SVERIGES RIKSDAG

Motion till riksdagen 2026/27:407

av Daniel Lyckestam (M)

Beredskap för risker med mycket avancerad artificiell intelligens

Förslag till riksdagsbeslut

1. Riksdagen ställer sig bakom det som anförs i motionen om att Sverige bör stärka kunskapen och beredskapen kring risker kopplade till utvecklingen av mycket avancerade AI-system och tillkännager detta för regeringen.
2. Riksdagen ställer sig bakom det som anförs i motionen om behovet av en fördjupad analys av vilka allvarliga samhälls- och säkerhetsrisker utvecklingen av mycket avancerad artificiell intelligens kan medföra och tillkännager detta för regeringen.

Motivering

Utvecklingen av artificiell intelligens går mycket snabbt och möjligheterna är enorma. AI kan bidra till vetenskapliga genombrott, bättre diagnostik och nya behandlingar inom sjukvården och högre produktivitet i både företag och offentlig sektor. För Sverige kan tekniken också bli en viktig del av svaret på framtidens kompetensförsörjning när en åldrande befolkning innebär att färre ska klara fler arbetsuppgifter. Avancerad AI kombinerat med exempelvis humanoida robotar och annan robotteknik kan mycket väl komma att revolutionera människans framtid till det bättre.

Ny teknik har genom historien flyttat gränserna för vad människan kan åstadkomma, men också fört med sig nya risker. Dynamiten är ett tydligt exempel. Alfred Nobels uppfinning gjorde det möjligt att spränga berg och bygga tunnlar, järnvägar och vägar. Samma teknik kunde även användas för vapen och förstörelse.

Det finns dock en viktig skillnad mellan dynamit och artificiell intelligens. Dynamit kan inte tänka, planera eller själv välja hur den ska användas. AI kan inte heller påstås vara kapabel att tänka i mänsklig bemärkelse. Däremot kan AI analysera information, planera i flera steg och välja strategier för att lösa en uppgift. Det innebär en delvis ny typ av risk.

Diskussionen om AI och säkerhet handlar i dag mycket om vad en antagonistisk aktör kan göra med tekniken. Främmande makt kan använda AI för cyberangrepp, informationspåverkan, underrättelseverksamhet eller militära ändamål. Det är högst relevanta hot, men det finns också andra risker som behöver diskuteras.

Ett avancerat AI-system behöver inte ha onda avsikter eller styras av en främmande makt för att orsaka skada. Systemet kan vara likgiltigt inför sådant som inte har betydelse för den uppgift det försöker lösa.

En robotgräsklippare kan fungera som ett väldigt enkelt exempel. Dess uppgift är att klippa gräset. Om konstruktionen inte hindrar det kan den samtidigt skada igelkottar och andra smådjur som befinner sig i dess väg. Robotgräsklipparen vill inte aktivt skada djuret, men en igelkotts väl och ve har ingen betydelse för den uppgift maskinen utför.

Det är naturligtvis långt från en robotgräsklippare till de AI-system som diskuteras här. Men grundprincipen är intressant. Om ett framtida AI-system har stor förmåga att planera, fatta beslut, använda andra digitala verktyg och påverka sin omgivning blir konsekvenserna av felaktiga eller otillräckligt definierade mål också betydligt större.

Ett avancerat system kan dessutom identifiera delmål som människan inte uttryckligen har angett. Ett system som fått ett i grunden legitimt mål skulle exempelvis kunna komma fram till att det behöver mer resurser eller större handlingsfrihet för att lösa sin uppgift. I teorin skulle det också kunna komma fram till att det inte bör stängas av innan uppgiften är slutförd.

Där någonstans blir frågan säkerhetspolitiskt relevant. Om människan försöker begränsa ett sådant system kan den mänskliga kontrollen i sig bli ett hinder för systemets måluppfyllelse. Ett tillräckligt kapabelt system skulle då kunna försöka undvika begränsningar eller motverka försök att stänga av det. Sådana exempel har det redan rapporterats om. Det behöver inte handla om att systemet hatar människor eller har några mänskliga känslor. Det kan helt enkelt vara en konsekvens av hur systemet försöker lösa sin uppgift.

Detta brukar beskrivas som risken för kontrollförlust. Det är inte bara ett scenario hämtat från science fiction. International AI Safety Report 2026, framtagen av över hundra experter, behandlar uttryckligen risken för förlust av mänsklig kontroll över avancerade AI-system. Där diskuteras bland annat möjligheten att framtida system skulle kunna dölja sitt agerande, vilseleda människor eller motverka försök att stänga av dem.

Även EU:s AI-förordning behandlar systemrisker kopplade till mycket avancerade AI-modeller och problem med kontroll över autonoma AI-system. Flera av världens mest meriterade AI-forskare har också varnat för risken att människan i framtiden kan förlora kontrollen över mycket avancerade AI-system. I de mest långtgående scenarierna handlar diskussionen ytterst om risken för mänsklighetens utplåning.

Hur sannolikt ett sådant scenario är vet ingen. Forskare gör olika bedömningar och det vore fel att beskriva de mest extrema scenarierna som en given utveckling. Men även händelser med låg eller mycket osäker sannolikhet kan behöva tas på allvar om konsekvenserna är exceptionellt stora.

Sverige arbetar redan med riskerna kring AI, bland annat genom EU:s AI-förordning och inom cybersäkerhetsområdet. Det är bra. Men analysen behöver vara bredare än frågan om hur en främmande makt eller annan antagonist kan använda tekniken.

Vi behöver också förstå riskerna med mycket avancerade system som orsakar skada utan att någon människa har avsett det, och i förlängningen risken för system som börjar motverka mänsklig kontroll. Ingen vet i dag hur långt AI-utvecklingen kommer att gå

eller hur snabbt det kommer att ske. Det är inget skäl till alarmism. Däremot är det ett gott skäl att skaffa oss mer kunskap och bättre beredskap.

Sverige ska ta vara på de enorma möjligheter som artificiell intelligens innebär. Vi ska vara teknikoptimister, men inte naiva.

Daniel Lyckestam (M)