyabra (2023) *1 in 4 profiles Are Pro-Hamas Fake Accounts. The Online Battlefront*. 11. oktober 2023. Hentet fra: https://cyabra.com/1-of-4-pro-hamas-profiles-are-fake-the-online-battlefront/

Dagens Næringsliv (2023) *Retningslinjer for redaksjonell bruk av kunstig intelligens (KI) I Dagens Næringsliv*. 25. Oktober 2023. Hentet fra: https://www.dn.no/teknologi/retningslinjer-for-redaksjonell-bruk-av-kunstig-intelligens-ki-i-dagens-naringsliv/2-1-1463963

Dallison, Paul (2023) *Panic! Russia sends France into a tailspin over bedbugs*. Politico, 27. Oktober 2023. Hentet fra: https://www.politico.eu/article/panic-russia-sends-france-into-a-tailspin-over-bedbugs/

David, Emilie (2023) *OpenAI can't tell if something was written by AI after all*. The Verge, 25. Juli 2023. Hentet fra: https://www.theverge.com/2023/7/25/23807487/openai-ai-generated-low-accuracy

Den europeiske menneskerettighetsdomstol (1976) *Handyside v. The United Kingdom*. No. 5493/72, Dom 7. desember 1976. Hentet fra: https://hudoc.echr.coe.int/?i=001-57499

Den europeiske menneskerettighetsdomstol (2005) *Salov v. Ukraine*. No. 65518/01. Dom 6. september 2005. Hentet fra: https://hudoc.echr.coe.int/?i=001-70097

Den europeiske menneskerettighetsdomstol (2022) *NIT S.R.L v. The republic of Moldova*. No. 28470/12. Dom 5.april 2022. Hentet fra: https://hudoc.echr.coe.int/fre?i=002-13629

De Vynck, Gerrit (2023) *ChatGPT leans liberal, research shows*. The Washington Post, 16.august 2023. Hentet fra: https://www.washingtonpost.com/technology/2023/08/16/chatgpt-ai-political-bias-research/

Douek, E. (2021) *Governing Online Speech: From "post-as-trumps" to proportionality and probability*. Colombia Law Review 121(3). Hentet fra: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3679607

Elliott, Vittoria (2023) *Big AI Won't Stop Election Deepfakes With Watermarks*. Wired, 27. juli 2023. Hentet fra: https://www.wired.com/story/ai-watermarking-misinformation/

Everypixel Journal (2023) *AI Has Already Created As Many Images As Photographers Have Taken in 150 Years. Statistics for 2023*. Hentet fra: https://journal.everypixel.com/ai-image-statistics

EU-kommisjonen (2020) *Shaping Europe's digital future*. 19. februar 2020. Hentet fra: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/shaping-europes-digital-future_en

Europarådets konvensjon om beskyttelse av menneskerettighetene og de grunnleggende friheter (Den europeiske menneskerettskonvensjonen/EMK), 4. november 1950.

Europaparlamentet (2023a) *Artificial intelligence Act. Amendments adopted by the European Parliament on 14 June 2023.* Hentet fra: *https://artificialintelligenceact.eu/wp-content/uploads/2023/06/AIA-%E2%80%93-IMCO-LIBE-Draft-Compromise-Amendments-14-June-2023.pdf*

Europaparlamentet (2023b) *Political advertising: deal on new measures to crack down on abuse.* Pressemelding, 6.november 2023. Hentet fra: https://www.europarl.europa.eu/news/en/press-room/20230626IPR00819/political-advertising-deal-on-new-measures-to-crack-down-on-abuse

Europaparlamentet (2023c) *Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI.* Pressemelding, 9.desember 2023. Hentet fra: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai

Europol (2022) *Facing reality? Law enforcement and the challenge of deepfakes.* 28. april 2022. Hentet fra: https://www.europol.europa.eu/publications-events/publications/facing-reality-law-enforcement-and-challenge-of-deepfakes

Flannery O'Connor, Jennifer, Emily Moxley (2023) *Our approach to responsible AI innovation.* YouTube Official Blog, 14. November 2023. Hentet fra: https://blog.youtube/inside-youtube/our-approach-to-responsible-ai-innovation/

FNs konvensjon om sivile og politiske rettigheter (SP-konvensjonen), 999 UNTS 171 (16.09.1966, trådte i kraft 23.03.1976).

FNs menneskerettighetskomité (2011) G*eneral Comment No. 34* on Article 19: Freedoms of opinion and expression. Hentet fra: https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf

FNs menneskerettighetsråd (2012) *The promotion, protection and enjoyment of human rights on the internet.* Resolusjon 20/8, para 1. Hentet fra: https://ap.ohchr.org/documents/dpage_e.aspx?si=a/hrc/res/20/8

FNs menneskerettighetsråd (2013) *Annual report of the United Nations High Commissioner for Human Rights. Addendum, Annex.* A/HRC/22/17/Add4. Hentet fra: https://www.ohchr.org/sites/default/files/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf

Forbrukerrådet (2023) *Ghost in the machine – Forbrukerutfordringer ved generativ kunstig intelligens*. Juni 2023. Hentet fra: https://storage02.forbrukerradet.no/media/2023/08/fr-generative-ai-rapport-web-no-mindre.pdf

Freedom House (2023) *Freedom on the Net 2023 – the Repressive Power of Artifical Intelligence*. Hentet fra: https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

Fried, Ina (2023) *How AI will turbocharge misinformation – and what we can do about it*. Axios, 10.juli 2023. Hentet fra: https://www.axios.com/2023/07/10/ai-misinformation-response-measures

Greger, Mats With (2023) *NRK innfører nye retningslinjer etter "kabelgate"*. Journalisten. 9. desember 2022. Hentet fra: https://www.journalisten.no/nrk-innforer-nye-retningslinjer-etter-kabelgate/548527

Harari, Yuval (2023) *What does the AI revolution mean for our future*. Videointervju mellom Mustafa Suleyman og Yuval Harari, oktober 2023. Hentet fra https://www.youtube.com/watch?v=7JkPWHr7sTY&t=1092s

Heikkilä, Melissa (2023a) *Google DeepMind har launched a watermarking tool for AI-generated images*. MIT Technology Review, 29. august 2023. Hentet fra: https://www.technologyreview.com/2023/08/29/1078620/google-deepmind-has-launched-a-watermarking-tool-for-ai-generated-images/

Heikkilä, Melissa (2023b) *This new tool could protect you pictures from AI manipulation*. MIT Technology Review, 26 juli 2023. Hentet fra: https://www.technologyreview.com/2023/07/26/1076764/this-new-tool-could-protect-your-pictures-from-ai-manipulation/

Hern, Alex (2022) *AI-assisted plagiarism? ChatGPT bot says it has an answer for that*. The Guardian, 31. Desember 2022. Hentet fra: https://www.theguardian.com/technology/2022/dec/31/ai-assisted-plagiarism-chatgpt-bot-says-it-has-an-answer-for-that

Imperva (2023) *2022 Imperva Bad Bot Report – Evasive Bots Drive Online Fraud*. Imperva 2022. Hentet fra: https://www.imperva.com/resources/reports/2022-Imperva-Bad-Bot-Report.pdf

Isenstadt, Alex (2023) *DeSantis PAC used AI-generated Trump voice in ad attacking ex-president*. Politico, 17. Juli 2023. Hentet fra:

https://www.politico.com/news/2023/07/17/desantis-pac-ai-generated-trump-in-ad-00106695

Jakesch, Maurice, Advait Bhat, Daniel Buscheck, Lior Zalmanson, Mor Naaman (2023) *Co-Writing with Opinionated Language Models Affects Users' Views*. Hamburg, 23-28. April 2023. Hentet fra: https://mauricejakesch.com/assets/pdf/aimc_influence.pdf

Jones, Kate (2019) *Online Disinformation and Political Discourse, Applying a Human Rights Framework*. International Law Programme, November 2019. Hentet fra: https://www.chathamhouse.org/sites/default/files/2019-11-05-Online-Disinformation-Human-Rights.pdf

Jungherr, Andreas (2023) *Artificial Intelligence and Democracy: A conceptual Framework*. Sage Journals, juli-september 2023. Hentet fra: https://journals.sagepub.com/doi/epub/10.1177/20563051231186353

Kaye, David (2018) *Report on Artificial Intelligence Technologies and implications for freedom of expressing and the information environment*. UN Special Rapporteur on Freedom of opinion and expression. 29. August 2018. Hentet fra: https://www.ohchr.org/en/calls-for-input/report-artificial-intelligence-technologies-and-implications-freedom-expression-and

Kirchenbauer, John, Jonas Geiping, Yuxin Wen, Joanathan Katz, Ian Miers, Tom Goldstein (2023) *A Watermark for Large Language Models*. Cornell University, juni 2023. Hentet fra: https://arxiv.org/abs/2301.10226

Khan, Irene (2021) *Disinformation and freedom of opinion and expression*. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. Human Rights Council, A/HRC/47/25. Hentet fra: https://digitallibrary.un.org/record/3925306#record-files-collapse-header

Khan, Jeremy (2023) *Stigma of dating a chatbot will fade, Replika CEO predicts*. Fortune, 12.juli 2023. Hentet fra: https://fortune.com/2023/07/12/brainstorm-tech-chatbot-dating/

Knight, Will (2023) *It Costs Just $400 to Build an AI Disinformation Machine*. Wired, 29. August 2023. Hentet fra: https://www.wired.com/story/400-dollars-to-build-an-ai-disinformation-machine/

Kóváry Sólymos, Karin (2023) *Slovakia: Deepfake audio of Dennik N journalist offers worrying example of AI abuse*. International Press Institute,

31. Oktober 2023. Hentet fra: https://ipi.media/slovakia-deepfake-audio-of-dennik-n-journalist-offers-worrying-example-of-ai-abuse/

Kripos (2023) Generativ kunstig intelligens og cyberkriminalitet. August 2023. Hentet fra. https://www.politiet.no/aktuelt-tall-og-fakta/aktuelt/nyheter/2023/08/18/ny-rapport-om-generativ-kunstig-intelligens/

Liang, Weixin, Mert Yuksekgonul, Yining Mao, Eric Wu, James Zou (2023) *GPT detetors are biased against non-native English writers*. Cornell University, juli 2023. Hentet fra: https://arxiv.org/abs/2304.02819

Maruf, Ramishah (2023) *Lawyer apologizes for fake court citations from ChatGPT*. CNN, 28.mai 2023. Hentet fra: https://edition.cnn.com/2023/05/27/business/chat-gpt-avianca-mata-lawyers/index.html

Medietilsynet (2023) *De globale plattformenes etterlevelse av bransjenormen mot desinformasjon*. 26. mai 2023. Hentet fra: https://www.medietilsynet.no/regelverk/internasjonale-reguleringer-for-digitale-tjenester/bransjenormen-mot-desinformasjon/

Medium (2021) *History and present of Natural Language Processing*. 29.november 2021. Hentet fra: https://deep-talk.medium.com/history-and-present-of-natural-language-processing-f19280866497

Metz, Cade (2023) *Chatbots May 'Hallucinate' More Often Than Many Realize*. The New York Times, 6. November 2023. Hentet fra: https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates

Miller, Elizabeth J., Ben A. Steward, Amy Dawel (2023) *AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones*. Sage Journals, oktober 2023. Hentet fra: https://journals.sagepub.com/doi/epub/10.1177/09567976231207095

Mullin, Benjamin, Nico Grant (2023) *Google Tests A.I. Tool That is Able to Write News Articles*. The New York Times, 19. Juli 2023. Hentet fra: https://www.nytimes.com/2023/07/19/business/google-artificial-intelligence-news-articles.html

Murphy, Matt (2022) *The Dawn of A.I. Mischief Models*. Slate Magazine. Hentet fra: https://slate.com/technology/2022/08/4chan-ai-open-source-trolling.html

Myers, Andrew (2023) *Assessing Political Bias in Large Language Models*. HAI, Stanford University, 22.mai 2023. Hentet fra: https://hai.stanford.edu/news/assessing-political-bias-language-models

Natanson, Elad (2023) *Hyper-Personalization Is Already Here – Its Futur Is Even More Cutting- Edge*. Forbes, 1 juni 2023. Hentet fra: https://www.forbes.com/sites/eladnatanson/2023/06/01/hyper-personalization-is-already-here---its-future-is-even-more-cutting-edge/

NewsGuard (2023) *Tracking AI-enabled Misinformation: 566 'Unreliable AI-Generated News' Websites (and Counting), Plus the Top Falske Narratives Generated by Artificial Intelligence Tools*. 27.november 2023. Hentet fra: https://www.newsguardtech.com/special-reports/ai-tracking-center/

NOU 1999: 27 *Ytringsfrihet bør finde sted*. Hentet fra: https://www.regjeringen.no/no/dokumenter/nou-1999-27/id142119/

NOU 2022:9 *En åpen og opplyst offentlig samtale*. Ytringsfrihetskommisjonen 2022. Hentet fra: https://www.regjeringen.no/no/dokumenter/nou-2022-9/id2924020/

NOU 2023:17 *Nå er det alvor – rustet for en usikker fremtid*. Totalberedskapskommisjonen. Hentet fra: https://www.regjeringen.no/contentassets/4b9ba57bebae44d2bebfc845ff6cd5f5/no/pdfs/nou202320230017000dddpdfs.pdf

OpenAI (2023a) *ChatGPT can now see, hear, and speak*. Blogg, 25. September 2023. Hentet fra: https://openai.com/blog/chatgpt-can-now-see-hear-and-speak

OpenAI (2023b) *Introducing GPTs*. Blogg, 6.november 2023. Hentet fra: https://openai.com/blog/introducing-gpts

OpenAI (2023c) *GPT-4 Systems Card*. 23. Mars 2023. Hentet fra: https://cdn.openai.com/papers/gpt-4-system-card.pdf

OpenAI (2023d) *How should AI systems behave, and who should decide?* Blogg, 16. februar 2023. Hentet fra: https://openai.com/blog/how-should-ai-systems-behave

Paul, Katie (2023) *Meta bars political advertisers from using generative AI ads tools*. Reuters, 7. November 2023. Hentet fra: https://www.reuters.com/technology/meta-bar-political-advertisers-using-generative-ai-ads-tools-2023-11-06/

Pew Research Center (2023) *As AI Spreads, Experts Predict the Best and Worst Changes in Digital Life by 2025.* Juni 2023. Hentet fra: https://www.pewresearch.org/internet/2023/06/21/as-ai-spreads-experts-predict-the-best-and-worst-changes-in-digital-life-by-2035/

Project Origin (2023) Project Origin, Protecting Trusted Media, Overview. Hentet fra: https://www.originproject.info/about

Prosjekt Reynir (2023) *One pager prosjekt reynir*. Media City Bergen. Hentet fra: https://mediacitybergen.no/media/9869/one-pager-prosjekt-reynir.pdf

Robinson, David (2023) *Hamas Propaganda Analysis-over 85 % of engagement are bots.* 16 oktober 2023. Hentet fra: https://internet2-0.com/hamas-propaganda-analysis/

Roose, Kevin (2023) *Personalized A.I. Agents Are Here. Is the World Ready for Them*. The New York times, 10.november 2023. Hentet fra: https://www.nytimes.com/2023/11/10/technology/personalized-ai-agents.html

Saberi, Mehrdad, Vinu Sanker Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, Soheil Feizi (2023) *Robustness of AI-Image Detectors: Fundamental Limits and Practical Attacks.* Oktober 2023. Hentet fra: https://arxiv.org/pdf/2310.00076.pdf

Salman, Hadi, Alaa Khaddaj, Guillaume Leclerc, Andrew Illyas (2023) *Raising the Cost of Malicious AI-Powered Image Editing*. MIT, februar 2023. Hentet fra: https://arxiv.org/abs/2302.06588

Santurkar, Shibani, Esin Durmus, Faisal ladhak, Cinoo Lee, Percy Liang, Tatsunori Hashimoto (2023) *Whose Opinions Do Language Models Reflect*. Cornell University, Mars 2023. Hentet fra: https://arxiv.org/abs/2303.17548

Simon, Felix M., Sacha Altay, Hugo Mercier (2023) *Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown.* Harvard Kennedy School, oktober 2023. Hentet fra: https://misinforeview.hks.harvard.edu/wp-content/uploads/2023/10/simon_generative_AI_fears_20231018.pdf

Simonite, Tom (2023) *A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be.* Wired, 17.mars 2022. Hentet fra: https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/

Smith B. C. (2019). *The promise of artificial intelligence: Reckoning and judgment*. MIT Press.

Solheim, Une (2023) *Regjeringen vil merke KI-generert innhold: - Det er en fare.* TV2, 28. oktober 2023. Hentet fra: https://www.tv2.no/nyheter/innenriks/regjeringen-vil-merke-ki-generert-innhold-det-er-en-fare/16168912/

Teknologirådet (2022a) *Taleteknologi med kunstig intelligens.* Publisert desember 2022. Hentet fra: https://teknologiradet.no/publication/taleteknologi-og-kunstig-intelligens/

Teknologirådet (2022b) *Saken forklart: Ny EU-lov skal gjøre internett tryggere.* Publisert november 2022. Hentet fra: https://teknologiradet.no/publication/ny-eu-lov-skal-gjore-internett-tryggere/

Teknologirådet (2023) *Saken forklart: EU vil regulere kunstig intelligens.* Publisert juni 2023. Hentet fra: https://teknologiradet.no/publication/eu-vil-regulere-kunstig-intelligens/

Thompson, Alan D. (2022) *A Comprehensive Analysis of Datasets Used to Train GPT-1, GPT-2, GPT-3, GPT-NeoX20B, Megatron-11B, MT-NLG, and Gopher, An Independent Report*. LifeArchitect.ai, mars 2022. Hentet fra: https://s10251.pcdn.co/pdf/2022-Alan-D-Thompson-Whats-in-my-AI-Rev-0b.pdf

The White House (2023a) *FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI*. 21. Juli 2023. Hentet fra: https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

The White House (2023b) *FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.* 30. Oktober 2023. Hentet fra: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

Tiku, Nitasha, Will Oremus (2023) *The right's new culture-war target: 'woke AI'*. The Washington Post, 24. februar 2023. Hentet fra:

https://www.washingtonpost.com/technology/2023/02/24/woke-ai-chatgpt-culture-war/

Twomey, John, Didier Ching, Matthew Peter Aylett, Michael Quayle, Conor Linehan, Gillian Murphy (2023) *Do deepfake videos undermine epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine*. Oktober 2023. Hentet fra: https://doi.org/10.1371/journal.pone.0291668

University College London (2023) *'Deepfakes' ranked as most serious AI crime threat*. UCL, 4. August 2023. Hentet fra: https://www.ucl.ac.uk/news/2020/aug/deepfakes-ranked-most-serious-ai-crime-threat

University of East Anglia (2023) *Fresh evidence of ChatGPTs political bias revealed by comprehensive new study*. 17.august 2023. Hentet fra: https://www.uea.ac.uk/about/news/article/fresh-evidence-of-chatgpts-political-bias-revealed-by-comprehensive-new-study

Verma, Pranshu (2023) *AI fake nudes are booming. It's ruining real teens' lives*. The Washington Post, 5. November 2023. Hentet fra: https://www.washingtonpost.com/technology/2023/11/05/ai-deepfake-porn-teens-women-impact/

Verma, Pranshu, Will Oremus (2023) *ChatGPT invented a sexual harassment scandal and named a real law prof as the accused*. The Washington Post, 5. April 2023. Hentet fra: https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/

Vincent, James (2020) *OpenAI's latest breakthrough is astonishingly powerful, but still fighting its flaws*. The Verge, 30 juli 2020. Hentet fra: https://www.theverge.com/21346343/gpt-3-explainer-openai-examples-errors-agi-potential

Waaler, Ingrid Emilie, Ine Julia Rojahn Schwebs (2023) *Stadig flere KI-romansar: Kjærlei fri for risiko*. NRK, 11. september 2023. Hentet fra: https://www.nrk.no/norge/fleire-blir-kjaerastar-med-kunstig-intelligens-_ki_-1.16474942

Walker Rettberg, Jill (2023) *Lystløgneren ChatGPT*. NRK, 4.mars 2023. Hentet fra: https://www.nrk.no/ytring/lystlogneren-chatgpt_-1.16316248

Walters v. OpenAI LLC (2023) *No 23-A-04860-2*. Court of Gwinnett Country, state of Georgia. Hentet fra: https://www.courthousenews.com/wp-content/uploads/2023/06/walters-openai-complaint-gwinnett-county.pdf

Ward, Jake (2023) *Innlegg fra X (tidligere Twitter). "We pulled off an SEO Heist that stole 3,6M totalt traffic from a competitor. Here's how we did it"*. 24. November 2023. Hentet fra: https://twitter.com/jakezward/status/1728032634037567509?s=20

West, Darrell M. (2023) *How AI will transform the 2024 elections*. Brookings, 3. Mai 2023. Hentet fra: https://www.brookings.edu/articles/how-ai-will-transform-the-2024-elections/