Rådet for
Digital Sikkerhed

Kunstig intelligens og de danske perspektiver

Sammenfatning af resultaterne fra Rådets medlems-undersøgelse om forventningerne til udviklingen af kunstig intelligens i Danmark

Oktober 2024

Forord

Lad os få konkretiseret udfordringerne ved kunstig intelligens

Rådet for Digital Sikkerhed har gennem mange år arbejdet for et trygt og frit digitalt samfund. Med udviklingen inden for kunstig intelligens er det helt centralt, at vi får indfanget de særlige udfordringer, som implementeringen af AI kan medføre. Både når det gælder nye former for cyberangreb, potentiel destabilisering af demokrati og samfund samt pres på individets rettigheder og de menneskelige konsekvenser det i øvrigt kan have, at algoritmer i stigende grad former vores liv. Rådets undersøgelse er et bidrag til at konkretisere de områder, vi skal være opmærksomme på, når vi drøfter, hvordan vi bedst kan udnytte de oplagte potentialer ved kunstig intelligens.

Anne Dorte Bach, næstformand for Rådet for Digital Sikkerhed

Innovation med et kritisk blik. Sådan kan man sammenfatte budskabet i Rådet for Digital Sikkerheds medlemsundersøgelse om kunstig intelligens (AI), der blev gennemført i foråret 2024.

Undersøgelsen indikerer, at de meget store teknologispring inden for kunstig intelligens lader vente på sig, og at vi derfor har et godt fundament for en innovativ, men kritisk tilgang til teknologierne i både privat og offentligt regi.

Medlemskredsen forventer, at kunstig intelligens ikke vil slå igennem med samme styrke overalt. De mest gennemgribende forandringer ventes inden for udvalgte sektorer så som sundhed, finans, kultur og kommunikation, it-teknologi og -sikkerhed samt forsvar, mens forventningen er mindre udtalt inden for energi- og vandforsyning, transport, landbrug og offentlig forvaltning.

Undersøgelsen udpeger flere kritiske udfordringer, der bør have stor opmærksomhed i den fortsatte udvikling:

- 'Gennemsigtighed og legitimitet' ved anvendelse af persondata. Det handler fx om gennemsigtighed i forhold til, hvilke persondata, der indgår i AI-træningsmodellerne, og i forhold til det digitale output, der indgår i afgørelser i forhold til borgere og kunder. Opmærksomheden samler sig også om algoritmebaseret bias, det vil sige risikoen for, at AI-systemer giver 'stigmatiserende' output i forhold til bestemte befolkningsgrupper.
- 'Cybersikkerhed' står tilsvarende højt på temalisten i kølvandet på kunstig intelligens. Risikoen for 'dataforurening', der handler om udnyttelse af svagheder i de anvendte datasæt og træningsmodeller, blev fremhævet som et væsentligt tema i undersøgelsen. Deciderede 'input-angreb', hvor fejlbehæftede instruktioner til et AI-system kan have ødelæggende konsekvenser for modellen, blev også fremhævet. Det samme blev sikringen mod såkaldte 'AI hallucinationer', hvor AI-systemet leverer selvopfundne og ukorrekte svar på forespørgsler.
- 'Det generelle tillidsniveau' i samfundet bør ifølge undersøgelsen være et vigtigt fokusområde i den fortsatte udbygning af det digitale samfund. Det handler om dagligdagens samspil mellem borgere, virksomheder og den offentlig forvaltning, hvor AI-udviklingen rummer risiko for erosion af den menneskelige kontakt og tillid. Det handler også om AI-teknologiens betydning i forhold til udbredelsen og gennemslagskraften af misinformation i det digitale miljø.

Rådets undersøgelse er selvsagt et øjebliksbillede, og ingen kender fremtidens innovationer. Så udviklingen inden for kunstig intelligens bør følges tæt. Rådets medlemsundersøgelse præsenterer vigtige pejlemærker og vil forhåbentlig bidrage til såvel den fortsatte samfundsdebat som forsvarlig digital udvikling i privat og offentligt regi.

1. Introduktion

Rådet for Digital Sikkerheds bestyrelse besluttede i foråret 2024 at gennemføre en undersøgelse om medlemmernes forventninger til udviklingen af kunstig intelligens[1] i Danmark. Efter et forarbejde i Rådets AI-faggruppe blev initiativet tilrettelagt som en online-spørgeskemaundersøgelse, som blev udsendt anonymt til hele medlemskredsen[2] – i alt ca. 80 respondenter. Der var indkommet i alt 38 besvarelser ved deadline den 15. maj. Undersøgelsens resultater blev første gang præsenteret på Rådets medlemsmøde den 20. juni.

Formålet med undersøgelsen er at indsamle rådsmedlemmernes vurdering og perspektiv på fremtidsudsigterne for AI, primært en dansk kontekst.

Undersøgelsen belyser følgende temaer, der også vil danne overskrifterne i denne sammenfatning:

- AI's gennemslagskraft og forandringspotentiale i forskellige sektorer
- Udvikling og anvendelse af forskellige typer af AI i privat og offentligt regi
- De teknologiske fremtidsudsigter inden for AI
- AI's betydning i forhold til det digitale trusselsbillede
- AI og risikobilledet i forhold til cybersikkerhed
- AI og risikobilledet i forhold til persondatasikkerhed

Undersøgelsen har afsæt i Rådets vision om et trygt og frit digitalt samfund og belyser de mere kritiske aspekter ved AI-udviklingen frem for at udpege teknologiens udviklingspotentialer i privat og offentligt regi. Det er kort sagt Rådets opfattelse, at vi i takt med at tage de nye digitale teknologier i brug bør forholde os konstruktivt til innovationspotentialet samtidig med, at vi har et kritisk blik på de mulige afledte effekter. Rådet lægger således stor vægt på, at undersøgelsen giver kvalificerede bud på opmærksomhedspunkter i forhold til udbredelse og anvendelse af kunstig intelligens.

Hermed er også sagt, at undersøgelsens resultater er et øjebliksbillede, der skal bidrage til den aktuelle debat. Det er derimod ikke hensigten, at undersøgelsen skal danne afsæt for politisk stillingtagen til hverken teknologien som sådan eller de lovinitiativer, der er undervejs i forhold til AI.

Sammenfatningen afrundes med en opsamling og perspektiver på Rådets videre indsats på AI-feltet.

1.1. Bemærkninger om spørgeramme og præsentation af resultaterne

Selvom spørgerammen ikke havde mange spørgsmål - 10 i alt med tilhørende mulighed for kommentarer – var undersøgelsen tidskrævende at besvare (den gennemsnitlige svartid var godt 20 min.). Det skyldes primært to forhold:

Først og fremmest er begrebet 'kunstig intelligens' ikke entydigt. Undersøgelsen tog afsæt i den kommende AI-forordnings brede definition, hvor det er et væsentligt aspekt, at AI overskrider ('transcends') almindelige databehandlingssystemer, idet systemet i kraft af maskinbaseret databehandling kan lære, ræsonnere og/eller modellere flere former for output baseret på flere former for datainput (tekst, video, billede eller lyd)[3].

[1] Forkortes efterfølgende 'AI' ('artificial intelligence')

[2] Besvarelserne er personlige, og ikke på vegne af den organisation, man har som primært arbejdssted

[3] Se ny betragtning 12 i Europaparlamentets vedtagne tekst til AI forordning (13. marts 2024): "The notion of 'AI system' … should be based on ***key*** characteristics of ***AI systems that distinguish it from simpler traditional software***

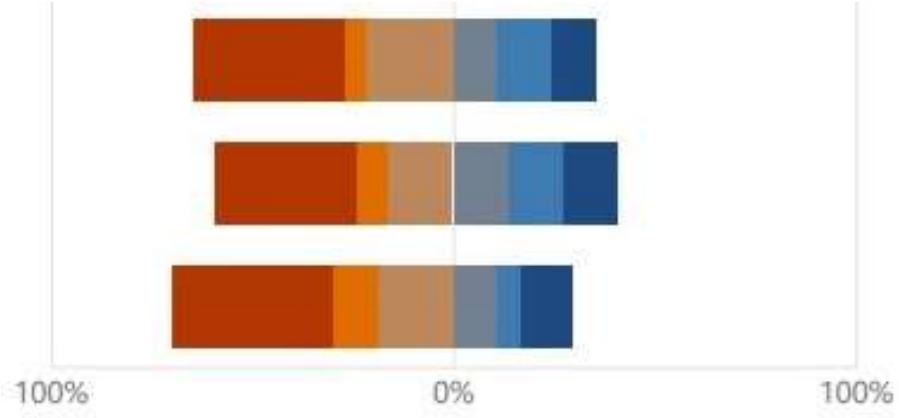
Spørgerammen havde således ikke fokus på bestemte AI-typer eller -udviklingsstadier. Kort sagt var flere af spørgsmålene meget åbne og lagde op til en kvalificeret vurdering fremtidsudsigterne.

For det andet var spørgerammen baseret på desktop-research af spørgsmålenes tematik: Definitioner, teknologiske og samfundsmæssige udviklingstendenser samt trusler og risici på AI-området. Flere spørgsmål havde en længere introducerende tekst, hvilket forøgede læsetiden (og kompleksiteten).

I tilknytning hertil bemærkes, at Rådet i sit forarbejde havde et stort ønske om, at undersøgelsen kunne medvirke til at kvalificere debatten om kunstig intelligens. Navnlig AI-faggruppen havde indgående drøftelser om temaer og vinkling af spørgsmål, hvilket afstedkom den omtalte research, der blev lagt til grund for de enkelte spørgsmål.

Sammenfattende skal det understreges, at undersøgelsen ikke er repræsentativ, men perspektiverende i sit fokus. Med en svarprocent på knap 50 % vurderes det dog, at undersøgelsen giver væsentlige pejlemærker for Rådets kommende indsats.

I forhold til præsentationen af resultaterne bemærkes, at hovedparten af spørgsmålene er opbygget omkring en afkrydsningsskala 1-3 eller 1-5 tilføjet svarmuligheden 'ved ikke'. Resultaterne angives typisk på følgende måde:



Eksemplet viser resultaterne fra 3 svarangivelser, hvor vurderingsskalaen har været 1-5 samt en 'ved ikke' kategori. De 6 farver afspejler hver især svarandelen inden for de enkelte skalatrin i procent (herunder 'ved ikke'). Hver 'bjælke' summer til 100 % (og er lige store). '0 %' - angivelsen er en markering af midtpunktet i besvarelsen. Den nederste bjælke i eksemplet indikerer således, at der er tydelig overvægt på skalatrin 1-3 (den røde, orange og brune farve), mens den midterste bjælke indikerer større spredning i besvarelserne. Den mørkeblå farve angiver i eksemplet antallet af 'ved ikke' besvarelser, og disse besvarelser burde rettelig

systems or programming approaches and should not cover systems that are based on the rules defined solely by natural persons to automatically execute operations. A key characteristic of AI systems is their capability to infer. This capability to infer refers to the process of obtaining the outputs, such as ***predictions, content***, recommendations, or decisions, which ***can*** influence ***physical and virtual environments, and to a capability of AI systems to derive models or algorithms from inputs or data. The techniques that enable inference while building an AI system include machine learning approaches that learn from data how to achieve certain objectives, and logic- and knowledge-based approaches that infer from encoded knowledge or symbolic representation of the task to be solved. The capacity of an AI system to infer transcends basic data processing, enables learning, reasoning or modelling. The term 'machine-based' refers to the fact that AI systems run on machines."*** (Texts adopted - Artificial Intelligence Act - Wednesday, 13 March 2024 (europa.eu))

ikke indgå i vægtningen i forhold til 0-punktet. Denne 'fejl' i præsentationerne vurderes dog ikke at have betydning i forhold til de resultater, som uddrages af spørgeundersøgelsen.

Spørgerammen og opsamlingen af undersøgelsens resultater er udarbejdet på grundlag af redaktion og databehandling i Microsoft Forms. Selve spørgerammen fremgår af bilag 1, mens samtlige figurer fra resultatopsamlingen fremgår af bilag 2.

2. Opsamling af undersøgelsens hovedresultater

2.1 Kort om respondentkredsen[4]

De 38 respondenter er alle fra organisationer, der er medlem af Rådet for Digital Sikkerhed. Langt hovedparten (godt 65 %) er fra private virksomheder fordelt ligeligt mellem virksomheder inden for og uden for it-branchen. Kredsen omfatter herudover universitetsansatte forskere, advokater, ansatte i forbrugerorganisationer og ansatte hos offentlige myndigheder og virksomheder.

Alle respondenter varetager typisk opgaver inden for cybersikkerhed og/eller persondatasikkerhed på deres ansættelsessted fx som sikkerheds- og complianceansvarlige, eksterne it-konsulenter, advokater, forskere eller politiske rådgivere.

Over halvdelen af respondenterne angiver i bemærkningerne, at deres ansættelsessted allerede anvender AI til interne eller bruger/kunderelaterede forretningsprocesser. I bemærkningerne angiver flere, at organisationen er afventende på grund af de tekniske, databeskyttelses- og lovgivningsmæssige udfordringer på feltet. Enkelte organisationer har iværksat pilotprojekter, hvor der fx eksperimenteres egenudviklede chatbots, og enkelte har allerede iværksat begrænsede chatbot-initiativer internt og i forhold til kunderne.

Vi er i en lettere afventende position. Afventende i forhold til det regulatoriske og i forhold til bedre basis-understøttelse (anonym)

Der eksperimenteres internt i firmaet på egenudviklede Chatbots (anonym)

Jeg har anbefalet, at administrationen kommer på kursus i de lovmæssige rammer (anonym)

Vi anvender kunstig intelligens 'baseret på syntetisk data' til interne eller bruger/kunderelaterede forretningsprocesser (anonym)

[4] Se bilag 2, spørgsmål 1 og 2.

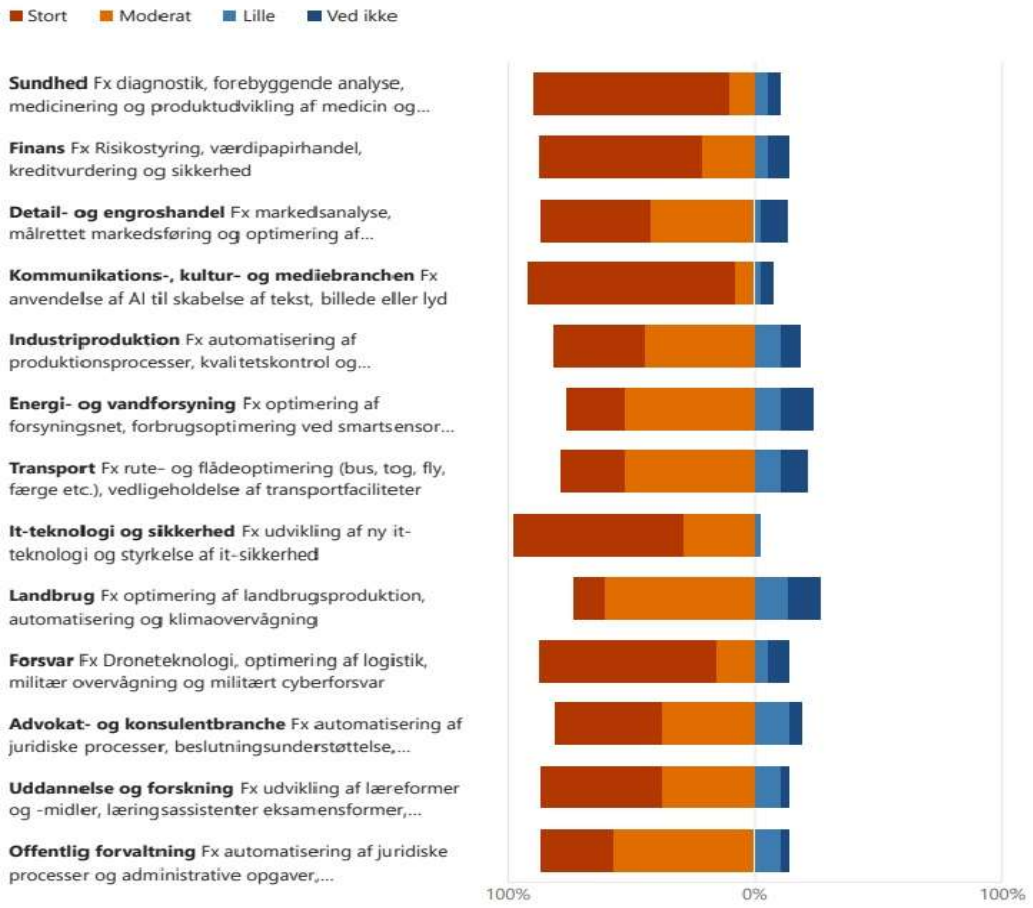
2.2 AI's gennemslagskraft og forandringspotentiale i forskellige sektorer

Det er tydeligt, at AI's store gennemslagskraft og samfundsmæssige betydning i privat og offentligt regi først ventes på længere sigt. På kort sigt (1-3 år) er den overvejende vurdering 'moderat' betydning, mens den 'store betydning' ventes inden for en længere tidshorisont på 5-10 år.[5]

Med afsæt i Danmarks Statistiks brancheinddeling (med tilføjelser) blev der spurgt mere specifikt om, hvilke sektorområder, hvor AI ventes et stort, moderat eller lille forandringsomfang inden for en længere tidshorisont.

Langt hovedparten af besvarelserne vurderer, at alle sektorer inden for den private og offentlige sektor vil blive berørt fra moderat til stort omfang.[6] Som det fremgår af figuren nedenfor, ventes de mest gennemgribende AI-relaterede forandringer inden for sektorerne: Sundhed, finans, kultur og kommunikation, it-teknologi og -sikkerhed samt forsvar, mens forandringsomfanget ventes moderat eller lille inden for energi- og vandforsyning, transport, landbrug og offentlig forvaltning.

6. **Hvordan vurderer du AIs forandringsomfang i Danmark inden for de angivne sektorområder inden for tidshorisont på 5-10 år?**



[5] Se bilag 2, spørgsmål 4 og 5.
[6] Se bilag 2, spørgsmål 6 (også gengivet i teksten, nummerangivelsen i figuren angiver spørgsmålets nummer).

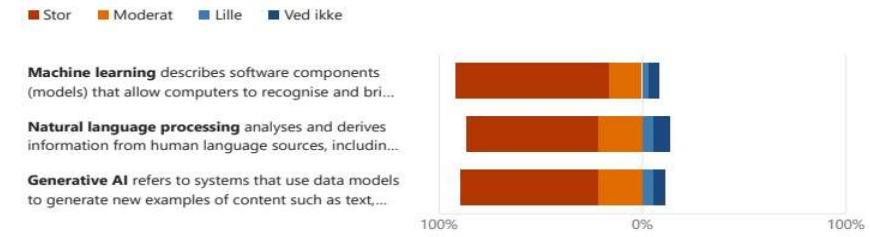
2.3 Udbredelse og anvendelse af forskellige typer af AI i privat og offentligt regi

På samme måde som der ikke findes en fast definition af AI, er det nærmest umuligt at etablere en typologi, som alle AI-modeller kan puttes ind i og som vil være stabil over tid. Imidlertid har det australske cybersikkerhedscenter (ACSC) i 2023 i samarbejde med myndigheder fra USA, Canada, UK, Tyskland, Norge og Sverige m.fl. opstillet en meget grov inddeling af forskellige typer af kunstig intelligens, idet det dog understreges, at typologien ikke er udtømmende. ACSC's AI-typologi[7]:

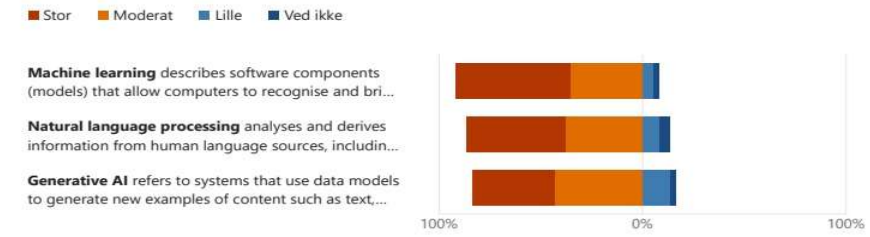
- **Machine learning** describes software components (models) that allow computers to recognize and bring context to *patterns in data without the rules having to be explicitly programmed by a human*. Machine learning applications can generate predictions, recommendations, or decisions based on statistical reasoning.
- **Natural language processing** analyses and derives information from human language sources, including text, image, video and audio data. *Natural language processing applications are commonly used for language classification and interpretation*. Many natural language processing applications not only process natural language but also generate content that mimics it.
- **Generative AI** refers to systems that use data models to generate new examples of content such as text, images, audio, code and other data modalities. Generative *AI applications are typically trained on large amounts of real-world data and can approximate human generated content from prompts*, even prompts that are limited or non-specific.

Alle AI-typer ventes at få moderat til stor udbredelse i både privat og offentligt regi. Dog forventer respondentkredsen, at anvendelsen af generative AI vil få større udbredelse i den private sektor.

8. **Med primo 2024 som nulpunkt - hvordan vurderer du de forskellige AI-typers udbredelse og anvendelse i den danske private sektor under ét inden for en tidshorisont på 5-10 år?**



9. **Med primo 2024 som nulpunkt - hvordan vurderer du de forskellige AI-typers udbredelse og anvendelse i den danske offentlige sektor under ét inden for en tidshorisont på 5-10 år?**



[7] Den australske typologi indgår i en vejledning om risici og sikkerhed i forhold til anvendelsen af AI-systemer – ACSC 2023: Engaging with Artificial Intelligence | Cyber.gov.au. Redaktionel kursivering.

Bemærkningerne i forhold til de forskellige AI-typers udbredelse betoner, at det i praksis er svært at skelne typerne fra hinanden. Typologien er 'teknisk' i sin udformning. I praksis vil AI-anvendelsen kombinere de forskellige AI-typer afhængigt af anvendelseskonteksten. Dog bliver det som nævnt indikeret, at navnlig inden for den offentlige sektor kan der være skepsis og lovgivningsmæssige barrierer for anvendelse af generative AI.

Når man taler om udviklingen af AI, tror jeg ikke, at det giver nogen mening at opdele det i machine learning, generativ AI eller natural language processing (anonym)

Tror GenAI i højere grad vil ramme det private pga. potentialet i markedsføring, underholdning mv, hvor denne type AI i det offentlige vil have mere begrænset formål og desuden være underlagt massiv regulering (anonym)

2.4 De teknologiske fremtidsudsigter inden for AI

De foregående afsnit belyste forventninger til anvendelsen af AI i en dansk kontekst. Medlemmer blev imidlertid også spurgt om deres vurdering af fremtidsudsigterne for den teknologiske udvikling på AI-området generelt set. Kernen i spørgsmålet er, hvor avanceret man forestiller sig fremtidens AI-teknologi – om AI-teknologiens formåen decideret vil kunne komme til at efterligne og måske overgå menneskelig tankevirksomhed (menneskets kognitive kapacitet).

I forhold til AI-teknologiens formåen (kapacitet, kompleksitet og potentielle anvendelse) skelnes der typisk mellem 3 'udviklingsstadier', hvor de former for AI-anvendelse, der aktuelt udbreder sig i forskellige funktionssammenhænge over hele verden, hører til første stadie ('weak AI'). De tre udviklingsstadier, som der er almindelig enighed om, er følgende[8]:

- **Narrow AI (også kaldet 'weak AI')** focuses on one narrow task and cannot perform beyond its limitations. It targets a single subset of cognitive abilities and advances in that spectrum. Narrow AI applications are becoming increasingly common in our day-to-day lives as machine learning and deep learning methods continue to develop.
- **General AI (også kaldet 'strong AI')** can understand and learn any intellectual task that a human being can. It allows a machine to apply knowledge and skills in different contexts. AI researchers have not been able to achieve strong AI so far. They would need to find a method to make machines conscious, programming a full cognitive ability set.
- **Super AI** surpasses human intelligence and can perform any task better than a human. The concept of artificial superintelligence sees AI evolved to be so akin to human sentiments and experiences that it doesn't merely understand them; it also evokes emotions, needs, beliefs, and desires of its own. Its existence is still hypothetical. Some of the critical characteristics of super AI include thinking, solving puzzles, making judgments, and decisions on its own.

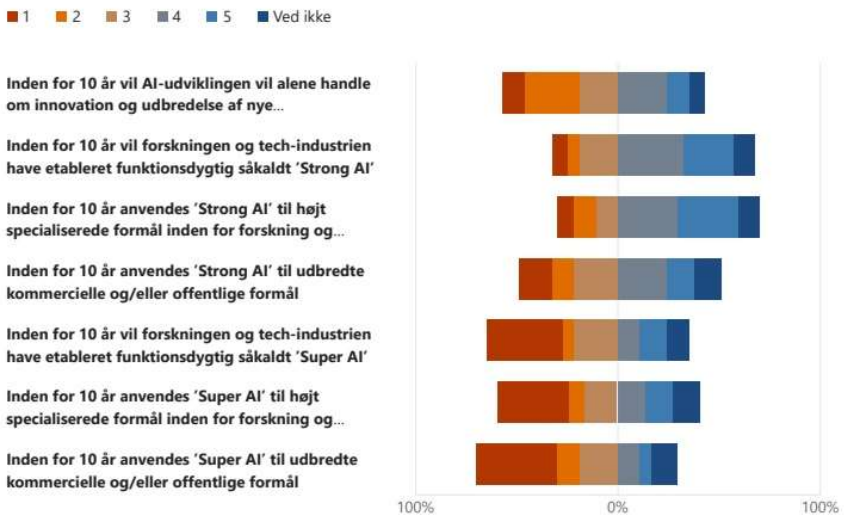
[8] De angivne udviklingsstadier er refereret fra Simplilearn 2024: Types of Artificial Intelligence That You Should Know in 2024 (simplilearn.com). Læs eventuelt mere her: 7 Types Of Artificial Intelligence (forbes.com)

I forhold til de nævnte udviklingsstadier blev medlemmerne spurgt om følgende 4 udviklingstrin inden for tidsramme på 10 år:

1. AI-udviklingen vil alene handle om innovation og udbredelse af nye anvendelsesformer inden for 'weak AI'
2. Forskningen og techindustrien have etableret funktionsdygtig såkaldt 'Strong AI' / 'Super A'
3. 'Strong AI' / 'Super AI' anvendes til højt specialiserede formål inden for forskning og afgrænsede forsvarsmæssige og/eller industrielle formål
4. 'Strong AI' / 'Super AI' anvendes til udbredte kommercielle og/eller offentlige formål

Der er stor spredning i svarene i forhold til, om AI-teknologien vil rykke fra stadie 1 til 2. Det kan hænge sammen med en forventning om, at innovation på AI-området først og fremmest vil handle om stadigt flere anvendelsesformer inden for rammen af 'Weak AI'. Den kan også skyldes tvivl i forhold til, om det rent teknisk vil lykkes at udvikle supercomputere og datagrundlag mv., der kan danne grundlag for teknologispring til mere avancerede former for AI. Respondenterne er dog mindre i tvivl i forhold til springet mellem 'Weak AI' og 'Strong AI', mens udviklingen af 'Super AI' er en fremtidshorisont, der pt.er mindre tro på samlet set.

11. **På en skala fra 1-5 bedes du angive i hvor høj grad, at du er enig i de angivne udsagn, hvor 1 helt uenig og 5 fuldstændig enig (fra et globalt perspektiv).**



Lang hovedparten af den AI der udbydes til programmeringsformål, der jo skal være en driver for udviklingen, minder mest af alt om gammeldags statistik moduler. Og f.eks. chat GPT hallucinerer og serverer skrupforkerte svar, der intet hold har i virkeligheden. (anonym)

10 år er stadig kort sigt når det kommer til udviklingen på AI fronten. (anonym)

2.5 AI's betydning i forhold til det digitale trusselsbillede

Med henblik på at tegne konturerne til udviklingen af det generelle digitale trusselbillede blev respondenterne bedt om deres vurdering af 3 overordnede trusselsaspekter ved øget anvendelse af AI:

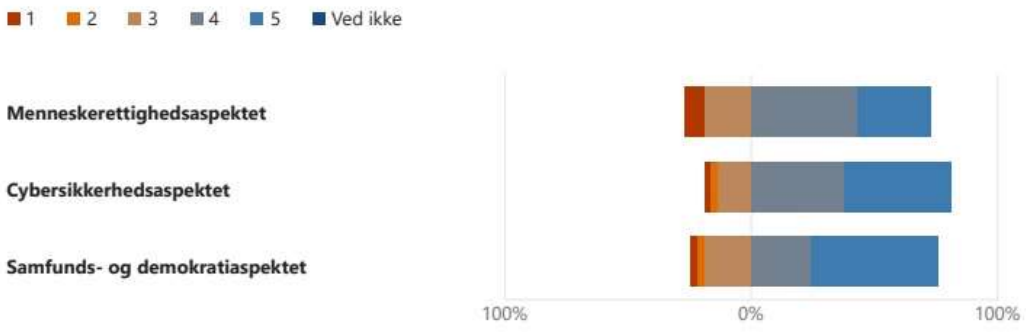
- **Det menneskeretlige aspekt.** Omfatter fx trusler i forhold til persondatabeskyttelse, gennemsigtighed i beslutningsprocesser, der har betydning for den enkelte, risici i forhold til ligebehandling, fair og retfærdige afgørelser, øget overvågning og kommerciel/offentlig profilering.
- **Cybersikkerhedsaspektet.** Omfatter fx trusler mod kritisk infrastruktur, andre danske virksomheder og offentlig sektor eller trusler i forhold til it-kriminalitet overfor borgerne.
- **Samfunds- og demokratiaspektet.** Omfatter fx trusler i forhold til det generelle tillidsniveau i samfundet (tab af tillid til digitale transaktioner, kommunikation og informationsindhentning), udbredelse af politisk relateret misinformation og vildledning.

Aspekterne er inspireret af de temaer, der er fremhævet i den såkaldte Bletchley deklaration[9], som blev vedtaget på et topmøde i London i begyndelsen af november 2023 af mere end 25 lande fra hele kloden. Bletchley-deklarationen er den første af sin art omkring AI på globalt plan og udgør et aktuelt bud på de trusselsaspekter, der knytter sig til kunstig intelligens.

Det er tydeligt, at respondentkredsen vurderer, at det samlede digitale trusselsbillede vil blive forstærket i takt med AI-udviklingen. Det bemærkes også, at såvel cybersikkerheds- som samfunds- og demokratiaspektet i særlig grad er et opmærksomhedsfelt i forhold til den fortsatte AI-udvikling, mens der er noget større spredning i svarene om AI-udviklingens betydning i forhold til udviklingen af nye menneskerettighedsrelaterede problemstillinger. Samlet set bør disse forskelle i vægtningen dog ikke tillægges stor betydning – den samlede vurdering af et forstærket trusselbillede er den overvejende tendens.

13. **AI og det generelle trusselsbillede: Hvordan vurderer du trusselsniveauet i forhold til nedenstående trusselsaspekter i Danmark inden for en tidshorisont på 5-10 år, som følge af udbredelsen af kunstig intelligens?**

    Skalaen er 1-5, hvor 1 angiver lille/ubetydelig forandring i forhold til aktuelt niveau og 5 er omfattende forøgelse af trusselsniveauet, som følge af øget anvendelse af AI.



[9] Bletchley-deklarationen 2023: The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023 - GOV.UK (www.gov.uk)

2.6 AI og risikobilledet i forhold til cybersikkerhed

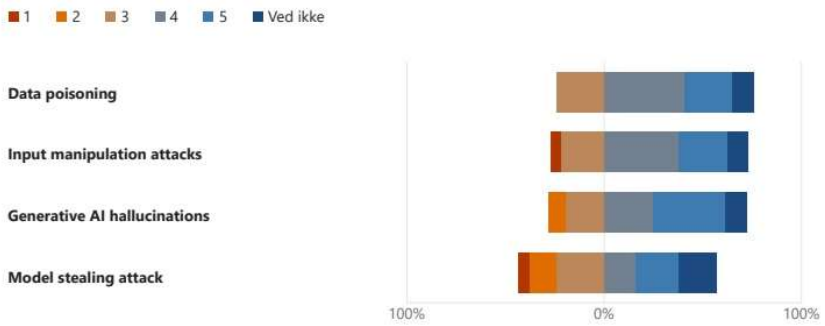
For at tydeliggøre det generelle trusselsbillede blev Rådets medlemmer spurgt mere specifikt om deres vurdering i forhold til udvalgte AI-relaterede cybersikkerhedsrisici. Cybersikkerhedsaspekterne, der redaktionelt var udvalgt til spørgerammen, er baseret på den føromtalte ACSC-vejledning.[10]

- **Data Poisoning of an AI Model.** These challenges include the potential for manipulation of training data and exploitation of model vulnerabilities (also known as adversarial AI), which can adversely affect the performance of machine learning tools and systems.
- **Input manipulation attacks.** Prompt injection is an input manipulation attack that attempts to insert malicious instructions or hidden commands into an AI system. Prompt injection can allow a malicious actor to hijack the AI model's output and jailbreak the AI system.
- **Generative AI hallucinations.** Outputs generated by an AI system may not always be accurate or factually correct. Generative AI systems are known to hallucinate information that is not factually correct. Organisational functions that rely on the accuracy of generative AI outputs could be negatively impacted by hallucinations.
- **Model stealing attack.** A model stealing attack involves a malicious actor providing inputs to an AI system and using the outputs to create an approximate replica of it. AI models can require a significant investment to create, and the prospect of model stealing is a serious intellectual property concern.

Navnlig de tre første typer træder tydeligt frem som væsentlige trusler inden for cybersikkerhedsområdet på længere sigt. Groft oversat, er der stor opmærksomhed omkring 'forurening af træningsdata', 'disruption af AI-modeller gennem input-angreb' og 'AI-hallucinationer' i form af ukorrekte og 'selvopfundne' svar fra AI-systemer. Der er større spredning i forhold til vurderingen af muligheden for at 'stjæle' en given AI-model gennem kopiering og dermed brud på ophavsretten.

15. **AI og cyberrisici: Hvordan vurderer du de angivne risicis betydning inden for en tidshorisont på 5-10 år, som følge af udbredelsen af kunstig intelligens i Danmark?**

Skalaen er 1-5, hvor 1 angiver lille/ubetydelig cybersikkerhedsrisiko i forhold til aktuelt niveau og 5 angiver en væsentlig udfordring i dansk kontekst



[10] ACSC 2023: Engaging with Artificial Intelligence | Cyber.gov.au. ACSC-vejledningens typologi refererer til US National Institute of Standards and Technologys (NIST) AI Management Framework fra 2023: AI 100-2 E2023, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations | CSRC (nist.gov)

2.7 AI og risikobilledet i forhold til persondatasikkerhed

Undersøgelsens sidste spørgsmål tog afsæt i Datatilsynets vejledning om offentlige myndigheders brug af AI[11], hvor det fremhæves som en væsentlig problemstilling, at der ikke nødvendigvis er klarhed over, om formålet med en given AI-model er tydeligt afgrænset og i overensstemmelse med det oprindelige formål med at indsamle de anvendte persondata. Datatilsynet fremhæver også:

"Borgere vil sjældent opleve direkte konsekvenser af, at deres oplysninger bruges til at udvikle en AI-løsning. Enhver behandling af personoplysninger indebærer dog risici for de borgere, hvis oplysninger det drejer sig om. Ved udvikling af AI-løsninger kan det bl.a. være en risiko for unødvendig dataophobning, da der oftest vil blive genereret særskilte træningsdatasæt baseret på myndighedens eksisterende registre mv. Det kan også være en risiko for, at myndigheden ikke tilvejebringer det samme tilsvarende behandlingssikkerhedsniveau for de træningsdata, som det er tilfældet for produktionsdata."[12]

I lyset af Datatilsynets vejledning blev respondenterne bedt om deres vurdering af følgende risici:

- **Gennemsigtighed og legitimitet** ved AI-løsninger, der anvender af persondata
- **Ophobning af persondata** i AI-løsninger
- **Databehandlingssikkerhed** i AI-løsninger, der behandler persondata

Datatilsynets perspektiv blev suppleret med bidrag fra en amerikansk undersøgelse[13] af AI og privacy:

- **Skabelse af følsomme og potentielt skadelige persondata** ved sammensætning af tilsyneladende uskadelige persondata
- **Algoritmebaseret bias**, stereotypisering og diskrimination i forhold til udvalgte grupper

Det er tydeligt i besvarelserne, at opmærksomheden samler sig om problemstillingen i tilknytning til 'gennemsigtighed og legitimitet' ved anvendelse af persondata i AI-modeller. Der er fokus på, at fremtidens AI-løsninger kan blive ugennemsigtige og illegitime fx i forhold til afgrænsning af, hvilke persondata, der indgår i AI-træningsmodellerne; i forhold til hvilket output, der potentielt kan skabes og danne grundlag for vurderinger og afgørelser samt i forhold til forskydning af formålet fra den oprindelige – og legitime - indsamling af data til ny AI-baseret anvendelse.

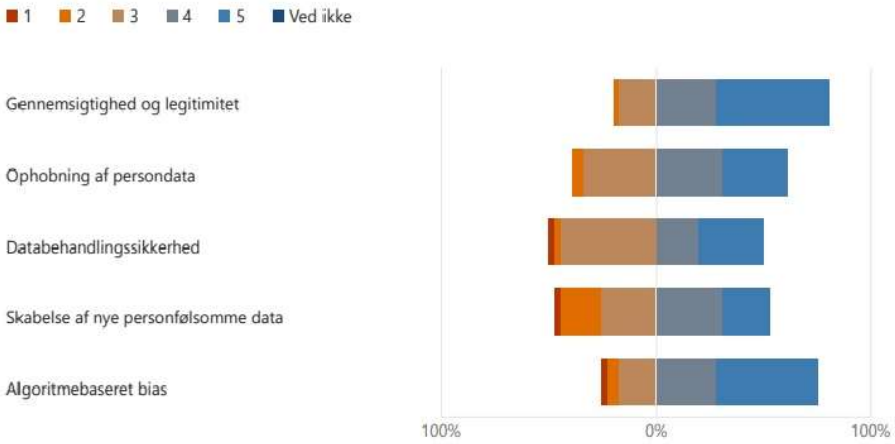
Opmærksomheden samler sig også om 'algoritmebaseret bias', det vil sige risikoen for, at AI-systemer giver misvisende eller decideret forkert output (fx vurderinger og afgørelser) på grundlag af en forvrængning i systemets databehandling og/eller træningsmodel.

Noget større spredning i svarene knytter sig til de tre øvrige datasikkerhedsmæssige problemstillinger: 'unødig ophobning af persondata', 'reduceret databehandlingssikkerhed' og 'skabelse af nye persondata ved sammenstilling af eksisterende persondata'. Det er dog en samlet vurdering, at disse problemstillinger efter respondenternes opfattelse også bør påkalde sig opmærksomhed i den fremtidige udvikling af AI-systemer.

[11] Datatilsynet 2023: Ny vejledning om offentlige myndigheders brug af AI og kortlægning af AI på tværs af den offentlige sektor (datatilsynet.dk)
[12] Datatilsynet 2023
[13] Transcend 2023: Examining Privacy Risks in AI Systems | Transcend | Data Privacy Infrastructure

17. **AI og privacy: Hvordan vurderer du de angivne risicis betydning inden for en tidshorisont på 5-10 år, som følge af udbredelsen af kunstig intelligens i Danmark?**

Skalaen er 1-5, hvor 1 angiver lille/ubetydelig privacyrisiko i forhold til aktuelt niveau og 5 angiver en væsentlig udfordring i dansk kontekst



3. Opsamling og perspektiv

I Rådets AI-faggruppe har drøftelserne på mange stræk afspejlet den generelle samfundsdebat om kunstig intelligens. Drøftelserne har derfor taget farve af de mere eller mindre dystre profetier om menneskets og samfundets endeligt. Det være sig overflødiggørelse af menneskelig arbejdskraft på selv videnstunge arbejdsområder; robotbaseret kundebetjening og forvaltning, hvor tilliden til privat og offentlig virksomhed reduceres til et spørgsmål om AI-systemernes fejlmarginer samt ikke mindst frygten for eskalering af den digitale svindel og cyberangreb som led i fremmede magters hybridkrigsførelse. Selv i forhold til det unikt menneskelige – det sociale liv og vores demokrati– er det blevet gisnet om erosion af menneskelige relationer ved udvikling af robotvenskaber og AI-baserede misinformationskampagner, der ødelægger den demokratiske dialog.

Rådets medlemsundersøgelse har udspring i Rådets vision om 'et trygt og frit digitalt samfund'. Selvom undersøgelsen ikke direkte forholder sig til de store digitale samfundsdagsordener og dystopierne, så har det været hensigten at bidrage til at få AI-debatten 'ned på jorden' ved at adressere væsentlige aspekter i udviklingen af kunstig intelligens.

Undersøgelsen har belyst Rådets forventninger og perspektiver i forhold til to hovedtemaer:

- Udviklingstendenserne på AI-området i Danmark og teknologisk set
- Trusselsbilledet med særligt fokus på cybersikkerhed og -persondatasikkerhed

Udviklingstendenserne på AI-området

Rådets undersøgelse indikerer, at udviklingen inden for kunstig intelligens først for alvor tager til inden for en længere tidshorisont. Som en af respondenterne bemærkede, er 10 år kort sigt, når talen handler om AI.

Den teknologiske udvikling inden for AI er aktuelt på udviklingsstadie, som kort kan betegnes 'weak AI'. AI-systemerne kan uanset deres træningsmodel ikke 'af sig selv' overskride deres fastlagte ramme, selvom der fx kan være indlejret avancerede sprogmodeller eller statistiske modeller i systemerne. Weak AI overskrider simple input-output regnemodeller, men maskinlæringskapaciteten er begrænset og kan ikke måle sig med menneskelig intelligens.

Respondenterne udtrykker en tvivl eller skepsis i forhold til meget store teknologispring fra 'weak AI', henover 'strong AI' til 'super AI', hvor træningsmodellerne bliver så avancerede, at de ikke blot kan efterligne menneskelig intelligens, men overskrider menneskets kognitive kapacitet.

Undersøgelsens beherskede forventning lægger op til en vis grad af besindighed eller sund kritisk tilgang til AI-udviklingen. Kræfterne bør bruges på at adressere spørgsmålet om forsvarlig AI-anvendelse, hvor der tages tydelig stilling til datagrundlaget, sikkerheden samt ikke mindst, hvad kunstig intelligens kan bidrage og ikke bidrage til i den givne kontekst. De dystopiske scenarier, der præger debatten, skal selvfølgelig have plads, for de kan medvirke til at sætte spot på de store potentielle samfundsforandringer i kølvandet på AI-udbredelsen; men de bør ikke være altoverskyggende temaer. Kort sagt indikerer undersøgelsen et credo om en innovativ, men kritisk tilgang til den fortsatte udvikling af kunstig intelligens.

I en dansk kontekst ventes de store samfundsmæssige forandringer nærmest selvsagt på længere sigt. Inden for en kortere tidshorisont på 1-3 år ventes moderate forandringer i den private og offentlige sektor. De mest markante forandringer på længere sigt ventes inden for udvalgte brancher så som sundhed, finans, kultur og kommunikation, it-teknologi og -sikkerhed samt forsvar, mens forventningen er mindre udtalt inden for energi- og vandforsyning, transport, landbrug og offentlig forvaltning.

Det brancheopdelte forventningsbillede antyder samlet set, at kunstig intelligens vil skabe forandring alle steder i erhvervslivet og den offentlige sektor, men også at de mere avancerede former for læringsmodeller og de mest omfattende konsekvenser ikke vil slå igennem i samme omfang overalt. Det kan skyldes, at innovationspotentialet ikke er oplagt, men også barrierer fx i form af meget store omstillingsopgaver, kompetencemangel, skepsis eller lovregulering.

Undersøgelsen analyserer imidlertid ikke innovationspotentialet i de enkelte sektorer, forretningsudviklingen, de organisatoriske konsekvenser eller barriererne. Det er et pejlemærke for Rådets fremtidige indsats at se nærmere på disse temaer på branche- og/eller organisationsniveau.

Trusselbilledet

Undersøgelsens belysning af trusselsbilledet, herunder cybersikkerheds- og databeskyttelsesrisiciene i kølvandet på AI-udviklingen, danner også et forventningsbillede, der giver ledetråde til Rådets kommende arbejde og fokus.

Det overordnede trusselsbillede rummer forskellige aspekter, som undersøgelsen sammenfattede under temaerne: 'Menneskerettigheder', 'Cybersikkerhed' og 'Samfund- og demokrati'. Undersøgelsen understregede forventningen til, at det samlede digitale trusselsbillede vil påkalde sig stor opmærksomhed i fremtiden, og at alle trusselsaspekter er i spil.

I tilknytning til menneskerettighedsaspektet samler opmærksomheden sig navnlig omkring 'gennemsigtighed og legitimitet' ved anvendelse af persondata i AI-modeller. Fremtidens AI-løsninger rummer risiko i forhold til fx klarhed om, hvilke persondata, der indgår i AI-træningsmodellerne, og tolkningen af det output, der potentielt kan danne grundlag for vurderinger og afgørelser i forhold til borgere og kunder. Opmærksomheden

samler sig også om 'algoritmebaseret bias', det vil sige risikoen for, at AI-systemer giver misvisende og 'stigmatiserende' output.

Cybersikkerhedsaspektet blev også nærmere undersøgt. Risikoen for 'dataforurening', der handler om udnyttelse af indbyggede svagheder i de anvendte datasæt og træningsmodeller, blev fremhævet som et væsentligt tema. Tilsvarende blev deciderede 'input-angreb', der handler om fejlbehæftede instruktioner til et AI-system, vurderet som et væsentligt tema. Endelig blev de såkaldte 'AI hallucinationer', hvor AI-systemet leverer ukorrekte og selvopfundne svar på forespørgsler, betonet som et væsentligt tema, der påkalder sig opmærksomhed.

Der blev ikke spurgt nærmere ind til samfunds- og demokratiaspektet; men det står klart, at Rådet i lyset af undersøgelsen vil have fokus på udviklingen i samfundets generelle tillidsniveau på det digitale område, herunder navnlig kunstig intelligens og teknologiens betydning i forhold til udbredelsen af misinformation i det digitale miljø.

Det bør bemærkes, at undersøgelsen ikke har adresseret holdninger og udfordringer i forhold til de mange lovgivningsinitiativer, der har betydning for udrulningen af kunstig intelligens i Danmark og Europa. Centralt i denne sammenhæng bør GDPR-forordningen, NIS2-direktivet (og tilgrænsende sikkerhedslovgivning) samt den nyligt vedtagne Forordning om kunstig intelligens (AI Act) fremhæves som væsentlige reguleringstiltag, der får stor betydning for udviklingen og den strategiske planlægning i privat og offentligt regi.

Rådet for Digital Sikkerheds vision om 'et trygt og frit digitalt samfund' er ikke indfriet, men kræver et vedvarende kritisk blik på den digitale udvikling. Rådet er grundlæggende positivt indstillet på digital innovation og ser det som en væsentlig opgave at være en 'teknologikritisk vagthund' for netop at medvirke til en 'tryg og fri' udvikling af det digitale samfund. Det er håbet, at medlemsundersøgelsen har medvirket til at belyse centrale udviklingstendenser og vigtige udfordringer i forhold til kunstig intelligens, og at der er udpeget væsentlige nedslagspunkter til den fortsatte debat om, hvordan vi i Danmark får omsat de oplagte potentialer til gavn for borgere, virksomheder og samfund.