18.09.2019

Forslag til nasjonal strategi for KI

Budskap

For å forberede oss på truslene og risikoene tilknyttet (ondsinnet) bruk av kunstig intelligens, så må vi forstå dem bedre. Internasjonale forskningsinstitusjoner har kommet med en rekke anbefalinger som både kan og bør følges opp også i Norge, fra regjeringens side, og i dette forslaget presenteres seks ulike måter å gjøre det på.

Bakgrunn

Kunstig intelligens (KI) kan medbringe store fordeler som påvirker mange mennesker både i Norge og i verden. Likevel, KI er en tosidig teknologi: Den kan bli brukt til både gode og skadelige formål. Kunstig intelligens påvirker trussellandskapet på spesielt tre måter: Det eskalerer eksisterende trusler, introduserer nye og kan endre truslenes karakter. Omfanget av trusler tilknyttet ondsinnet bruk av KI er stort, og begrenses ikke av sektorer eller av kommunes- og fylkesgrenser. Ei heller begrenses truslene av nasjonale grenser, og er dermed et globalt problem. The Centre for the Study of Existential Risk mener for eksempel at ondsinnet bruk av KI er blant menneskehetens ti største potensielle trusler. På samme liste finner vi klimaendringer, atomvåpen og pandemier. Norge har dermed både et nasjonalt og globalt ansvar for å forstå disse risikoene bedre og finne gode tiltak til å imøtekomme dem.

Eksempler på trusler

Å beskytte for ondsinnet bruk av KI-systemer innebærer å sikre tre sikkerhetsdomener: digital, fysisk og politisk sikkerhet. Å bevare digital sikkerhet omhandler trusler som for eksempel nettfisking, manipulasjon gjennom falsk tale eller tekst, og automatisert datahacking. Slike angrep er normalt arbeids- og tidkrevende, noe som reduserer bruken av dem, men dette kan endres ved bruk av kunstig intelligens og maskinlæring. Videre kan KI bli brukt til å true fysisk sikkerhet, for eksempel gjennom droner og automatiserte våpen. Fysiske trusler kan økes ved at flere og flere av systemene blir automatisert, for eksempel ved at noen overtar styringen av selvkjørende biler. Dersom kunstig intelligens brukes til å automatisere oppgaver innen overvåking og påvirkning kan dette ha store konsekvenser på den politiske sikkerheten. Det kan oppstå angrep på innsamlede data om menneskelig atferd og oppfatninger, som videre kan misbrukes til å undergrave demokratiske staters offentlige politiske debatter slik som vist med Cambridge Analytica og påvirkelse av det amerikanske presidentvalget i 2016.

Videre lesning

Mer detaljer om hvorfor og hvordan KI også kan ha negative effekter for Norge og resten av verden kan leses i en svært grundig rapport om ondsinnet bruk av KI utgitt av OpenAI, the Future of Humanity Institute, Centre for the Study of Existential Risk, Center for a New American Study og Electronic Frontier Foundation. Rapporten gir et mer dyptgående bilde av trussellandskapet, i tillegg til en rekke forslag til tiltak og videre forskning. For flere mer konkrete tiltak anbefales The Centre for the Study of Existential Risk sin policy-serie om å håndtere globale risikoer.

Forslag

Det er først når vi har fått en forståelse av risikoene at vi kan vite hvordan å best redusere dem, forberede oss til dem og legge handlingsplaner for å imøtekomme dem. Forslagene nedenfor bidrar til dette. De er basert på artikler, rapporter og samtaler fra Future of Humanity Institute, Centre for Study of Existential Risk, Centre for the Governance of AI, OpenAI, Google DeepMind og Leverhulme Centre for the Future of Intelligence. I den nasjonale strategien bør det legges opp til at regjeringen skal:

1. Legge frem en Stortingsmelding om risiko og sårbarheter tilknyttet ondsinnet bruk av kunstig intelligens, samt en visjon for å sikre mot slike trusler.
2. Bestille en kartlegging av økonomisk skade og eventuelle budsjettallokeringer tilknyttet risikoscenarier om ondsinnet bruk av kunstig intelligens. Samt inkorporere negative eksternaliteter og langsiktig diskonteringsrente i budsjettprosessene tilknyttet KI.
3. Opprette en gruppe i etterretningstjenesten som gjennomfører analyser og simuleringer av angrep av risikoer tilknyttet utvikling og bruk av kunstig intelligens.
4. Finansiere forskning innen disse tre områdene:
    - *Lærdommer fra Cyber-feltet*. Det er stor sammenheng mellom cybersikkerhet og KI-relaterte angrep, og vi bør derfor bruke lærdommer fra det etablerte feltet.
    - *Utforske ulike nivåer av åpenhet*. Det er farlig å publisere all forskning om kunstig intelligens, men det er likevel viktig å kunne dele informasjon og kunnskap. Derfor bør det forskes mer på hvordan normer og institusjoner kan brukes for å sikre et passende nivå av åpenhet.
    - *Utvikle tekniske og politiske løsninger*. Vi trenger mer forskning innen tekniske løsninger på ondsinnede angrep ved bruk av kunstig intelligens, og hva slags politikk som best egner seg for å forebygge slike risikoer.
5. Utvikle et rammeverk for risikovurdering og -håndtering av ondsinnet bruk og uforutsette hendelser tilknyttet utvikling og bruk av kunstig intelligens (se for eksempel Rammeverk for autentisering og uavviselighet i elektronisk kommunikasjon med og i offentlig sektor). Et slikt rammeverk bør inkludere
    - et sett med nasjonale "verdier" (kategorisert under menneskelig kapital, sosial kapital, økonomisk kapital og naturressurser),
    - en delt forståelse av midler og aktører som beskytter disse verdiene, (typ systemer, retningslinjer, prosesser, organisasjoner og infrastrukturer),
    - og en analyse av sårbarheter tilknyttet disse midlene.
6. Gi finansiering til sentre og institutter som studerer politikkdrevet forskning innen katastrofale risikoer, deriblant tilknyttet utvikling og bruk av kunstig intelligens.

Ta kontakt om vi kan bidra ytterligere, for eksempel gjennom å introdusere dere til forskningssentre og -institusjoner i vårt nettverk som har skrevet rapporter og forslag til blant annet EU, Storbritannia og USA.

Med vennlig hilsen,
Eirin Evjen
Generalsekretær i Effektiv Altruisme Norge
eirin@effektivaltruisme.no