Tillidsskabende AI

A. INDLEDENDE OPLYSNINGER	
Aktivitetsområde	Indsatsområdet Digital sikkerhed, tillid og dataetik
Institut	Alexandra Instituttet
Titel *Dækker indholdet af aktiviteterne*	Tillidsskabende AI
Nummerering *Af beskrivelsen*	3
Version	1
Periode *Forventet start og slut*	1/1 2022-31/12 2022
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?*	Målet med denne aktivitet er at indfri målsætningen om at sikre tillid til AI-løsninger under indsatsen *Digital Sikkerhed, Tillid og Dataetik*. For tillid til teknologien er fundamentalt afgørende for om de kæmpe samfunds- og forretningsmæssige potentialer, som følger med AI-teknologier, kan realiseres. For hvis ikke vi har tillid til, at et system er sikkert og handler i overensstemmelse med etiske overvejelser og domæneviden, så vil vi ikke udnytte det godt nok - hvis overhovedet. Aktiviteten har derfor bredt fokus på at understøtte dansk erhvervslivs udvikling af ansvarlige AI-løsninger med rådgivning og teknologisk service i alle led af udviklingsprocessen fra indledningsvis eksplorativ dataanalyse til idriftsætning og monitorering af et AI-system. Men aktiviteten har også fokus på at løsningerne skal kunne baseres på et sprog som alle danske borgere forstår. Derfor bidrager aktiviteten også med udvikling af dansk sprogteknologi og NLP med henblik på at styrke mulighederne for at udvikle AI-løsninger på dansk med sproglig forståelse og formuleringer der henvender sig til alle danskere - også personer der ikke har særlige teknologiske forudsætninger. Konkret bidrages der til de overordnede mål-indikatorer med to eller flere caseforløb sammen med virksomheder, videreudvikling af teknologisk service indenfor AI, kompetenceopbygning, udbygning af samarbejde med videnspartnere, etablering af samarbejde med én ny videnspartner, samt vidensprednig af resultater til dansk erhvervsliv og andre interesserede i form af eksempelvis indlæg på konferencer, webinarer, formidlingsrapporter, artikler, blogindlæg, m.v.
2. Indhold **Hvad skal der ske?** **Hvilke(n) konkret(e) aktiviteter udføres?**	Med afsæt i ovenstående målbeskrivelse udføres følgende aktiviteter i 2022:

	State-of-the-industry analyse af arbejdsprocesser og brug af digitale værktøjer i dansk erhvervsliv relateret til Ansvarlig AI[1] Arbejdet med state-of-the-industry analysen blev påbegyndt i 2021 og fortsættes i 2022. Formålet med analysen er at afdække hvordan danske virksomheder arbejder med digital ansvarlighed når de udvikler AI-løsninger, samt deres forventninger til fremadrettede behov og indsatser indenfor området. Resultater af analysen opsamles i en formidlingsrapport, der deles frit i relevante digitale kanaler. Viden fra analysen/aktiviteten bruges til at opdatere og videreudvikle Alexandras services indenfor AI i TDU'en, som for eksempel rådgivning om anvendelse af XAI-metoder, algoritmisk fairness og håndtering af bias i datasæt. **Udarbejdelse af køreplan for mulig fremtidig AI certificering** Aktivitetens omdrejningspunkt er at undersøge om og hvad der skal til for at Alexandra Instituttet kan opnå godkendelse til at certificere AI systemer efter implementering af EUs AI-forordning, der er under udvikling og forventes færdig i 2023. I 2022 tages de første skridt i arbejdet. Her vil aktiviteten være at hjemtage og bearbejde viden om lovarbejdet der pågår. Formålet vil være at udvikle og forberede rådgivning og services i relation hertil frem mod 2023, hvor forordningen forventes at blive implementeret – i stil med GDPR-forordningen fra 2018. I forlængelse af denne hjemtagning og bearbejdning udarbejdes en køreplan for hvordan og evt. på hvilke måder Instituttet kan blive godkendt til at certificere AI-systemer når reguleringen træder i kraft.[2] I forbindelse med arbejdet vil der blive trukket på viden og erfaringer der opbygges i forbindelse med aktiviteten *Standarder og kritiske systemer*. **DaNLP udvides med komponenter til talegenkendelse og -analyse** I 2022 udvikles minimum fem nye komponenter til DaNLP-platformen. Det drejer sig hovedsageligt om datasæt og maskinlæringsmodeller, men et udsnit af komponenterne kan også være tutorials og guidelines. To eller flere komponenter vil fokusere på talegenkendelse og -analyse, og derved åbne op for at danske virksomheder kan komme i gang med at anvende taledata. **Udvikling af danske sprogmodeller – tale og/eller tekst** Best practice indenfor anvendt NLP er forsat at benytte store såkaldte prætrænede sprogmodeller, der tilrettes en specifik opgave. Trods mange danske aktørers store arbejde de seneste år er der forsat stor forskel i mulighederne på mindre sprog som dansk og store sprog som engelsk. I tillæg er den mest anvendte danske sprogmodel fra 2019[3] og senest har en ny norsk sprogmodel vist sig at kunne løse opgaver på dansk bedre end den bedste danske model.[4] Der er med andre ord stort og tydeligt behov for at træne nye danske open source-sprogmodeller alle kan benytte sig af. I den forbindelse er det vigtigt samtidigt at sikre en smidig måde løbende at opdatere modellerne på med ny data, så de vedbliver relevante over tid. Et markant kvalitetsspring opad i performance vil desuden betyde at anvendelsesmulighederne for dansk NLP bredes ud til flere brugsområder og derved flere danske virksomheder.

[1] Refererer til indsatsbeskrivelsen (s. 2), hvor ansvarlig/ansvarlighed defineres som samlebegreb for *"(...) etiske, cybersikre, privacy-venlige, ikke-biased, tillids- og tryghedsskabende løsninger*".
[2] I det nuværende forslag skal AI-løsninger, der kategoriseres som "High-Risk AI-systems", leve op til en række forpligtigelser herunder auditering og muligvis CE-mærkes før de kan markedssættes.
[3] https://huggingface.co/Maltehb/danish-bert-botxo
[4] https://scandeval.github.io/pretrained/

	Derfor påbegyndes denne aktivitet, der handler om på sigt at blive i stand til ikke kun at udvikle, men også vedligeholde større danske sprogmodeller, som er frit tilgængelige for alle danske virksomheder via open source. Og som del heraf også udvikle en platform og arbejdsgange, så modellerne efterfølgende kan opdateres med ny data. Arbejdet i 2022 på aktiviteten vil fokusere på første etape i udviklingsprocessen og primært handle om at samle og konstruere valide datasæt i de rette mængder og kvalitet, som kan bruges til at træne robuste modeller, der baserer sig på data med bred sproglig repræsentation, hvad end det er tekst og/eller tale. Af erfaring ved vi at indsamling af data til offentlig udstilling er problemfyldt. Eksempelvis rettigheder og usikkerhed vedrørende compliance er medvirkende årsager til at virksomheder og organisationer afholder sig fra at stille data til rådighed. Derfor er der indledningsvist behov for en bred ramme for aktiviteten ved opstart. Både i forhold til modelvalg og datamodalitet. Endelig valg af model(ler) og datamodalitet foretages derfor på baggrund af dialog med relevante aftagervirksomheder, aktører i det danske open source-community, dataejere, rettighedshavere og/eller offentlige myndigheder, samt det datagrundlag det viser sig muligt at etablere. I forbindelse med aktiviteten vil der blive trukket på viden, kompetencer og resultater fra aktiviteten *Sikker brug af følsomme data*. Det drejer sig særligt om metoder og tilgange, der har potentiale til at lette arbejdet med at skabe adgang til data såsom privacy-bevarende metoder og anonymisering af ustruktureret data – i denne sammenhæng tale- og tekstdata. En mindre del af arbejdet på aktiviteten i 2022 vil desuden handle om at etablere synergiprojekter med eksterne partnere, hvor der målrettet arbejdes på indsamling af data. **Styrket samarbejde med vidanspartnere** Samarbejde med NLP-grupperne ved henholdsvis ITU og DIKU forsættes og udbygges i 2022. Med ITU kan samarbejdet eksempelvis handle om videreudvikling af NLP-baserede løsninger til moderering af hadtale på sociale medier eller en udvidelse af Danish Gigaword. Med DIKU kan samarbejdet eksempelvis udvikles gennem samarbejde om identificering af misinformation på internettet eller forsættes via det nuværende samarbejde om udvikling af danske open source-komponenter til dialogsystemer. Derudover afsøges mulighed for at indlede samarbejde med DIRECs arbejdsgruppe for AI/NLP.
3. Aktører *Hvem udfører aktiviteterne? Hvilken afdeling af instituttet? Evt. hvilke eksterne parter er med (videninstitutioner, virksomheder, erhvervsorganisationer, myndigheder, klyngeorganisationer eller andre.)*	Alexandra Instituttets medarbejdere på tværs af tre afdelinger vil udføre aktiviteterne: *Artificial Intelligence and Data Analytics*, *Human Insights* og *R&I -Business and Governance.* Centrale eksterne samarbejdspartnere er: • DIKU, Datalogisk Institut ved Københavns Universitet • ITU, IT-Universitet i København • DIREC, Dansk nationalt forskningscenter for digitale teknologier • Konsortiet bag AI Denmark; Teknologisk Institut, Danmarks Tekniske Universitet, Aalborg Universitet, ITU, DIKU samt Innovation Centre Sillicon Valley
4. Sammenhæng med andre projekter *Indgår aktiviteten i andre eksternt finansierede projekter?*	Der er synergi med projektet AI Denmark hos 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. Desuden forsætter og afsluttes Innovationsfondsprojektet AutoAI4CS i 2022, hvor Alexandra Instituttets rolle er at udvikle open source-komponenter til dansk samt anvendeliggøre og formidle projektresultater omkring udvikling af state-of-the-art chatbots. I sidstnævnte projekt anvendes aktiviteten som egenfinansiering.
5. Følgegruppe *Har følgegruppen forholdt sig til aktiviteten? I så fald hvordan?*	I den forgående periode har vi været i løbende dialog med både virksomheder, rådgivere, offentlige organisationer og universiteter gennem indsatsens netværks- og følgegrupper. Her har vi bredt identificeret markedsmæssige udfordringer og problemområder, som i

	dag ikke bliver dækket af teknologiske services. I den kommende aktivitetsperiode vil vi etablere 3 mindre følgegrupper (én for hver aktivitetsbeskrivelse), der vil være mere agile i forhold til den løbende dialog omkring markedssituationen og indsatsområdets aktiviteter. I denne aktivitet etableres en specifik følgegruppe med fokus på AI. Følgegruppen sammensættes af aktører på tværs af private virksomheder, offentlige organisationer og vidensinstitutioner, der kan bidrage positivt ind i samtlige aktiviteter, men med særligt fokus på opgaven med indsamling af data. Konkurrencesituationen vil desuden vendes løbende på møder med følgegruppen, så det sikres at der ikke udvikles komponenter og services der er konkurrenceforvridende. Parallelt med følgegruppen vil der løbende være fokus på engagement og deltagelse i det danske open source-community for data science og AI, som eksempelvis AI Denmarks community og foreningen Dansk Data Science Community. Aktivitetsbeskrivelsen er sendt i "e-mail/telefonisk høring" i følgegruppen, forinden den er uploadet på Bedreinnovation.dk. Hen over perioden præsenteres følgegruppen for fremdrift på aktiviteten på følgegruppemøderne og i den løbende ad hoc dialog.
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 fra aktiviteten udstilles via Instituttets website for DaNLP, selve DaNLP GitHub-siden, samt de to blogsider *Ansvarlig AI* og *DaNLP*. Derudover indtænkes resultaterne fra aktiviteten i Instituttets generelle vidensformidling og i arbejdet omkring den videre etablering af TDU'en. OPDATERING: Væsentlige aktiviteter og resultater opnået i 2022: • Fem komponenter udviklet og udstillet på Hugging Face (https://hugging-face.co/alexandrainst). Blandt andet har vi udviklet ny model til identificering af stødende tale (opnået SOTA), samt det første Question Answering datasæt på dansk, norsk og svensk. • State-of-industry undersøgelse af ansvarlig AI i praksis færdiggjort. • Roadmap for mulig fremtidig certificering udarbejdet. • Grand solution ansøgning (CoRal) med bl.a. Københavns Universitet om produktion af danske taledata bevilget af Innovationsfonden.