Etisk rapport: COMPAS

Algoritmen COMPAS kan bidra til effektivisering av rettssystemet. Effektivisering frigjør ressurser, som vil gagne alle parter i samfunnet – spesielt de mest utsatte: de innsatte. Ressurser frigjøres fra rettssaler til rehabilitering slik at ansatte kan flytte fokus fra administrering til menneskeverd. Samtidig, vil denne effektiviseringen føre til dårligere «odds» for de arresterte? Det kan være nyttig å undersøke hvor automatisert dagens praksis er fra før. Om prosessen allerede følger faste regler, så vil kanskje lite, hva angår rettferdighet, endre seg ved å ta i bruk en algoritme – og dermed vil ikke effektiviseringen være problematisk.

Å bruke COMPAS kan også bidra til å avdekke bias og urettferdigheter i rettsvesenet som rettsvesenet som system og institusjon selv ikke er bevisst på. COMPAS kan altså brukes som et analyseverktøy: hva ligger i forskjellen mellom avgjørelsene COMPAS tar og avgjørelsene jurister tar? Dersom COMPAS brukes på denne måten står vi overfor et verktøy som kan skape et mer rettferdig rettssystem. Ut ifra målet om et rettferdig rettssystem følger det uunngåelige spørsmålet: hva er rettferdighet?

Hvordan vil en rettferdig algoritme se ut?

Idealet om en objektiv definisjon på rettferdighet virker å forbli et ideal. Selv om rettferdighet kan assosieres med nøytralitet, vil en definisjon aldri komme fra et nøytralt ståsted. I oppgaven det er å definere rettferdighet, ligger makt og autoritet. Derfor kan vi heller enn å spørre oss for *hva* rettferdighet er, spørre oss *hvem* som burde definere rettferdighet. Burde det være mennesker som allerede har stor autoritet i samfunnet, som jurister og professorer, eller burde det være mennesker som direkte påvirkes av rettferdighetsbegrepet i rettssystemet, være det den fornærmede eller den siktede? Det vil være de som definerer rettferdighetsbegrepet som vil avgjøre hvilke premisser en rettferdig algoritme skal bygges på, og uansett hvem det er som definerer rettferdighetsbegrepet, vil det være deres subjektive begrep – og er det problematisk?

Er det subjektive subjektet rettferdighetens fiende her, i bunn og grunn? Vi er som mennesker unektelig subjektive vesener, og det er vi som konstituerer samfunnet. For at et samfunn skal være rettferdig, synes det å være rimelig å tenke at vesenene som konstituerer det skal kunne fatte rettferdige beslutninger. Om subjektivitet er en fiende for rettferdigheten

betyr det dermed at vi i utgangspunktet har et urettferdig samfunn. I en tilstand hvor kun objektivitet kan fostre rettferdighet, er vi til enhver tid lammet fra å ta ansvar for våre handlinger og fra å dømme dem. Det virker lite plausibelt at det er slik. Dermed må et subjektivt rettferdighetsbegrep kunne eksistere som et gyldig begrep.

Hvis subjektivitet ikke er en fiende for rettferdige beslutninger, er vi i utgangspunktet nødvendigvis ikke i en tilstand av et urettferdig samfunn. Hvis vi legger denne forståelsen til grunn er det mulig for de frie subjektene å ta rettferdige eller urettferdige beslutninger. Da kan vi gå rett til spørsmålet om å ta i bruk algoritmer i samfunnet.

Vi kan spørre: hvor lang tid bør algoritmen trene før den ansees som "god nok"? Spørsmålet hviler på antagelsen om at en algoritme kan perfeksjoneres. Vi står på dørterskelen til en ny æra med et kraftig verktøy. Hvorfor ikke vente til det er optimalt? Det er mulig å vente med å implementere en algoritme til den har trent i 10 år, slik at den verifiseres. Med dette garanterer man algoritmens validitet. Og kanskje er det ikke nødvendig å ha en algoritme som er "god nok" slik trollene "er seg selv nok" i Peer Gynt?

Videre kan man spørre seg hvilket datasett algoritmen trenes med? Med et fullstendig datasett kan kanskje en algoritme kompensere for skjevheter i samfunnet. I COMPAS-algoritmen, hvor dataene handler om menneskers oppførsel, fremstår dataenes holdbarhet som ferskvare. Da bør algoritmen kanskje ideelt sett endre seg over tid, slik menneskers oppførsel endrer seg fra tid til tid. Uten slik kontinuerlig mating av algoritmen er det tilsynelatende vanskelig å sørge for en optimalisert og "rettferdig" algoritme som er dagsaktuell.

En rettferdig algoritme er bevisst. Den ubevisste tar beslutninger ut ifra hvordan ting er, og vil være fokusert på rettferdighet i prosessen. Den bevisste tar beslutninger ut ifra et fokus på rettferdighet i resultatet.

Hvis det finnes eksisterende urettferdigheter i samfunnet, er det mulig for en algoritme å gi rettferdige utslag?

En algoritme burde klare å korrigere eksisterende urettferdighet i et samfunn, men å gi en algoritme ansvar for å rette opp eksisterende urettferdigheter i samfunnet virker som en overfladisk løsning på dype problemer grunnet i kolonialisme, rasisme, klasseskiller og patriarkatet – for å nevne noe. Kort sagt hjelper det ikke å sminke grisen. Det faktiske

problemet, nemlig strukturell urettferdighet i det amerikanske samfunnet, gjenstår uberørt. Dette kan, og burde kritiseres, men samtidig: alle kan ikke stå på barrikadene for å bekjempe den strukturelle urettferdigheten hver eneste dag – selv et strukturelt urettferdig samfunn skal gå rundt, og mennesker som eksisterer i dette samfunnet skal blant annet dømmes til varetekt. Dersom det finnes en mulighet for å bli dømt av en algoritme som retter opp i eksisterende urettferdighet virker det absurd å skulle velge å ikke benytte seg av denne på prinsipielt grunnlag.

Selv om det blir feil å anta at man kan oppnå «rettferdige utslag» med en «rettferdig algoritme», da algoritmen er trent på et urettferdig samfunn og med regler fra et uperfekt rettssystem, kan COMPAS allikevel brukes som et verktøy til endring og forbedring for å gjøre rettssystemet mindre urettferdig.

Hvis resultatet skal være rettferdig og gyldig, må det også bety at det er sant. Det er slik at datasettet er mangelfullt i noen tilfeller, og dette kan ha betydning for sannhet. Dette viser seg angående neglisjerte grupper. Dette er et problem for dem som tenker med et dataistisk tankesett, som vil si at "så lenge man har data på noe, så er det sant". Det impliserer at når man ikke har data på noe, så kan det ikke tas med i sannhetsberegningene – og med det ikke vite om det er sant. Hvis man regner ut ifra sett med manglende data (selv om man sletter noe av dataen for å jevne ut) betyr det at resultatet på sitt beste bare blir tilstrekkelig sant, og ikke gyldig. Eventuelt at man ikke kan avgjøre om noe er sant eller usant. Å likevel holde fast på at man kan lage rettferdige algoritmer med rettferdig resultat med et dataisisk tankesett, er naivt – litt som å skulle insistere på å bygge en asfaltvei fem meter i løse luften.