SENTER FOR LANGSIKTIG POLITIKK

30.11.23

INNSPILL TIL NY NASJONAL DIGITALISERINGSSTRATEGI

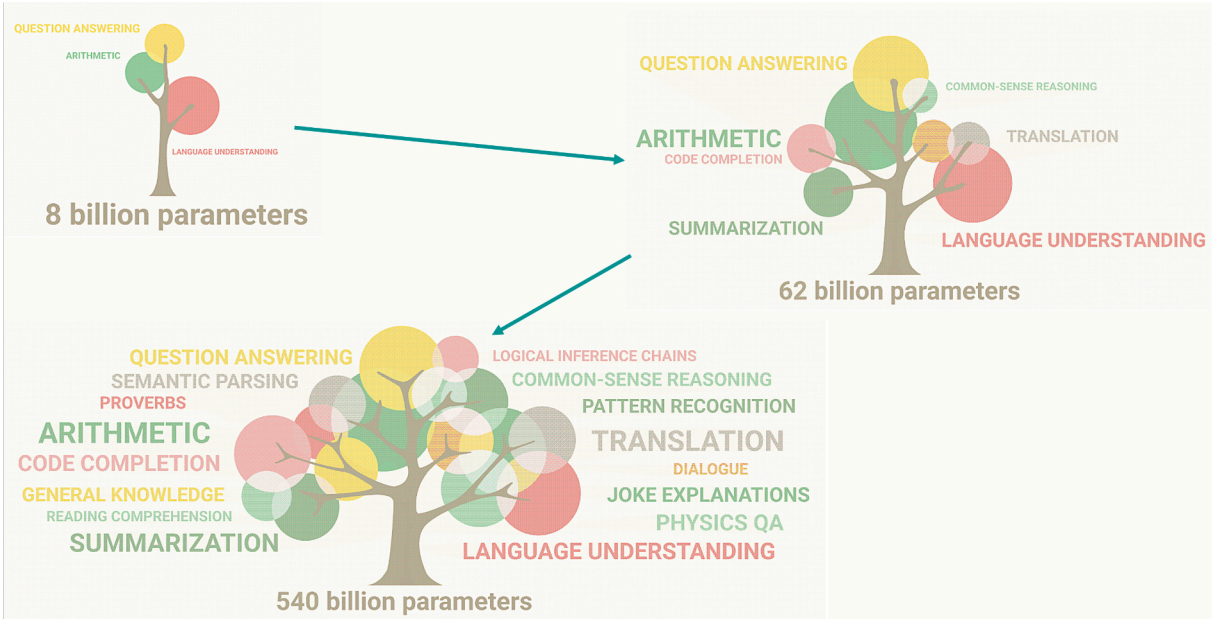
SENTER FOR LANGSIKTIG POLITIKK

Innspill til ny nasjonal digitaliseringsstrategi

Vi i Senter for langsiktig politikk takker for anledningen til å gi innspill til digitaliseringsstrartegien. Å sikre trygg og samfunnsnyttig bruk av nye digitale teknologier, særlig kunstig intelligens (KI), er en av regjeringens viktigste oppgaver.

Er det drivkrefter og utviklingstrekk som vil påvirke samfunnet generelt, og som strategien bør ta høyde for?

Selv om vi forstår treningsprossessen for KI-modeller godt, forstår vi selve KI-modellene vi skaper dårlig. Fremskrittene de siste årene kommer hovedsakelig fra å bruke stadig mer data og datakraft og ikke fra ny kunnskap om hvordan modellene fungerer. Større modeller er ikke bare bedre på det samme, de får kvalitativt nye ferdigheter.[1] Det gjør at gapet mellom hvor kraftig KI vi klarer å lage og hvor godt vi forstår hva vi lager, blir større og større.



Figur 1[2]: Googles illustrasjon av hvordan modellen deres, PALM, ikke bare blir bedre, men også får nye evner når de øker antall parametre.

Dette gjør det også vanskelig å forutsi hva neste generasjon modeller kan få til. På spørsmål om hva GPT-5 kan få til sa Sam Altman nylig at "Frem til vi trener modellen, er det som en morsom gjettelek for oss".[3] Selv lenge etter publisering kan det bli oppdaget nye egenskaper i en modell som de som lagde modellen ikke var klar over.[4] At vi ikke forstår hva som skjer inne i modellene gjør også at vi ikke har robuste metoder for å sikre at de oppfører seg som vi ønsker i nye situasjoner. [5],[6]

Vi må forvente rask og uforutsigbar utvikling fremover. At kraftige modeller rulles ut uten at vi vet helt hvor sikre de er eller hva de kan bringe, gir både store muligheter og store risikoer samtidig.

Hva er de viktigste KI-utfordringene fremover?

KI-utviklingen bringer tre typer utfordringer:

- **Misbruksrisiko** er risiko for at ondsinnede aktører bruker KI til å volde skade. KI kan senke terskelen for å gjennomføre avanserte **cyberangrep[7], utvikle kjemiske[8] eller biologiske[9] våpen** og styrke stater og selskapers evne til å **overvåke**[10],[11] og **manipulere** enkeltpersoner og demokratiske prosesser.[12]
- **Ulykkesrisiko** er risiko for ulykker forårsaket av KI-systemer som ikke oppfører seg som ment. Hastigheten og uforutsigbarheten til KI kan øke risikoen for alvorlige ulykker ved for eksempel bruk i **kritisk infrastruktur[13], autonome våpen**[14] eller **atomvåpensystemer**.[15]
- **Strukturelle problemer** som følge av at KI kan skape **systematisk diskriminering**[16] av utsatte grupper, skape økonomiske **omstillingsutfordringer**[17] [18], **utfordre demokrati og tillit**[19], samt forårsake **destabiliserende kappløpsdynamikk**[20] mellom stater og selskaper.

Alle utfordringene vil skalere i takt med at KI blir kraftigere. KI-systemene har blitt stadig mer generelle de siste årene og på mange områder gjør de det bedre enn mennesker. Vi burde forvente at trenden fortsetter. Mange eksperter er redde for at dersom systemene blir veldig kapable uten at vi klarer å sikre dem, vil det kunne få svært alvorlige konsekvenser for menneskeheten.[21] [22]

Hvordan kan regjeringen bidra til å løse disse utfordringene gjennom denne strategien?

Bygge KI-kompetanse i statsapparatet
Vi må forstå utviklingen og respondere raskt og robust på problemstillingene den medfører.

Norge burde:

- Gjøre Digdir til en KI-hub. Ambisjonen for Digitaliseringsdirektoratet om å være verdensledende på digitalisering bør utvides til også å inkludere bruk av kunstig intelligens. Direktoratet bør fungere som en kompetansebank for kunstig intelligens i offentlig sektor, og målet må være å heve kvaliteten og brukervennligheten av offentlige tjenester.
- Utvide mandatet og kompetansen til Datatilsynet dersom de skal føre algoritmetilsyn. KI er mye mer enn personvern.
- Etablere et KI-register for risikable systemer og KI-ulykker. Et register for bruk av KI-systemer innenfor sensitive områder som helse, kritisk infrastruktur og finans vil uansett bli et krav fra EU og bør implementeres så fort som mulig. Vi bør også innføre et register for ulykker som kan skyldes KI. Ved alvorlige ulykker bør det være praksis at disse etterforskes av offentlige myndigheter, eksempelvis etter modell fra Havarikommisjonen.

Satse på målrettet sikkerhetsarbeid
KI bringer et bredt spektrum av risikoer, og kan virke geopolitisk og økonomisk destabiliserende. Sikkerhetsarbeidet burde integreres tett med digitaliseringsarbeidet, da digitalisering vil bringe nye sårbarheter. Vi håper digitaliseringsstrategien bygger på NSMs rapporter: Nasjonalt digitalt risikobilde 2023 og Sikkerhetsfaglig råd - Et motstandsdyktig Norge, og satser på omfattende sikkerhetsarbeid tilknyttet fremvoksende teknologier som KI.

Vi burde se etter muligheter for å bruke KI innen områder som minsker trusler, som cyberforsvar, monitorering av kritisk infrastruktur, detektering av farlige DNA-bestillinger og avsløring og merking av falskt og KI-generert medieinnhold.

Norge burde:

- Styrke cybersikkerheten etter Sikkerhetsfaglig råds anbefalinger (kap 10)
- Oppnevne en NOU for å få oversikt over sikkerthetsimplikasjonene av KI-utviklingen. Arbeidet kan kombinere NSMs aktørfokuserte analyser med et bredere fokus som inkluderer ulykkesrisiko og strukturell påvirkning fra KI, samt scenarioanalyser for fremtidig KI-utvikling.

Støtte forskning og innovasjon rettet mot trygg KI
Vi trenger mer forskning på og utvikling av tekniske løsninger for å overvåke trening av kraftig KI, teste KI-systemer før de publiseres og sikre at KI-systemer oppfører seg som de er ment. Dette krever finansiering eller lovverk som stimulerer til tekniske sikkerhetsløsninger. En rekke forskere har tatt til orde for at **minst en tredjedel** av staters og selskapers FoU-budsjetter for KI burde gå til å sikre trygg og etisk KI.[23]

Norge burde:
- Følge Sikkerhetsfaglig råds anbefaling om å utvide det nyetablerte senteret for anvendt kryptografi til et innovasjonssenter for sensitive teknologier.
- Vinkle KI-milliarden mot forståelighet- og sikkerhetsforskning.

Styrke datadeling for samfunnsnyttig KI
Bevissthet rundt risikoer må ikke stoppe oss fra å hente ut gevinster fra KI. De fleste anvendelsesområder for KI har lav risiko. Digitaliseringsnivået i Norge gir oss tilgang på unikt verdifulle datasett, men det hjelper ikke om de er utilgjengelige for forskerne som kan bruke dem. Særlig innen helse, der KI har stort potensiale, men det er svært vanskelig for både forskere og bedrifter å få tilgang til data, trenger vi enklere tilgang. Vi trenger en offensiv politikk som kan sikre samfunnsnyttig bruk av data uten at det går på bekostning av folks privatliv og kontroll over data som berører dem.

Norge burde:
- Senke kravene for bruk av helsedata innen forskning og utvikling.
- Vurdere å legge til rette for et marked for data, som sikrer enkeltpersoner kontroll over sine data og kompensasjon ved bruk.

Tilpasse og håndheve lovverket for å ansvarliggjøre KI-utviklere for ulykker eller misbruk av modellene deres
Det hersker bekymring for at EUs AI Act (KI-forordningen) blant annet ikke vil dekke de største modellene og ikke skiller tilstrekkelig mellom de som bruker og de som påvirkes av KI-systemer. Det er to av flere grunner til at vi ikke utelukkende lene oss på EU.

Norge burde:
- Sikre at utvalget som ser på KI-forordningen, KI og norsk lov har nok ressurser til å gjøre en grundig gjennomgang av lovverket, slik at tvilstilfeller kan avklares og eventuelle tilpasninger som kreves blir oppdaget.
- Avklare hvilke hendelser vi kan forvente at utviklere av KI-modeller står ansvarlig for. En slik avklaring vil gjøre det mulig for rettsvesenet å ansvarliggjøre utviklere for ulykker og misbruk gjennom objektivt ansvar.
- Vurdere mer spesifikk ansvarsfordeling for verdikjeden til KI og KYC-krav for tilbydere.

Ta aktiv del i internasjonale samarbeid for å regulere utvikling av de kraftigste KI-modellene

De kraftigste modellene lages utenfor Norge, og en reell sikkerhetspolitikk for kraftig KI krever internasjonal koordinering. Opp mot AI Safety summit i Storbritannia har det kommet flere forslag til internasjonale organer og avtaler for å redusere risiko knyttet til kunstig intelligens.[24] [25] [26] [27] [28] [29] Blant forslagene er å innføre et CERN-aktig internasjonalt forskningssamarbeid for trygg KI, IAEA-aktig samarbeid for håndhevning av internasjonale avtaler og felles sikkerhetsavtaler, monitorering og regulering av tilgang til datakraft. Norge må involvere seg i disse diskusjonene, for å lære fra internasjonal ekspertise og for å passe på at norske interesser blir ivaretatt i epokegjørende beslutninger.

Norge burde:

- Etablere en global ambassadør for KI og en avdeling på KI i UDs seksjon for global sikkerhet.
- Utarbeide en tydelig posisjon på internasjonal KI-sikkerhet innen Summit for the Future i 2024.

Vi i Senter for langsiktig politikk publiserer snart en rapport som kartlegger utfordringene fra KI, og stiller gjerne opp for utdype synspunktene våre ytterligere.

Sluttnoter

1. Wei et al. (2022), *Emergent Abilities of Large Language Models*
2. Google (2022), *Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance*
3. Financial Times (2023), *OpenAI chief seeks new Microsoft funds to build 'superintelligence'*
4. Wei (2022), *Chain-of-thought prompting elicits reasoning in large language models*
5. Casper et al. (2023), *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*
6. Hendrycks et al. (2021), *Unsolved Problems in ML Safety*
7. Fritsch et al. (2022), *An Overview of Artificial Intelligence Used in Malware*
8. The Verge (2022), *AI suggested 40,000 new possible chemical weapons in just six hours*
9. Soice et al. (2023), *Can large language models democratize access to dual-use biotechnology?*
10. Qiang (2019), *The Road to Digital Unfreedom: President Xi's Surveillance State*
11. Zuboff (2019), *The Age of Surveillance Capitalism*
12. NSM (2023), *Nasjonalt digitalt risikobilde 2023*
13. Laplante et al. (2020), *AI and Critical Systems: From hype to reality*
14. https://autonomousweapons.org/the-risks/
15. Boulanin (2020), *Artificial Intelligence, Strategic Stability and Nuclear Risk*
16. Forbrukerrådet (2023), *GHOST IN THE MACHINE*
17. Acemoglu & Johnson (2023), *Power and Progress*
18. Sterri (2022), *AI and the Transition Paradox*
19. Schick (2020) *Deepfakes: The coming infocalypse*
20. Boulanin (2020), *Artificial Intelligence, Strategic Stability and Nuclear Risk*
21. Machine Intelligence Research Institute (2022), *2022 Expert Survey on Progress in AI*
22. https://www.safe.ai/statement-on-ai-risk
23. Bengio et al. (2023), Managing AI Risks in an Era of Rapid Progress
24. Bengio et al. (2023), Urging an International AI Treaty: An Open Letter
25. Bengio et al. (2023), Managing AI Risks in an Era of Rapid Progress
26. Maas og Villalobos (2023), International AI Institutions: A Literature Review of Models, Examples, and Proposals
27. https://www.governance.ai/research
28. https://thefuturesociety.org/resources/
29. Det hvite hus (2023), FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence