Prosjektplan

Navn på sandkasseprosjekt: Forebygging av seksuelle overgrep mot mindreårige på nett.

Navn på virksomhet(er): Politihøgskolen (PHS), Universitetet i Agder (UiA), Politiets IT-tjeneste (PIT).

Deltakere fra virksomhet:

Navn	Rolle	E-post	Telefon
Inger Marie Sunde - PHS	Prosjektleder, jurist		
Jens Erik Paulsen – PHS	Filosof, teknolog		
Lei Jiao - UiA	Data scientist		
Thomas Ibsa Beka – PIT	Data scientist		

Deltakere fra Datatilsynet:

Navn	Rolle	E-post	Telefon
Arild Opheim	Prosjektleder, kommunikasjon		
Jan Henrik Mjønes Nielsen	Prosjektmedarbeider, juridisk seniorrådgiver		
Tanja Czelusniak	Prosjektmedarbeider, juridisk seniorrådgiver		
Narjes Al-Sultan	Prosjektmedarbeider, teknologisk rådgiver		
Eirik Gulbrandsen	Faglig ressurs, senioringeniør		

1. Kort beskrivelse av prosjektet og formålet med planen

KI har vist seg å kunne identifisere mening i tekst og å kunne si noe om forfatterens sinnsstemninger og egenskaper. Formålet med prosjektet i sandkassen er å kartlegge de rettslige og etiske rammene for utvikling, testing og bruk av teknologien, og å utvikle en felles forståelse av disse gitt tverrfagligheten i prosjektet.

Tanken er at politiet skal kunne bruke teknologien for å avdekke grooming hvor en voksen overgriper kommuniserer med en mindreårig på nett. KI-modellens prediksjoner skal varsle politioperatøren og bidra til en beslutning av om det er grunnlag for å gripe inn, f.eks. med en advarsel. Erfaringer fra norske straffesaker viser at overgriper og barn kommuniserer på norsk. Teknologien må følgelig baseres på norske tekstdata.

Flere forutsetninger må være oppfylt for realisering:

1. Det må finnes tilstrekkelig mengde norske tekstdata til å kunne utvikle KI-modellen. På tidspunktet for sandkassen er det usikkert om denne forutsetningen er oppfylt, men den vil kunne realiseres over tid, gitt at teknologien/metoden er formålstjenlig. Sandkassen vil bidra til å belyse dette.

2. Groomingen må foregå på arenaer hvor politiet har lovlig tilgang.

3. Utviklingen og bruken av teknologien for det gitte formålet, må være lovlig.

4. Utviklingen og bruken av teknologien for det gitte formålet, må være i samsvar med etiske normer for politiets arbeid og ansvarlig KI.

2. Mål og forventede leveranser i prosjektet

Som trenings- og testdata i utviklingsfasen trengs taushetsbelagte tekstdata fra norske straffesaker om nettovergrep. Prosjektet har allerede en mindre mengde slike data i kraft av en tillatelse fra Riksadvokaten, jf. politiregisterloven § 33, som gir adgang til å oppheve taushetsplikten for *forskningsformål*. Spørsmål som reiser seg, er:

- For å kunne bruke tilsvarende fremtidige data i en senere *driftsfase* for å oppdatere/videreutvikle/korrigere KI-modellen, kan det være behov for lovendring. Hvilke hensyn gjør seg gjeldende for en slik lovendring?

- Hva skal til for å respektere person-/opplysningsvernet til deltakerne i tekstsamtalene, når dataene brukes i utvikling og testing av KI modellen? Spørsmålet gjelder både dataene som er omfattet av samtykket for forskningsformålet i utviklingsfasen, og fremtidige data som skal brukes i en driftsfase.

Prosjektet ønsker å kartlegge hva som kreves for at teknologien skal kunnes anses som «ansvarlig», slik at den oppfyller etiske krav og gir grunnlag for tillit til politiets bruk av den.

Til sist ønskers en diskusjon i sandkassen om rettslige og etiske sider ved bruken av teknologien, slik at politiet kan ta en informert beslutning om hvorvidt man ønsker å bruke den for det gitte formålet.

Målene kan dermed konkretiseres slik:

1. Avklare rettslige krav til behandling av tekstdata i utviklingsfasen
2. Konkretisere hva «ansvarlig KI» betyr når teknologien brukes av politiet for å analysere kommunikasjon på internett, kanskje særlig med fokus på forklarbarhet
3. Siden teknologien er generell og kan brukes av forskjellige aktører, ønsker vi å synliggjøre hensyn som kjennetegner politiets bruk av teknologien sammenlignet med andre aktørers bruk av den (tjenestetilbydere og deres underleverandører, sluttbrukere (foreldre, barn) …).

De to første målene har hovedprioritet. Det tredje blir tatt med mot slutten av sandkasseprosessen om tiden tillater det.

3. Aktiviteter, metode og arbeidsfordeling

Det vil være hensiktsmessig med workshops for å belyse utfordringene som følger med målet for sandkasseprosjektet. Med workshop mener vi et møte for med mulighet for å inkludere flere deltakere enn de som allerede er i prosjektet og i Datatilsynets sandkasse, f.eks. fra politiets nettpatrulje, personer med ekspertise på trender i grooming på nett (f.eks. fra politiet og Redd Barna), personer som kan støtte gjennomføring av prosjektet uten selv å være forskere (f.eks. politistudenter som kan bistå med «labeling» av data) mv.

Workshopene må følges opp med aktiviteter som bringer prosjektet fremover. Noen aktiviteter er nevnt i planen nedenfor, men vi antar at de må konkretiseres nærmere med Datatilsynet. Siden prosjektet er enestående i politiet kunne det vurderes å holde et avsluttende seminar og dele erfaringene med relevante interessenter og bidragsytere.

Oppstartsmøte med Datatilsynet.

Workshop 1: Funksjonsbeskrivelse av teknologien og retningslinjer for bruk

Aktivitet etter workshop 1: Følge opp kartlagte handlingspunkter.
(Rapportering til Datatilsynet?)

Workshop 2: Avklare muligheter og begrensninger i tilgangen til trenings- og testdata. Hvilke behov må imøtekommes for å kunne bruke dem i utvikling av KI modellen? Hvem kan gis tilgang og under hvilke betingelser?

Aktivitet etter workshop 2: Samarbeid mellom UiA og PIT for å etablere en sikker løsning.
(Rapportering til Datatilsynet?)

Workshop 3: Etiske problemstillinger knyttet til prosjektet.

Aktivitet etter workshop 3: Følge opp kartlagte handlingspunkter. (Rapportering til Datatilsynet?)

Avsluttende seminar for erfaringsdeling. Beslutning om veien videre.

4. Tidsplan for sandkasseaktivitetene

Vi kan begynne i mai (oppstartmøte), og trolig følge opp med workshop 1 i juni (denne måneden begynner å bli veldig tett). Deretter må vi over sommeren, f.eks. workshop 2 i slutten av august. Siden oppfølgningen av denne antas å være krevende, bør workshop 3 helst legges til oktober. Avsluttende seminar kan legges til slutten av året, eller januar 2024.

Prosjektets formål og problemstillinger kan publiseres eksternt med angivelse av institusjonene som deltar i sandkasseprosjektet.