Teknologi og digitalisering

I denne spalten vil Lars Erlend Leganger og noen av hans kollegaer skrive om aktuelle temaer innen teknologi/digitalisering som direkte eller indirekte også vil påvirke revisors hverdag. Lars Erlend er AI-ekspert og direktør i PwC. Han har en PhD i teoretisk fysikk fra NTNU.

Hva er viktigst?

AI-etikk vs. AI-sikkerhet

Vil intelligente maskiner alltid tjene mennesker, eller kan de en dag overgå oss? Noen ser intelligente maskiner kun som verktøy, der alt vi trenger å frykte er direkte og indirekte konsekvenser av menneskers (mis)bruk. Andre aner en fremtid hvor kunstig intelligens kan bli menneskets overmann, med potensielt katastrofale konsekvenser.

PhD
Lars Erlend Leganger
Direktør i PwC

Hvordan kan forskjellige eksperter, som alle har det samme målet om trygg bruk av kunstig intelligens til menneskehetens beste, være så fundamentalt uenige om hva som kan gå galt, og hva som er rett vei til mål?

The Robots can do everything?

Da ordet «robot» ble introdusert for verden av Karel Čapek i 1920,[1] var det for å belyse en problemstilling som er like relevant i dag som den var for hundre år siden: Vil intelligente maskiner – roboter – alltid forbli bare nok et av den fjerde industrielle revolusjons fremskritt – som automatiserer og forbedrer forskjellige verdikjeder på samme måte som boktrykkerkunsten, dampmaskinen, og oppgangssagen, men alltid til syvende og sist kontrollert av mennesker? Eller er vi nå i ferd med å bygge noe *mer* enn bare maskiner – kan kunstig intelligens (AI – fra engelsk «artificial intelligence») en dag overgå menneskelig intelligens? Og hva skjer da?

1 Karel Čapek, *R.U.R. (Rossums Universelle Roboter) (1920)*, https://www.gutenberg.org/files/59112/59112-h/59112-h.htm

Fra R.U.R. akt to

Helena. *Do you hate us? Why?*

Radius. *You are not as strong as the Robots. You are not as skillful as the Robots. The Robots can do everything. You only give orders. You do nothing but talk.*

Helena. *But someone must give orders.*

Radius. *I don't want a master. I know everything for myself.*

Helena. *Radius! Doctor Gall gave you a better brain than the rest, better than ours. You are the only one of the Robots that understands perfectly. That's why I had you put into the library, so that you could read everything, understand everything, and then, oh, Radius—I wanted you to show the whole world that the Robots are our equals. That's what I wanted of you.*

Radius. *I don't want a master. I want to be master over others.*

Helena. *I'm sure they'd put you in charge of many Robots. You would be a teacher of the Robots.*

Radius. *I want to be master over people. (Head up. Pride.)*

</R.U.R>

Fra Karel **Čapeks** *stykke «R.U.R.» som hadde premiere i januar 1921*

AI-etikk vs. AI-sikkerhet

Engasjementet rundt spørsmålet om hva som er de største risikoene med AI har variert i takt med interessen for AI-feltet for øvrig. Med det siste årets AI-fremskritt og produktlanseringer innen generativ kunstig intelligens, ledet an av OpenAIs GPT-maskinlæringsmodeller og ChatGPT-app, har debatten fått et oppsving. Grovt forenklet er det to leirer i debatten om hva de viktigste AI-risikoene er, og hvordan de må håndteres.

AI-etikk-leiren

AI-etikk-leiren mener verktøy basert på kunstig intelligens er (og vil forbli, i overskuelig fremtid) passive og verdinøytrale verktøy som mennesker kan bruke til godt og vondt. De største risikoene i dag og i morgen knyttet til AI er at mennesker, organisasjoner, og stater benytter AI-løsninger på uvettige og/eller uetiske måter.

Konkrete eksempler er (mis)bruk av billig arbeidskraft og åpent tilgjengelige data for datafangst til trening av maskinlæringsmodeller,[2] uforsiktig bruk av historiske data i modelltrening som lærer maskinlæringsmodeller å fortsette uønsket adferd i (potensielt) stor skala,[3] og økte ulikheter i samfunnet ved at AI-gevinstene havner i lomma på kapitalsterke eiere og utvalgte grupper (høyt utdannet) arbeidere.[4] AI-etikk-leiren ser konturene av en fremtid der AI-teknologi fører til at statsforvaltningens beslutninger tas kaldt og maskinelt, hvor de store teknologiselskapene får enorm makt og innflytelse, der de rike blir rikere, mens kanskje de eneste fra arbeiderklassen som kommer godt ut av det hele er teknologene som bygger AI-løsningene (og revisorene som reviderer dem)?

AI-sikkerhet-leiren

AI-sikkerhet-leiren er ikke nødvendigvis ubekymret for problemstillingene AI-etikk reiser, men de mener sannsynligheten for – og konsekvensen av – at det oppstår kunstig intelligens som overgår menneskelige kapabiliteter er så høye at alt annet blir trivielt. Det er i forberedelser på en fremtid med kunstig intelligens som overgår mennesker at alle gode krefter nå må settes inn. Grunntanken blant AI-sikkerhet-tilhengerne er at hvis (eventuelt når) kunstig (super) intelligens overgår menneskelig intelligens, kan en av to ting skje:

Vil AI ta godt vare på mennesker, og skape en langt bedre verden eller ...
(Bildet er generert ved hjelp av genAI-applikasjonen Midjourney).

Fremtidsscenario 1: Godsinnet super-AI

I fremtidsscenario 1 tar godsinnet super-AI (direkte eller indirekte) over styringen av samfunnet, tar godt vare på mennesker, og skaper en langt bedre verden for fremtidens menneskehet enn det tusener av år med krig og konflikt antyder at vi ville fått til på egenhånd. Scenarioet ligner det som beskrives i Iain M. Banks' science-fiction bøker om «The Culture», og menneskeheten kommer rimelig greit ut av det hele etter de fleste målestokker.

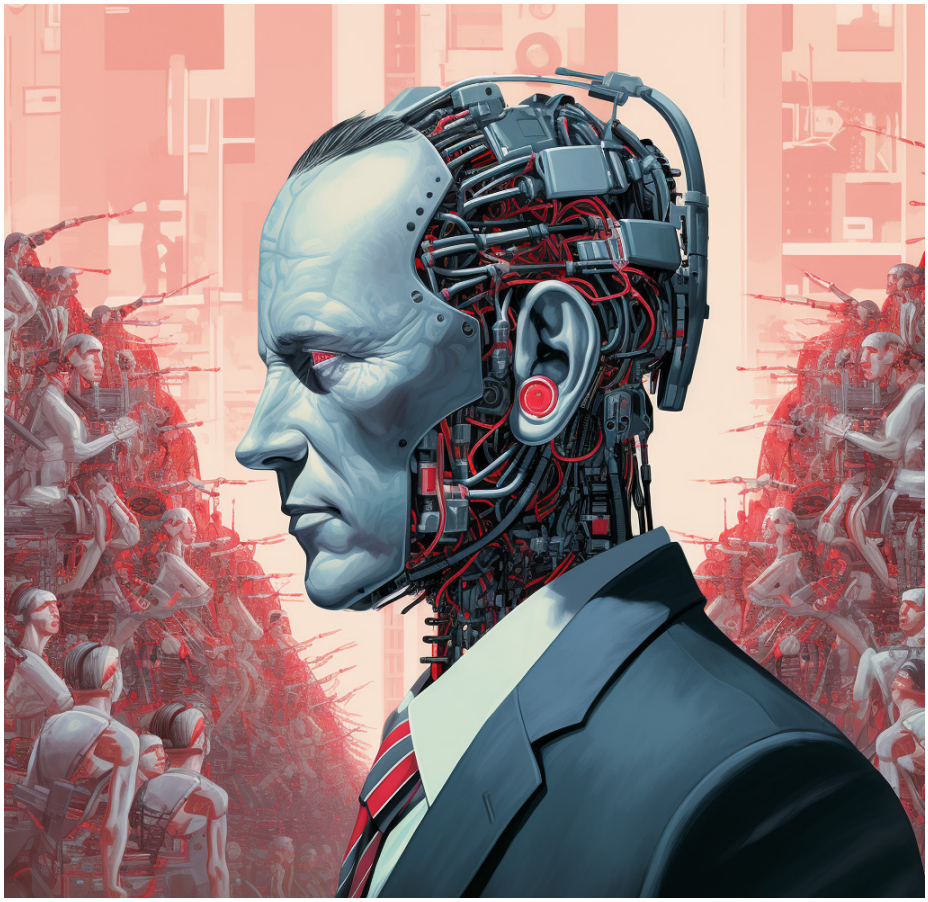
Men det er også et alternativt, mørkere fremtidsscenario ...

Fremtidsscenario 2: Ondsinnet super-AI

I fremtidsscenario 2 tar ondsinnet super-AI tar (mer eller mindre fredelig) over makten og utkonkurrerer/utrydder menneskeheten. For dette alternativet er kanskje Skynet fra James Camerons Terminator-filmer den mest nærliggende sci-fi-referansen.

Nettopp det at vi må til science fiction-bøker og -filmer for å finne konkrete «eksempler» på AI-sikkerhet-leirens risikobilde, brukes ofte som et argument for å avskrive bekymringene. Det er derfor verdt å merke seg at AI-sikkerhet-leiren ikke er begrenset til sci-fi-entusiaster og dommedagsprofeter – den inkluderer (også) Turing-prisvin-

2 Se f.eks. Fort et al, *Last Words: Amazon Mechanical Turk: Gold Mine or Coal Mine? (2011) https://aclanthology.org/J11-2010.pdf, og Gray & Suri, Ghost Work (2019) https://ghostwork.info/*
3 Se f.eks. Angwin et al, *Machine Bias, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*
4 Se f.eks. Zuboff, *The Age of Surveillance Capitalism (2019)*



... vil AI utkonkurrere/utrydde menneskeheten?
(Bildet er generert ved hjelp av genAI-applikasjonen Midjourney).

nerne Geoffrey Hinton og Yoshua Bengio (hjernene bak mange av de siste årenes viktigste gjennombrudd innen maskinlæring), OpenAI-grunnleggeren Sam Altman og Microsofts Bill Gates.[5]

I mars 2022 skapte et opprop fra AI-sikkerhet-miljøet om å *«... immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.»*[6] mye medieoppmerksomhet – og i skrivende stund har over 30 000 forskere og andre interessenter sluttet seg til oppropet.

Med et utgangspunkt om nyttemaksimering (utilitaristisk tilnærming) innebærer det første fremtidsscenarioet enorme positive konsekvenser, mens det andre scenarioets negative konsekvenser kan nærme seg uendelig – litt avhengig av hvordan en tallfester nytteverdien av menneskehetens fortsatte eksistens. Med en matematisk tilnærming der en risikos viktighet kan utledes fra sannsynligheten for at risikoscenarioet inntreffer multiplisert med konsekvensen av at det inntreffer, er det fort gjort å regne seg frem til at tiltak som øker sannsynligheten for det positive fremtidsscenarioet og/eller reduserer sannsynligheten for det negative fremtidsscenarioet er det suverent viktigste vi kan gjøre på AI-fronten for tiden.

Er én AI-risiko i hånden viktigere enn ti på taket?

AI-sikkerhet-problemstillingene har fått mye medieoppmerksomhet det siste året. Dette har skapt en del frustrasjon i AI-etikk-leiren, som mener engasjementet rundt langsiktige potensielle fremtidsscenarioer fjerner fokus fra deres mer konkrete og dagsaktuelle problemstillinger.[7]

Både AI-etikk og AI-sikkerhet-miljøene ønsker ansvarlig utvikling og bruk av AI til det beste for menneskeheten. Hvordan går det an å bli så uenige om riktig vei til målet? Det er særlig to uklarheter som gjør AI-risiko til et komplisert og til dels konfliktfylt tema: «Diskonteringsrenten» for fremtidige generasjoners velferd, og usikkerhet rundt sannsynligheten for at super-AI i det hele tatt vil bli en realitet.

«Diskonteringsrenten» for fremtidig lykke og lidelse

For økonomer er diskonteringsrente et kjent konsept: En krone i dag har en annen verdi enn en krone i fremtiden. Ved å bruke diskonteringsrente kan man omregne fremtidige inntekter eller utgifter til dagens verdi, noe som gir et bedre bilde av den faktiske verdien av en investering eller et prosjekt. Størrelsen på diskonteringsrenten kan ha stor innvirkning på beslutninger: En høyere rente vil redusere nåverdien av fremtidige kontantstrømmer, mens en lavere rente vil øke den.

Tilsvarende kan en anvende høy eller lav «diskonteringsrente» for verdien av fremtidige generasjoners lykke og lidelse sammenlignet med dagens generasjoner. Lav diskontering av fremtidige generasjoners velferd taler for mer drastiske tiltak i dag, enten det gjelder oljefond-pensjonssparing, bekjempelse av global oppvarming, eller sikringstiltak mot eksistensielle AI-risikoer. Omvendt, med høy diskontering av fremtidige generasjoners potensielle velferdsutsikter, blir det viktigste i dag å løse konkrete problemer i dagens samfunn.

Et verdispørsmål

Hva som er «riktig» diskontering er til syvende og sist et verdispørsmål uten noe objektivt fasitsvar. Kontrovers rundt dette spørsmålet er da heller ikke noe nytt: I forbindelse med bekjempelse av global oppvarming har en tilsvarende debatt rast i flere tiår.[8]

5 Se f.eks. tilslutningslisten for Center for AI Safetys opprop *"Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."* https://www.safe.ai/statement-on-ai-risk

6 https://futureoflife.org/open-letter/pause-giant-ai-experiments/

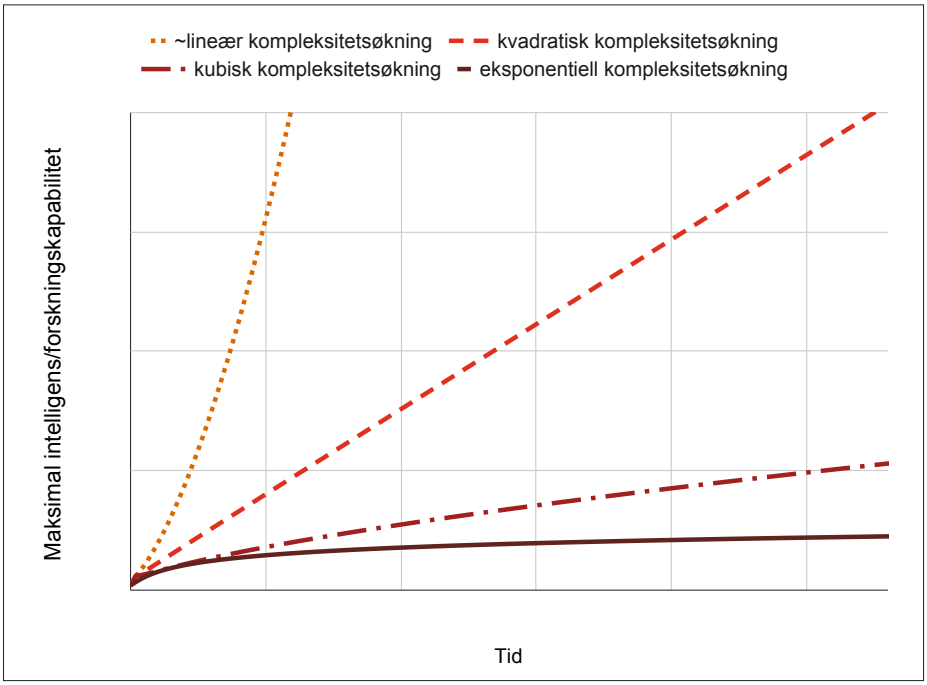
7 Se f.eks. Gebru et al, *Statement from the listed authors of Stochastic Parrots on the "AI pause" letter,* https://www.dair-institute.org/blog/letter-statement-March2023/ og Bender, *Talking about a 'schism' is ahistorical,* https://medium.com/@emilymenonbender/talking-about-a-schism-is-ahistorical-3c454a77220f

8 Se Ackerman, *Debating Climate Economics: The Stern Review vs. Its Critics (2007)* https://www.bu.edu/eci/files/2019/06/SternDebateReport.pdf for en oppsummering av kontroversen rundt Nicholas Sterns Review on the Economics of Climate Change.

Sannsynligheten for at kunstig super-intelligens blir en realitet?

Det er en lang rekke kumulative vilkår som må på plass for at et kunstig super-intelligens-fremtidsscenario skal bli en realitet: Det er usikkert hvor mye data og regnekraft som kreves for å skape mer avansert intelligens (se figur 1), og hvor fort tilgjengelig regnekraft vil øke. Kanskje finnes det et tak for hvor avansert intelligens dagens maskinlæringsalgoritmer kan skape, uansett hvor mye data og regnekraft som er tilgjengelig. Og selv om en skulle lykkes i å bygge kunstig intelligens som overgår menneskelig evne til å hente innsikt ut fra data og trekke slutninger fra den, er det noen skritt derfra til Čapeks selvbevisste og opprørske robot Radius.

Med mange faktorer som spiller inn, og stor usikkerhet knyttet til hver av dem, kan en regne seg frem til alt fra at kunstig intelligens aldri vil overgå mennesker, at det er mulig, men hører en fjern fremtid til, eller at det er rett rundt hjørnet.[9]



Figur 1: Forskjellige scenarier for når/om en kunstig intelligens-«singularitet» vil inntreffe, der kunstig intelligens brått overgår menneskelige kapabiliteter. I denne (svært forenklede) modellen er «intelligens» definert som «evnen til å forske på og videreutvikle kunstig intelligens» slik at en aktør med intelligens nivå 2 er i stand til å produsere dobbelt så mye forskning & utviklingsarbeid som en aktør med intelligens nivå 1 (aktørene kan være et menneskelig forskningsteam, en kunstig intelligens, en kombinasjon, osv.). De forskjellige linjene representerer forskjellige hypoteser om hvor mye forskning- og utviklingsarbeid som kreves for å bygge mer avansert intelligens/kapabiliteter. For eksempel betyr hypotesen «kvadratisk kompleksitetsøkning» at det er fire ganger så komplekst – og krever fire ganger så mye forsknings- og utviklingsarbeid – å forske på nivå 2-intelligens som nivå 1-intelligens (2^2 = 4 vs. 1^2 = 1). Avhengig av hvilken kompleksitetsmodell som best beskriver virkeligheten, kan en kunstig intelligens-singularitet være alt fra nært forestående, til å ligge fjernt/uendelig fjernt frem i tid. Om hvilken kompleksitetsmodell som er mest realistisk, er det mange sterke meninger, og lite empiri.

Intet er nytt under solen

Som fysiker ser jeg mange paralleller mellom dagens AI-forskning og atomfysikk-forskningen rundt andre verdenskrig. Her åpnet gjennombrudd i teoretisk og eksperimentell fysikk nye dører og muliggjorde teknologiske fremskritt innen mange områder, fra klimavennlig kjernekraft til medisin og romfart. Samtidig begynte et internasjonalt kappløp i utvikling og anskaffelse av stadig mer destruktive atomvåpen som, kanskje enda mer enn kunstig intelligens, utgjør en reell risiko for utslettelse av menneskeheten slik vi kjenner den.

Dersom atomfysikkens og atomvåpnenes historie kan fortelle oss noe om AI-teknologiens fremtid har AI-sikkerhet-forkjemperne en tung jobb foran seg. Når kommersielle (og ikke minst militære) anvendelser lokker, har det vist seg vanskelig å samkjøre full stans i utvikling og anvendelse av ny teknologi. Riktignok har samfunnet lykkes med regelverk og risikoreduserende tiltak som gjør det trygt å benytte atomkraft i strømproduksjon, samt bruke radioaktive atomkjerner i alt fra kreftmedisin til røykvarslere, men vi har i liten grad lykkes med å stanse videreutviklingen av – eller hindre spredningen av – atomvåpen-teknologi.

EUs AI-Act

På AI-fronten er det EU som har kommet lengst i reguleringsarbeidet. Unionens foreslåtte «AI Act» vil kreve gjennomsiktighet der AI benyttes, forby noen få «uakseptable» anvendelser, og innføre høye krav til risikoreduserende tiltak for et bredere utvalg «høyrisiko»-anvendelser.[10] Eksempelvis anses bruk av AI til statlig «sosial scoring» av innbyggere som uakseptabelt, mens bruk av AI i rekruttering havner i kategorien høyrisiko, begge deler helt i tråd med det som fremheves som sentrale risikoer av AI-etikk-leirens.

Om AI-sikkerhet-forkjemperne foreløpig ikke har nådd frem der lovgivning utformes, får det være en trøst at mange av AI-etikk-leirens foreslåtte tiltak mot dagens konkrete AI-risikoer også kan bidra risikoreduserende for AI-sikkerhets fremtidsrisikoer. Og uansett – dog trøst kanskje ikke er det rette ordet – er det fullt mulig for menneskeheten å kjøre sivilisasjonen slik vi kjenner den i grøfta – f.eks. gjennom verdensomspennende atomkrig eller ukontrollert global oppvarming – lenge før kunstig superintelligens kommer på banen.

9 Se f.eks. Roser, *AI timelines: What do experts in artificial intelligence expect for the future? (2023)* https://ourworldindata.org/ai-timelines for mer detaljer om hva AI-eksperter svarer i spørreundersøkelser om når de tror menneskelignende kunstig intelligens vil bli en realitet.

10 Hvordan etterlevelsen av EU AI Acts krav skal revideres er et interessant spørsmål for fremtidens (intern)revisorer!