Data & AI governance

A. INDLEDENDE OPLYSNINGER	
Aktivitetsområde	Indsatsområdet Digital sikkerhed, tillid og dataetik
Institut	Alexandra Instituttet
Titel *Dækker indholdet af aktiviteterne*	Data & AI governance
Nummerering *Af beskrivelsen*	3
Version	1
Periode *Forventet start og slut*	01.01.2024 – 31.12.2024
Kontaktperson	Kristian Krämer

B. ÆNDRINGER
Angiv her, hvis en planlagt aktivitet er ændret i forhold til den forudgående version af beskrivelsen.

C. BESKRIVELSE	
1. Mål *Hvorfor? Hvad er målet for aktiviteterne? Hvordan bidrager de til det overordnede mål for indsatsområdet?*	Kunstig intelligens er et område i hurtig udvikling, men i et voksende og innovativt felt, er sikkerhedsbetragtninger ikke altid første prioritet. Der kan derfor opstå et hul imellem den nye teknologi på den ene side og love, regulering og sikkerhed på den anden. Kunstig intelligens er i høj grad baseret på data og det er derfor afgørende for løsninger som er bygget på kunstig intelligens, at datagrundlaget er troværdigt, tilgængeligt og at følsom information ikke lækkes til uvedkommende. Det overordnede mål for dette aktivitetsområde er derfor at arbejde med datasikkerhed, med henblik på at højne sikkerhedsniveauet i virksomheders arbejde med følsomme data, og i forlængelse heraf, virksomhedernes arbejde med data som input til kunstig intelligens bl.a. dataadgang og anonymisering. Disse aktiviteter relaterer sig direkte til udviklingen af ansvarlige løsninger, som indgår i det overordnede indsatsområdes målsætning. Aktiviteterne omkring datasikkerhed vil fokusere på flere aspekter • Organisatorisk datasikkerhed, også benævnt data governance • Kryptografiske tiltag til beskyttelse, præcision, pålidelighed, tillid og distribution af følsomme data, og herunder data til brug i kunstig intelligens. • Tekniske løsninger til tillid i kunstig intelligens Data governance er de processer og regler som bruges til at håndtere og administrere data på en forsvarlig måde. Effektiv data governance en udfordring, fordi governance forudsætter at virksomheden bestrider flere discipliner som hurtigt kan blive resursekrævende. Sikkerheden omkring kunstig intelligens er udfordret af en kombination af utilstrækkelig regulering og et hurtigt udviklende marked. Dette skaber et behov for at have et skærpet fokus på hvordan datasikkerhed skal gribes an, så virksomheder får de fornødne redskaber

	til at drive data governance på en effektiv måde, uden at gå på kompromis med konkurrencedygtighed. Foruden de organisatoriske udfordringer, er det vigtigt at finde gode teknologiske løsninger der muliggør anvendelsen og distribution af følsomme data, uden at gå på kompromis med hverken personers privatliv eller proprietærer rettigheder. Det er desuden vigtigt at værdien af data bevares, også hvis de anonymiseres eller beskyttes på en anden teknisk-måde: Data kan og skal bruges som input til modeller for på den måde at kunne bruges til at generere information til at træffe beslutninger. Men når data processeres, kan det også risikere at blive misbrugt, eller at afgive unødvendigt meget information. Et centralt element i aktiviteten vil være at etablere og udføre cases, der kan demonstrere ovenstående. Konkret bidrages der til de overordnede mål-indikatorer med et eller flere caseforløb sammen med virksomheder, videreudvikling af teknologisk service indenfor AI, kompetenceopbygning, udbygning af samarbejde med vidanspartnere, samt videnspredning af resultater til dansk erhvervsliv og andre interesserede i form af eksempelvis indlæg på konferencer, webinarer, formidlingsrapporter, artikler, m.v. I forlængelse af indsatsen for at fremme AI- og datasikkerhed, er det afgørende at adressere tillidsaspektet i udviklingen af maskinlæringsmodeller. Tillid i maskinlæring er fundamentalt for at sikre, at modellerne ikke blot er præcise, men også troværdige og transparente i deres beslutningsprocesser. Dette er særligt relevant i konteksten af sprogmodeller, hvor faktuel korrekthed og forståelse af nuancer i sproget er afgørende. For at opnå dette, skal tekniske løsninger udvikles, der kan integrere og validere faktuelle data i træningsprocessen af maskinlæringsmodeller. Dette indebærer udvikling af metoder til at verificere og korrigere information, som modellerne genererer, og sikre, at de ikke reproducerer eller forstærker fejlagtige data. En sådan tilgang vil ikke alene forbedre modellernes præcision, men også deres pålidelighed og dermed brugernes tillid til disse systemer. Det er essentielt, at disse tekniske løsninger er designet til at være interoperable og kan integreres i eksisterende og fremtidige sprogmodeller. Dette vil muliggøre en bredere anvendelse og sikre, at tillid og faktuel korrekthed bliver en integreret del af maskinlæringsmodeller på tværs af sprog og anvendelsesområder. I sidste ende vil fokus på tekniske løsninger for tillid i maskinlæringsmodeller bidrage til at styrke Danmarks position inden for udvikling af ansvarlige AI-systemer, og understøtte et sikrere og mere pålideligt digitalt økosystem.
2. Indhold *Hvad skal der ske? Hvilke(n) konkret(e) aktiviteter udføres?*	Med afsæt i ovenstående målbeskrivelse udføres følgende aktiviteter i 2024: **Governance af følsomme data** Administration af data er en vigtig disciplin som skal opnå flere forretningsmæssige mål, bl.a. komplians med lovgivning, opretholde kvalitet, tilgængelighed og pålidelighed af data og sikre at data ikke lækkes til uvedkommende. Men fordi governance er så bred en disciplin, er det også ressourcekrævende at udføre, og særligt for virksomheder hvor der er mange manuelle processer. I forlængelse af arbejdet med følsomme data fra forrige projektperiode, lægges der i denne periode af indsatsen vægt på organisatoriske udfordringer med datasikkerhed. Denne aktivitet fokuserer på governance af følsomme data og kommer bl.a. til at arbejde med at: • Identificere specifikke udfordringer hvad angår data governance og herunder governance af følsomme data og AI data. • Opbygge erfaring og viden så der kan produceres vejledning og værktøjer til identifikation, klassifikation og risikovurdering af følsomme data, og særligt til AI data.

	• Producere vejledning til virksomheder i at udføre struktureret governance af følsomme data. **Tekniske løsninger på anvendelse af følsom information** Ved anvendelse af følsomme data er det bydende nødvendigt at sikre at der ikke sker et utilsigtet læk af information, da et sådant læk kan have katastrofale konsekvenser for både virksomheder og individer. Et sådant utilsigtet læk af data kan antage mange forskellige former, men mest prominent er: Læk af input data samt utilsigtet meget information i output. Dette er langt fra trivielt at forhindre og der er talrige eksempler hvor dette er gået galt[1]. Ydermere, er der mange praktiske anvendelser hvor input data er distribueret imellem flere parter, og det er ofte ikke hensigtsmæssigt at samle disse data da dette i sig selv udgør en risiko. Datadrevede beslutninger udbredes mere og mere, og derved bliver korrektheden og dokumenterbarheden af beslutningsgrundlaget i stigende grad vigtigt. Korrektheden af et datasæt indbefatter blandt andet at alle interessenter har samme datagrundlag, samt at datasættet er retvisende ift. de seneste ændringer, mens dokumenterbarheden omfatter at ændringer i datasættet kan dokumenteres for eftertiden. I denne aktivitet vil vi fokusere på at udvikle og teste metoder til at anvende tekniske løsninger som kan opnå en præcis kontrol med den information som afgives i forbindelse med komplekse anvendelser af følsomme data, og tekniske løsninger til at sikre korrekthed af data og beregninger. Der er både kryptografiske metoder og avancerede anonymiseringsteknikker, og vi vil tage udgangspunkt i cases, det kan bl.a. være med udgangspunkt anvendelse af sundhedsdata i forbindelse med sundhedsforskning. **Troværdighed i sprogmodeller** I forlængelse af TrustLLM-projektets målsætninger om at udvikle troværdige og bæredygtige store sprogmodeller (LLMs), vil dette projekt fokusere på at styrke tilliden til maskinlæringsmodeller gennem tekniske løsninger, der sikrer integration og validering af faktuel data i træningsprocessen. Projektet vil adressere udfordringerne ved at bevare faktuel korrekthed i sprogmodeller, hvilket er afgørende for at modellerne kan træffe beslutninger baseret på præcise og pålidelige data. Da LLMs er et nyt område findes der meget lidt viden om hvordan dette kan opnås, hvilket gør at de fleste aktiviteter vil være baseret på konkrete forskningsspørgsmål. Aktiviteterne kunne omfatte: - Udvikling af metoder til at indarbejde faktakontrolmekanismer i træningsprocessen for sprogmodeller, således at de genererede outputs er baseret på verificeret information. - Implementering af algoritmer, der kan identificere og korrigere fejlagtige eller forældede data, for at forhindre reproduktion og forstærkning af misinformation. - Skabelse af værktøjer til at øge gennemsigtigheden i beslutningsprocesserne for sprogmodeller, så brugerne kan forstå og stole på de resultater, de leverer. Disse tekniske løsninger vil blive designet med henblik på interoperabilitet, så de kan integreres i både eksisterende og fremtidige sprogmodeller, og dermed fremme en bredere anvendelse. Ved at fokusere på faktuel korrekthed og tillid vil projektet bidrage til at styrke Danmarks position inden for udvikling af ansvarlige AI-systemer og understøtte et mere sikkert og pålideligt digitalt økosystem.
3. Aktører *Hvem udfører aktiviteterne? Hvilken afdeling af instituttet?*	Alexandra Instituttets medarbejdere på tværs af fire afdelinger vil udføre aktiviteterne: *Artificial Intelligence & Data Analytics*, *Insights, Security lab.* og *Strategic Business & Governance.*

[1] https://latanyasweeney.org/JLME.pdf og https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

Evt. hvilke eksterne parter er med (videninstitutioner, virksomheder, erhvervsorganisationer, myndigheder, klyngeorganisationer eller andre.)	Eksterne samarbejdspartnere: • Universiteterne: I dette aktivitetsområde vil vi samarbejde med flere af de danske universiteter. Vi har samarbejder i gang med både AU, DTU, ITU og AAU, og vil både forsøge at udbygge samarbejdet med disse og indlede nye. • Klynger: Hovedsageligt DigitalLead og CenSec, men også˚andre klynger hvor denne aktivitet kan støtte op med teknologiske løsninger • Dansk Industri • Nationalt Forsvarsteknologisk Center • Security Tech Space
4. Sammenhæng med andre projekter *Indgår aktiviteten i andre eksternt finansierede projekter?*	RK-indsatsen medfinansierer og sikrer sammenhæng til følgende igangværende projekter: • CRUCIAL, støttet af Innovationsfonden, formål i dette projekt er bl.a. at anvende forretningsfølsomme data fra kritisk infrastruktur på en sikker måde. • TRUST-LLM, Horizon Europe - udvikling af Germansk sprogmodel • CoRAL, Grand Solution, DK - Indsamling af taledata og udvikling af dansk tekst-til-tale model • AI-Matters, Digital Europe, TEF, hvor formålet er at udvikle en europæisk testplatform, hvor fremstillingsvirksomheder kan teste sine AI teknologier. • European Digital Innovation Hubs: TechCircle (CD-EDIH/Midtjylland), SEDIH (Sydjylland), AI Boost (GC-EDIH/Hovedstaden) Ovenstående projekter leverer 1) behovsafdækning, udvikling, modning og pilottest af relevante TDU services i RK-indsatsen som er under udvikling og 2) relevant videnspredning til målgruppen. Øvrige projekter, der bl.a. bidrager med viden og adgang til RK indsatsens målgruppe: • AI Denmark projektet, støttet af Industriens Fond, der har til formål at understøtte SMV'er med at komme hurtigere i gang med at udnytte data og AI-værktøjer i deres forretning.
5. Følgegruppe *Har følgegruppen forholdt sig til aktiviteten? I så fald hvordan?*	For følgegruppen for Data & AI Governance, har vi haft dialog med repræsentanter fra Aalborg Universitet, Data For Good Foundation, Partisia og BlockDeamon. Vores løbende dialog har været centrale for udviklingen af aktivitetsbeskrivelsen for 2024.
6. Formidling af resultater *Hvordan/hvor kan interesserede virksomheder m.fl. få viden om resultaterne af aktiviteterne?* *Anføres/tilføjes hvis det ikke allerede fremgår af beskrivelsen ovenfor, f.eks. ved links til konferencer, hjemmeside, publikationer etc.*	Resultater og erfaringer fra denne aktivitet vil primært blive formidlet gennem aktiviteten 'Demonstration af anvendelighed og værdiskabelse' inden for indsatsområdet Digital sikkerhed, tillid og dataetik. Formidlingen laves gennem forskellige formater såsom white papers, blogindlæg samt præsentationer på konferencer og webinarer. Desuden vil materialer være tilgængeligt gennem TDU-AI og TDU-Cybersikkerhed på alexandra.dk. For at nå ud til en bredere målgruppe vil en betydelig del af formidlingen foregå i samarbejde med andre aktører og samarbejdspartnere.