INSTITUT FOR MENNESKE RETTIGHEDER

NÅR ALGORITMER SAGSBEHANDLER

RETTIGHEDER OG RETSSIKKERHED I OFFENTLIGE MYNDIGHEDERS BRUG AF PROFILERINGSMODELLER

NÅR ALGORITMER SAGSBEHANDLER
Rettigheder og retssikkerhed i offentlige myndigheders brug af profileringsmodeller

Forfattere: Marya Akhtar (kapitel 1-8), Frej Klem Thomsen (kapitel 3 og 7-8), Rikke Frank Jørgensen (kapitel 1 og bidrag til kapitel 4-8), Pernille Boye Koch (kapitel 1 og bidrag til kapitel 2 og 6)

Redaktion: Marya Akhtar og Rikke Frank Jørgensen (Institut for Menneskerettigheder), Trine Jørgensen (Summarium)

En særlig tak til Birgitte Kofod Olsen (Carve) for værdifuld feedback

e-ISBN: 978-87-7570-025-7

Layout: Hedda Bank

Danmarks Nationale Menneskerettighedsinstitution
Wilders Plads 8K
1403 København K
Telefon 32 69 88 88
www.menneskeret.dk

Vi tilstræber, at vores udgivelser bliver så tilgængelige som muligt. Vi bruger f.eks. store typer, korte linjer, få orddelinger, løs bagkant og stærke kontraster.
Læs mere om tilgængelighed på: www.menneskeret.dk/tilgaengelighed

INDHOLD

KAPITEL 1

OM DENNE RAPPORT

”I de kommende år vil kunstig intelligens vende op og ned på, hvordan virksomheder agerer, hvordan den offentlige sektor leverer velfærd til borgerne, og hvordan vi alle lever vores daglige liv. Ligegyldigt om det er løsninger på klimakrisen, forbedret velfærd og sundhed og bedre uddannelse, vil kunstig intelligens være en del af svaret”[1]. Sådan lyder det i en rapport fra Innovationsfonden, der konkluderer, at teknologien potentielt kan gøre Danmark 35 mia. kr. rigere om året.

Innovationsfondens forudsigelser om betydningen af kunstig intelligens deles af mange, og brugen af kunstig intelligens er de seneste år kommet øverst på den politiske dagsorden. Det har allerede nu udmøntet sig i en lang række strategier, signaturprojekter og forskningsprogrammer for kunstig intelligens på både nationalt og EU-plan. Det er værd at bemærke, at den nye teknologi ikke kun forventes brugt af virksomhederne i den private sektor. Der antages også at være et stort potentiale for brugen af kunstig intelligens i den offentlige sektor og i mødet mellem borgerne og myndighederne.

Da de første AI-signaturprojekter i efteråret 2019 blev indstillet til at modtage tilskud for i alt 67 mio. kr., understregede Digitaliseringsstyrelsen, at kunstig intelligens kan bruges i den borgernære velfærd til alt fra hospitaler til borgerservicecenteret. Kunstig intelligens skal bruges, ”hvor det giver værdi for borgerne”, og altid etisk ansvarligt.[2] Samtidig har regeringen i sin strategi for kunstig intelligens fremhævet, at brugen af teknologien ikke må underminere tilliden mellem borger og offentlig myndighed, og at borgere skal være trygge ved offentlige myndigheders anvendelse af kunstig intelligens, særligt når myndigheder anvender den til at understøtte beslutninger og afgørelser. F.eks. skal der sikres en rimelig, ansvarlig og gennemsigtig anvendelse af kunstig intelligens som grundlag for at træffe beslutninger.[3]

Hermed er skitseret nogle af de centrale dilemmaer, som denne rapport skal behandle. For på den ene side er brugen af kunstig intelligens i den offentlige sektor et felt, som er i rivende udvikling, og der er mange ønsker om at bruge teknologien på forskellige velfærds- og forvaltningsområder. På den anden side mangler der et klart overblik over, hvordan teknologianvendelsen griber ind i borgerens rettigheder.

I Danmark, der ofte betegnes som et digitalt foregangsland, har det politiske fokus indtil videre mest været rettet mod alle de fordele, som borgerne og den offentlige sektor kan opnå igennem udbredt brug af kunstig intelligens, f.eks. bedre diagnosticering i sundhedsvæsenet, hurtigere sagsbehandling i kommunen og mere målrettede sociale tilbud til den enkelte. Spørgsmålet om, hvordan brugen af kunstig intelligens udfordrer borgerens rettigheder, er derimod kun mere sporadisk blevet adresseret i Danmark. På internationalt plan har beskyttelsen af borgernes rettigheder haft større fokus, idet både EU[4], Europarådet[5] og FN[6] har udtrykt bekymring over, at konkrete anvendelser af kunstig intelligens potentielt underminerer beskyttelsen af den enkelte borger. Særligt Philip Alston, tidligere FN specialrapportør, har peget på de menneskeretlige udfordringer, som brugen af kunstig intelligens rejser i den offentlige sektor. Han anfører, at myndighederne i varetagelsen af deres opgaver i stigende grad behandler store mængder af data fra en lang række kilder, anvender automatiserede værktøjer til at forudse borgeres adfærd og fjerner sagsbehandlerens menneskelige skøn. Dette fører til en verden, hvor borgere bliver stadig mere gennemsigtige for myndighederne, men myndighederne stadig mere uigennemsigtige for borgerne.[7]

Senest har FN's højkommissær for menneskerettigheder anbefalet, at stater ikke indfører kunstig intelligens, hvor der er uklarhed om teknologiens indvirkning på rettigheder.

ANALYSE

Når der ikke tidligere er foretaget større danske analyser af de rettighedsmæssige perspektiver ved brugen af kunstig intelligens i den offentlige sektor, er det ikke kun, fordi der er tale om et relativt nyt område. En af hovedårsagerne er formentlig også, at området er utrolig komplekst, og at det inkluderer flere fagligheder, der er meget forskellige af natur. Området kræver, at der bygges bro mellem noget meget teknisk, nemlig kunstig intelligens, herunder udvikling og brug af algoritmiske profileringsmodeller, og noget juridisk. Tilmed er også det juridiske felt komplekst og dækker flere forskellige retsområder og reguleringer på både nationalt, EU- og internationalt niveau. Indtil nu har der fra politisk side været meget fokus på etiske principper for kunstig intelligens[8], mens forholdet mellem etisk ansvarlig brug af teknologi og borgernes rettigheder og retssikkerhed ikke har fået samme opmærksomhed.[9]

Emnets faglige karakter kan imidlertid betyde, at der er et ekstra behov for analyser som denne. I praksis er den kunstige intelligens allerede ved at bevæge sig ind i den offentlige sektor, men ofte mangler aktørerne viden om hinandens fagligheder. Der sidder f.eks. dataloger og IT-folk og udvikler algoritmer og profileringsmodeller, men som ikke nødvendigvis har blik for de styringsmæssige logikker i den offentlige forvaltning eller de juridiske principper og rammer. Der sidder også jurister på forskellige niveauer, som savner forståelse for det genstandsfelt, som reglerne skal rumme, fordi det er teknisk vanskeligt at forstå principperne bag udvikling og brug af modeller. Og endelig sidder der politikere, forvaltningschefer og andre beslutningstagere, som er med til at træffe nogle meget afgørende

beslutninger om vores offentlige sektor, men som også kan have svært ved at overskue og gennemskue, hvordan kunstig intelligens kan og bør virke i praksis.

På den baggrund er denne rapport et forsøg på at give et lettilgængeligt overblik over både teknologiens egenskaber, begrænsninger og udvikling, og de rettigheds- og retssikkerhedsmæssige aspekter, som brugen af teknologien rejser i den offentlige sektor. Rapporten behandler problemstillingerne ud fra et tværfagligt perspektiv, hvor vi både kommer rundt om teknologi og jura. Vi bruger derfor i rapporten plads på at redegøre for den teknologi, der ligger bag udvikling og anvendelse af profileringsmodeller. Det tværfaglige perspektiv er efter vores opfattelse nødvendigt i en mere helhedsorienteret analyse af feltet. Der er således også en selvstændig pointe i at fremhæve og illustrere, at myndighedernes kendskab til (og uddannelse i) profileringsmodellers udvikling og anvendelse er en afgørende forudsætning for, at de kan overholde deres forpligtelser og sikre borgernes rettigheder.

Rapporten tager udgangspunkt i borgernes rettigheder og retssikkerhed i forhold til den forvaltningsretlige afgørelse. Det betyder, at vi ikke ser på de mange øvrige områder, hvor myndigheder kan anvende kunstig intelligens i kontakten med borgeren eller i varetagelsen af en myndighedsopgave. Til gengæld ser vi ikke kun på de regler, der gælder for sagsbehandlingen; vi inddrager også det myndighedsarbejde, der går forud for oprettelsen af en sag, f.eks. når en profilerings-model bruges til at vurdere, om en sag overhovedet skal oprettes. Algoritmisk profilering har nemlig, også i den tidligste fase for myndighedernes undersøgelser, en stor betydning for borgerens rettigheder og retssikkerhed. Man kan derfor sige, at rapporten undersøger forholdene "i og omkring" en forvaltningsafgørelse, og når vi, både i rapportens titel og undervejs i vores analyse, henviser til sagsbehandlingen, omfatter det også fasen forud for oprettelsen af en borgersag.

Med denne rapport er det første skridt taget til en tværfaglig afdækning af, hvordan offentlige myndigheders brug af profileringsmodeller griber ind i borgerens rettigheder. Det er en væsentlig pointe i rapporten, at profileringsmodellernes særlige udfordringer for rettigheder og retssikkerhed kræver særlige tiltag rettet mod modellernes brug. Rapporten munder derfor ud i en række anbefalinger til beslutningstagere, som skal sikre, at kunstig intelligens og profileringsmodeller i det offentlige anvendes ud fra en rettighedsbaseret tilgang.

Vi håber, at rapporten kan være med til at give et overblik over nogle af de grund-læggende udfordringer på området, men vi opfordrer også til, at der arbejdes videre med de forhold, som vi beskriver i rapportens enkelte kapitler – og ikke mindst de udfordringer, som af den ene eller anden grund ikke måtte have fundet vej til rapporten.

OPBYGNING

En analyse som den nærværende kan struktureres på mange forskellige måder. Rapportens tværfaglige tilgang betyder, at der ikke er en oplagt systematik for at kombinere det teknologiske felt med det juridiske. Vi har derfor valgt en tematisk tilgang, hvor vi opdeler rapporten i tre dele, der hver behandler forskellige aspekter af, hvad vi anser for væsentligt for, at en rettighedsbaseret brug af profileringsmodeller kan blive sikret i det offentlige.

I rapportens del I præsenterer vi de forhold om rettigheder og teknologi, som vi vurderer er nødvendige at kende til for at sikre en rettighedsbaseret tilgang til brug af profileringsmodeller. Del I består af kapitel 2 og 3.

I kapitel 2 ("Den retlige ramme: Analysens tre retsområder") præsenterer vi de tre retsområder, som vi har valgt at fokusere på. De tre områder er forvaltningsretten, databeskyttelsesretten og forbuddet mod diskrimination.

I kapitel 3 ("Den teknologiske ramme: Kort om profileringsmodeller?") gennemgår vi de tekniske forhold omkring profileringsmodeller, som vi anser for nødvendigt for myndighedspersoner og tilsyn at have indsigt i. Det betyder, at vi i kapitlet – og gennem hele rapporten – bruger tekniske fagtermer og beskriver udviklingsmetoder, som kan virke komplicerede for den uindviede læser, men formentlig vil fremstå forsimplede for specialisten. For at lette læsningen har vi udarbejdet en terminologiliste i bilag 1, hvor samtlige begreber markeret med fed er defineret.

I kapitel 3 præsenterer vi også fire cases, som gennem rapporten vil tjene til at illustrere myndighedernes konkrete anvendelse af profileringsmodeller og de valg og dilemmaer, der knytter sig hertil. De fire cases er desuden mere udførligt beskrevet i bilag 2 til denne rapport.

I rapportens del II præsenterer vi første del af vores analyse, som har til formål at udstikke nogle grundlæggende forudsætninger for brugen af profileringsmodellerne. Del II består af kapitel 4, 5 og 6.

I kapitel 4 ("Beslutning og styring") opstiller vi rammer for myndighedens beslutning om brugen af profileringsmodeller og præsenterer en række konkrete styringsværktøjer, der under hele profileringsmodellens livscyklus skal være med til at sikre borgeres rettigheder og retssikkerhed. Dette kapitel udgør på mange måder rapportens kerne, idet de øvrige tematiske dele er med til at forklare og uddybe, hvorfor sådanne styringsværktøjer er vigtige.

I kapitel 5 ("Modellens input: Nødvendige og korrekte data") fokuserer vi på kravene til de data, som modellen skal udvikles på, og de databeskyttelsesretlige risici, der knytter sig hertil.

I kapitel 6 (”Modellens output: Rammer for anvendelse”) fokuserer vi på selve den profilering af borgeren, som modellen foretager, herunder den måde, hvorpå profileringen dannes, og den måde, hvorpå den anvendes af myndighederne.

I rapportens del III behandler vi to særlige udfordringer ved profileringsmodeller, som ofte og ganske berettiget fremhæves som nogle af de væsentligste. Del III består af kapitel 7 og 8.

I Kapitel 7 (”Transparens i og om profileringsmodeller”) fokuserer vi på de udfordringer, brugen af profileringsmodeller giver for borgernes og offentlighedens mulighed for at få indblik i sagsbehandlingen – og de krav, der derfor må stilles til modellernes transparens.

I kapitel 8 (”Forbuddet mod diskrimination”) fokuserer vi på de udfordringer, brugen af profileringsmodeller rejser for borgernes ret til at blive beskyttet mod både direkte og indirekte diskrimination.

ANBEFALINGER

Rapportens analyser munder ud i en række anbefalinger til det politiske niveau, som skal sikre, at profileringsmodeller i det offentlige anvendes ud fra en rettighedsbaseret tilgang. Anbefalingerne optræder tematisk i kapitlerne 4-7 og er desuden samlet i nedenstående oversigt (se boks). At kapitel 8 om diskrimination ikke indeholder anbefalinger, hænger sammen med, at stort set alle de forudgående anbefalinger i rapporten har til formål også at beskytte mod diskrimination. Rapportens afsluttende kapitel fremhæver således nødvendigheden af alle de forudgående anbefalinger.

VI ANBEFALER

Vejledning (kapitel 4)
at Justitsministeriet med inddragelse af Datatilsynet og Digitaliseringsstyrelsen udsteder en vejledning om offentlige myndigheders brug af profileringsmodeller med fokus på de rettigheds- og retssikkerhedsmæssige udfordringer ved modellerne.

Krav om Artificial Intelligence (AI) konsekvensanalyser (kapitel 4)
at der i vejledningen stilles krav om konsekvensanalyser for kunstig intelligens (AI-konsekvensanalyser) tidligt i beslutningsfasen og periodisk herefter under modellens udvikling og anvendelse til både beslutningsstøtte og fuldautomatisering.

Konsekvensanalyserne bør som minimum omfatte vurdering af:
- Beslutning: Hjemmel og begrundelse for anvendelse af modellen, herunder om modellen anvendes til beslutningsstøtte eller fuldautomatiseret afgørelse, modellens egnethed og oplysninger om eventuelle alternative modeller eller løsninger.
- Risici: Om modellen rejser risici for usaglige og forkerte (ulovlige) afgørelser samt øvrige utilsigtede konsekvenser for borgerens rettigheder, om modellen overholder databeskyttelsesforordningens krav om dataminimering og sikrer brug af korrekte data, hvilke diskriminationsrisici modellen rejser, samt hvorledes samtlige risici vil blive imødegået
- Transparens: Om modellen lever op til krav om algoritmisk transparens.
- Tilsyn og kontrol: Hvordan tilsyn og kontrol med modellen sikres.

Gentræning af data (kapitel 5)
at der i vejledningen (se kapitel 4) stilles krav om periodisk gentræning af modellen på nye data, eller at modellen udvikles således, at den som udgangspunkt giver mindre vægt til data, jo ældre de er.

Partshøring ved brug af profileringsmodeller (kapitel 6)
at der i vejledningen (se kapitel 4) fastsættes krav, der sikrer en reel partshøring ved anvendelsen af profileringsmodeller til fuldautomatisering og til beslutningsstøtte, herunder at vejledningen fastsætter, hvornår og hvordan borgeren inddrages, og hvordan input fra borgeren indgår i sagen.

Automatiseringsbias (kapitel 6)
at vejledningen (se kapitel 4) adresserer risikoen for, at modellen får en uforholdsmæssig stor indflydelse på den endelige afgørelse og bliver beslutningsstyrende frem for beslutningsstøttende.

Tydeliggørelse af begrundelsespligten (kapitel 7)
at det i vejledningen (se kapitel 4) tydeliggøres, at profileringsmodellens konkrete vurdering af borgeren altid indgår i begrundelsespligten, herunder også ved myndigheders brug af profileringsmodeller til beslutningsstøtte.

Hjemmel ved fuldautomatiserede afgørelser (kapitel 4)
at Justitsministeriet tager initiativ til at indføre regler i forvaltningsloven med krav om udtrykkelig lovhjemmel for offentlige myndigheders brug af profilerings-modeller til fuldautomatiserede afgørelser.

Begrænsninger i brug af fuldautomatiserede afgørelser (kapitel 6)
at Justitsministeriet tager initiativ til at indføre regler i forvaltningsloven om, at myndighederne alene må anvende profileringsmodeller til fuldautomatisering, når afgørelsen er baseret på regler med entydige kriterier.

Ændring af databeskyttelsesloven (kapitel 7)
at Justitsministeriet tager initiativ til at ændre databeskyttelsesloven, således at myndigheden skal give besked om, hvilke oplysninger om borgeren der bruges i en profileringsmodel til at støtte eller træffe en afgørelse om borgeren.

Særregel om systemisk transparens i offentlighedsloven (kapitel 7)
at Justitsministeriet tager initiativ til at indføre regler i offentlighedsloven om aktindsigt i informationer om modellens træning, kvalitet, anvendelse og tilsyn, når myndigheder anvender profileringsmodeller til automatiseret beslutningsstøtte eller ved fuldautomatisering.

Offentligt tilgængeligt register over profileringsmodeller (kapitel 7)
at Digitaliseringsstyrelsen, som led i det fællesoffentlige arbejde med kommuner og regioner, opretter et offentligt register over samtlige offentlige myndigheders brug af **profileringsmodeller rettet mod borgere**.

at AI-konsekvensanalyserne eller et uddrag heraf offentliggøres i registret.

Styrket tilsyn (kapitel 4)
at Justitsministeriet med inddragelse af Datatilsynet, Digitaliseringsstyrelsen og Folketingets Ombudsmand sikrer et styrket samarbejde på tværs af de tilsyns-, rekurs- og kontrolinstanser, der forventes at kontrollere myndighedernes brug af profileringsmodeller.

at der fastsættes regler om, at AI-konsekvensanalyserne indrapporteres til en relevant myndighed, herunder at valget af myndighed sker under hensyntagen til EU's kommende forordning om kunstig intelligens.

Styrket teknologisk og rettighedsmæssig forståelse (kapitel 4)
at behovet for teknisk forståelse for profileringsmodeller og for de rettigheds- og retssikkerhedsmæssige udfordringer, som de rejser hos beslutningstagere og anvendere af profileringsmodeller samt hos tilsyns- og kontrolinstanser adresseres i vejledningen.

at Digitaliseringsstyrelsen inkluderer teknisk kendskab til profileringsmodeller og kendskab til de retlige udfordringer forbundet med profileringsmodeller i "Statens Digitaliseringsakademi".

NOTER

1 Duvold, Innovationsfonden (2019), Ambitiøs dansk satsning på kunstig intelligens kan gøre vores samfund rigere og vores liv bedre, tilgængeligt på: https://innovationsfonden.dk/da/nyheder-presse-og-job/ambitios-dansk-satsning-pa-kunstig-intelligens-kan-gore-vores-samfund-rigere
2 Digitalstyrelsen (2019), Kommuner og regioner skal afprøve kunstig intelligens for at løfte kvaliteten i den offentlige service, tilgængelig på: https://digst.dk/nyheder/nyhedsarkiv/2019/oktober/kommuner-og-regioner-skal-afproeve-kunstig-intelligens-for-at-loefte-kvaliteten-i-den-offentlige-service/
3 Regeringens nationale strategi for kunstig intelligens (2019), initiativ 51 tilgængelig på: https://digst.dk/media/19302/national_strategi_for_kunstig_intelligens_final.pdf
4 Se f.eks. Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final, der er en del af en større EU-indsats for at udvikle kunstig intelligens med mennesket i centrum
5 Committee of experts on internet intermediaries (MSI-NET) (2019), Algorithms and Human Rights, Study on the human rights dimensions of automated data processing techniques and possible regulatory implications, (CoE), tilgængelig på: https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5
6 Se rapport fra FN Specialrapportør for ekstrem fattigdom (2019) tilgængelig på: https://www.ohchr.org/EN/Issues/Poverty/Pages/DigitalTechnology.aspx Se også rapport fra FN Specialrapportør for racediskrimination mv (2020) tilgængelig på: https://undocs.org/en/A/HRC/44/57
7 Se rapport fra FN Specialrapportør for ekstrem fattigdom (2019) tilgængelig på: https://www.ohchr.org/EN/Issues/Poverty/Pages/DigitalTechnology.aspx
8 EU-kommissionens uafhængige Ekspertgruppe på Højt Niveau om Kunstig Intelligens (2019) Etiske retningslinjer for pålidelig kunstig intelligens tilgængelig på: https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/JURI/DV/2019/11-06/Ethics-guidelines-AI_DA.pdf
9 Wagner (2018) Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping? In M. Hildebrandt (Ed.), Being Profiling. Cogitas ergo sum. Amsterdam University Press.

DEL I

DEN RETLIGE OG TEKNOLOGISKE RAMME

I denne del af rapporten præsenterer vi de tre retsområder, som de efterfølgende analyser tager udgangspunkt i. Samtidig giver vi en indføring i teknologien, dens udviklingsmetoder og særlige kendetegn.

Fra et juridisk perspektiv er det ikke så ligetil at analysere profileringsmodeller. Udfordringerne med profileringsmodellerne går på tværs af klassiske juridiske discipliner, der på forskellig vis har til formål at sikre rettigheder og retssikkerhed i forvaltningen. Med enkelte undtagelser er de forskellige regelsæt ikke udformet med teknologien i tankerne, og der kan være uklarhed både om deres anvendelse og samspil, når profileringsmodeller introduceres i sagsbehandlingen. I kapitel 2 opridser vi den retlige ramme for analysen.

Herefter vender vi os mod den teknologiske redegørelse i kapitel 3, som tjener til både at forklare teknologien og til at præsentere, hvad vi anser for uomgængelig teknologisk forståelse, som det er nødvendigt, at beslutningstagere, retsanvendere og tilsynsmyndigheder besidder. Ligesom den retlige ramme er også teknologien ganske kompleks og kan være svær at omsætte til sprog og koncepter, som er mere alment forståelige. I kapitel 3 forsøger vi at bibringe ikke-teknikkyndige læsere en forståelse for området.

KAPITEL 2

DEN RETLIGE RAMME: ANALYSENS TRE RETSOMRÅDER

I dette kapitel redegør vi for de rettigheds- og retssikkerhedsmæssige forhold, som vores analyse bygger på. Det har været nødvendigt at træffe en række valg om denne retlige ramme, og disse valg redegør vi for samtidig med, at vi giver et overblik over de regler, som vi fokuserer på.

Det er relativt kompleks at foretage en juridisk analyse af offentlige myndigheders brug af profileringsmodeller. De relevante regelsæt har forskellig karakter og status og går ofte på tværs af klassiske juridiske discipliner. Der er både nationale og internationale regler, og de findes både inden for traditionel forvaltningsret, EU-databeskyttelsesret og menneskeretlig diskriminationslovgivning. Det betyder, at reglerne har forskellige formål, beskyttelsesinteresse og reguleringslogik, og det påvirker den måde, de anvendes og fortolkes på.

Vi har i denne rapport valgt et fokus på rettigheder og retssikkerhed ud fra både et individperspektiv og et strukturelt styringsperspektiv og med et mere tværdisciplinært udgangspunkt end hvad der er sædvanligt ud fra en snæver disciplinbetragtning. Det betyder forhåbentlig, at vores juridiske analyse bliver mere dækkende, uden at den dog kan betegnes som komplet.

DE TRE RETSOMRÅDER

Når vi i denne rapport beskæftiger os med rettigheder og retssikkerhed, har vi særligt udvalgt tre retsområder. For det første ser vi på rettigheder og retssikkerhed, som sikres i forvaltningsretten, idet genstandsfeltet for rapporten er offentlige myndigheders brug af profileringsmodeller. Vi vil nedenfor begrunde valget mere konkret og forklare, hvordan vi griber det an.

For det andet ser vi på den særegne beskyttelse af personoplysninger, der følger af EU's databeskyttelsesforordning (samt databeskyttelsesloven), hvorimod vi ikke kommer nærmere ind på den beskyttelse, der følger af internationale konventioner. Den nærmere forklaring på denne prioritering vil blive udfoldet nedenfor.

For det tredje fokuserer vi på diskriminationsforbuddet, som nok i højere grad kan kvalificeres som "klassisk" menneskeret. Vi giver et kort overblik over, hvilke regler, der på internationalt, EU og nationalt plan beskytter mod diskrimination. Derudover forklarer vi, hvorfor vi i kapitel 8 om forbuddet mod diskrimination har valgt at fokusere på den EU-retlige beskyttelse mod diskrimination snarere end de internationale konventioner.

Især forvaltningsretten og databeskyttelsesretten er for myndighederne tæt beslægtede områder, hvor databeskyttelsesretten både præciserer og supplerer krav til sagsbehandlingen efter forvaltningsretten. Hovedparten af vores rapport centrerer om disse to retsområder.

Diskriminationsforbuddet rejser særlige problemstillinger og udgør et særligt risikoområde, som behandles i rapportens sidste kapitel. Forbuddet mod diskrimination følger også delvist af og supplerer forvaltningsrettens grundprincip om lighed, ligesom der er en nær sammenhæng mellem beskyttelsen af personoplysninger og forbuddet mod diskrimination. Beskyttelseshensynet indgår derfor også løbende i rapportens analyser og anbefalinger.

Formålet med rapporten er at give et overblik over de væsentligste udfordringer inden for de tre retsområder fra et rettighedsperspektiv, og vi holder os derfor til de overordnede principper og regler uden at gå ned i særlovgivning, som måtte gælde for myndigheder på bestemte områder. Vores analyse omfatter forpligtelser, som samtlige danske myndigheder (statslige, regionale og kommunale) som minimum skal iagttage og vi kommer med en række anbefalinger, som vedrører dem alle.

Det er kendetegnende for retsområderne, at ingen af reglerne (med en enkelt undtagelse) er udformet netop med sigte på profileringsmodeller. Snarere er der tale om mere generelle principper fra de forskellige retsområder, som skal anvendes på det relativt nye fænomen, at kunstig intelligens i stadig stigende omfang bruges i den offentlige sektor over for borgerne.

Behovet for strukturelle regler om brugen af kunstig intelligens har den seneste tid ført til reguleringsmæssige tiltag på europæisk plan, som vi behandler til slut i kapitlet og undervejs i rapporten løbende anvender til perspektivering. Det er imidlertid vigtigt at fremhæve, at udfordringerne, som behandles, nødvendiggør et nationalt fokus på området – et fokus, som vi håber, at denne rapport kan bidrage med.

FORVALTNINGSRETTEN OG MENNESKERETTEN

Man kan spørge sig selv, hvorfor Institut for Menneskerettigheder beskæftiger sig med forvaltningsretten. Svaret er, at der er væsentlige sammenfald mellem forvaltningsretten og menneskeretten. Forvaltningsretten baserer sig på et retssikkerheds- og et retsstatsprincip. Det handler grundlæggende om at beskytte borgerne i forhold til myndighedernes ageren og dermed om borgernes retssikkerhed i mødet med myndighederne – hvilket også er formålet med beskyttelsen efter de internationale koventioner. Begge retsområder skal sikre, at myndighederne forvalter i overensstemmelse med lovgivningen, forfølger saglige hensyn og ikke træffer afgørelser på et vilkårligt grundlag.[1]

Umiddelbart indeholder menneskeretten relativt få nedskrevne og håndfaste krav til offentlige myndigheders sagsbehandling, dvs. en slags egentlig, menneske-retligt funderet forvaltningsret. Imidlertid er det vigtigt at være opmærksom på, at det fremgår af Den Europæiske Menneskerettighedsdomstols praksis, at

Domstolen i konkrete situationer – uafhængigt af de nationale regler – lægger vægt på, om processen ved de nationale myndigheder har ført til en passende involvering af borgeren i beslutningsprocessen og har givet denne en rimelig mulighed for at varetage sine interesser.[2] Domstolen har i sin praksis særligt fokuseret på procedurer i forhold til myndighedernes afgørelser, sagsoplysning, kontradiktion og den samlede prøvelse.[3]

Det bør også understreges, at Den Europæiske Menneskerettighedskonvention har det som udtrykkelig grundforudsætning, at indgreb i de relevante rettigheder kræver et betryggende retsgrundlag for at overholde menneskeretten. Det er et krav, som på mange måder ligner det forvaltningsretlige legalitetsprincip. Det forvaltningsretlige legalitetsprincip er centralt for retsstaten og indebærer, at forvaltningen er underlagt lovgivningsmagten og derfor ikke må handle i strid med lovene (det formelle lovs princip) og ikke må handle uden hjemmel i lovgivningen (hjemmelskravet).[4] Dette betyder for det første, at forvaltningen skal handle lovligt og for det andet, at den som udgangspunkt skal kunne henvise til en retsregel som grundlag for sin afgørelse.

Endelig skal det bemærkes, at forvaltningsretten både vedrører det grundlæggende retsgrundlag, som myndighederne opererer ud fra, og også skal bidrage til en mere systemisk sikring af rettigheder – ikke kun på individniveau, men på et overordnet plan. Forvaltningsrettens essens er derfor at sikre retssikkerheden, ligesom tilfældet er med den internationale menneskeret.

Retssikkerhed er ikke et entydigt begreb, men knytter sig til forestillingen om retsstaten, der bygger på "rule of law". Heraf kan udledes principper om f.eks. lighed for loven, forudsigelighed og fraværet af arbitrære afgørelser, klare og gennemskuelige regler, effektive kontrolinstanser (herunder domstolene). Mange af de forvaltningsretlige grundprincipper om f.eks. saglighed, lighed og proportionalitet, som myndighederne skal overholde i deres afgørelser, har således et betydeligt overlap med retssikkerheds- og retsstatsprincipper, der bl.a. følger af forfatningsretten og den internationale menneskeret.[5] I forvaltningsretten bruges retssikkerhed som en samlebetegnelse for en lang række materielle og formelle krav, der har det til fælles, at de har til hensigt at sikre borgerne en lovlig og fair behandling og så vidt muligt kompensere for det asymmetriske magtforhold, der eksisterer i mødet mellem den enkelte borger og det store magtapparat.

MATERIELLE OG FORMELLE KRAV TIL DIGITAL SAGSBEHANDLING

Der forhold, at sagsbehandlingen digitaliseres, har principielt set ikke betydning for de forvaltningsretlige forpligtelser. Forvaltningsretten er teknologineutral og samtlige forpligtelser gælder, uanset hvordan sagsbehandlingen gennemføres.[6]

Der kan imidlertid opstå en række forskellige udfordringer ved digitaliseret sagsbehandling, og ofte er myndighederne ikke tilstrækkeligt opmærksomme på de retssikkerhedsmæssige garantier i den digitale sagsbehandling, hvilket

Folketingets Ombudsmand ved flere lejligheder har påpeget som et alvorligt problem (se boks).

OMBUDSMANDEN: FORVALTNINGSRETLIGE KRAV TIL OFFENTLIGE IT-LØSNINGER

Som respons på flere sager om IT-løsninger i det offentlige har ombudsmanden bl.a. observeret, at:

"Myndigheder er ikke opmærksomme på, at en hel eller delvis automatisering af sagsbehandlingen på et område som udgangspunkt forudsætter, at afgørelserne eller de automatiserede dele af afgørelserne kan træffes efter rent objektive kriterier, og at myndighederne på forhånd kan afgrænse, hvilket faktum der vil være relevant i fremtidige sager."

"Myndigheder er ikke opmærksomme på retten til partshøring, hvor en IT-løsning indebærer, at der som led i behandlingen af en afgørelsessag automatisk indhentes oplysninger fra registre og tidligere sager."[7]

"Myndigheder forsømmer at sikre, at den relevante juridiske ekspertise er til rådighed i alle væsentlige faser af udviklingen af nye IT-systemer, bl.a. med henblik på at sikre, at formelle og materielle regler for behandlingen af de pågældende sager overholdes eller søges ændret."[8]

Både materielle og formelle forvaltningsretlige krav kan blive udfordret af, at profileringsmodeller introduceres i sagsbehandlingen.

De materielle forvaltningsretlige krav indebærer bl.a., at forvaltningen skal overholde det førnævnte legalitetsprincip, og skal foretage et skøn, hvor det er forskrevet. Herudover er de væsentligste materielle krav til afgørelsen, at den skal være saglig, proportional og i overensstemmelse med lighedsgrundsætningen.

Det materielle krav om saglighed er tæt knyttet til legalitetsprincippet og indebærer, at forvaltningen i sine afgørelser ikke må inddrage hensyn, som lovgiver ikke (direkte eller indirekte) har anerkendt som saglige og – efter omstændighederne – at forvaltningen er forpligtet til at inddrage visse hensyn, når de er relevante i en konkret sammenhæng.[9] Saglighedskravet følger hovedsageligt af ombudsmandspraksis og kaldes også forbuddet mod magtfordrejning. Hvilke hensyn, der er saglige, beror på loven. I nogle bestemmelser vil bestemte hensyn (som f.eks. alder) være saglige, mens de samme hensyn i andre bestemmelser vil være usaglige. Andre hensyn vil altid anses for usaglige (fx religiøs, politisk og seksuel overbevisning). Vi kommer nærmere ind på de materielle krav i kapitel 6.

Ved siden af de materielle krav til myndighederne, gælder der også en række formelle krav, som myndighederne skal iagttage under sagsbehandlingen. De fleste af disse krav tager form af garantiforskrifter, der betyder, at en afgørelse kan blive ugyldig, hvis ikke reglerne er iagttaget. De formelle krav kan have forskellige formål, men garantiforskrifterne sigter i sidste ende alle sammen mod, at myndigheden træffer en lovlig og korrekt afgørelse. De væsentligste sagsbehandlingsregler, som vi vil komme ind på i denne analyse, er reglerne om partshøring, officialprincippet, myndighedens begrundelsespligt, og reglerne om aktindsigt.

Udover ombudsmandsudtalelserne (se boks) er der udstedt en række vejledninger mv. fra myndighederne om offentlige digitale løsninger, herunder Digitaliseringsstyrelsens vejledning om digitaliseringsklar lovgivning samt Justitsministeriets notat om forvaltningsretlige krav til offentlige digitale løsninger.[10]

Vejledningen om digitaliseringsklar lovgivning forholder sig ikke eksplicit til rettighedsperspektivet, men bygger videre på en politisk aftale indgået mellem Folketingets partier i 2018, der lægger kimen til udviklingen af den digitale forvaltning.[11] Aftalen indebærer, at lovforslag skal stile efter og vurderes i forhold til om de er "digitaliseringsklar". Lovgivning anses for digitaliseringsklar, hvis den lever op til syv grundprincipper opstillet i aftalen: enkle og klare regler; digital kommunikation; muliggørelse af automatisk sagsbehandling; sammenhæng på tværs – ensartede begreber og genbrug af data; tryg og sikker datahåndtering; anvendelse af offentlig infrastruktur og endelig forebyggelse af snyd og fejl.

Det fremgår af aftalen, at digitaliseringen skal bidrage til større gennemsigtighed, bedre tilgængelighed for borgere og virksomheder og en mere ensartet sagsbehandling. Det er imidlertid også klart, at man med aftalen tilstræber en højere grad af digitalisering og øget brug af automatiseret sagsbehandling. Hermed øges også behovet for at vurdere de rettigheds- og retssikkerhedsmæssige konsekvenser af digitaliseringen, hvilket for så vidt erkendes i aftalen, som bemærker, at man skal sikre, at digitaliseringen respekterer borgerens rettigheder.

DATABESKYTTELSE

Databeskyttelse – eller beskyttelse af personoplysninger – er det andet overordnede retsområde, vi beskæftiger os med i denne rapport. Det følger af forskellige regelsæt og er tæt knyttet til den menneskeretlige beskyttelse af retten til respekt for privatliv. Retten til respekt for privatliv og beskyttelsen af personoplysninger sætter grænser for, hvor meget og hvordan staten må blande sig i og få indblik i individets forhold.

En stor del af databeskyttelsen er omfattet af reglerne om respekt for privatlivet, herunder FN's Konvention om Borgerlige og Politiske Rettigheder og Den Europæiske Menneskerettighedskonvention, men personoplysninger er

samtidig særligt beskyttet i en række regelsæt. Det gælder ikke mindst Den Europæiske Unions Charter om Grundlæggende Rettigheder (EU-chartret), hvor personoplysninger er beskyttet i en særskilt bestemmelse i artikel 8. Her fremgår det bl.a., at personoplysninger skal behandles rimeligt, til udtrykkeligt angivne formål og på grundlag af de berørte personers samtykke eller på et andet berettiget grundlag fastsat ved lov. Også artikel 16 i Traktaten om den Europæiske Unions Funktionsmåde omhandler databeskyttelse.

Ved siden af disse regler gælder EU-databeskyttelsesforordningen (GDPR), som er det regelsæt, som vi behandler i rapporten. Vores fokus skyldes, at forordningens regulering er den mest specifikke og udførlige i forhold til rapportens genstandsfelt. Forordningen indeholder bl.a. en række grundprincipper for behandling af personoplysninger (se boks) og krav til tekniske, organisatoriske og sikkerhedsmæssige foranstaltninger, som vi især fokuserer på i rapporten.

De rettigheder og forpligtelser, der fremgår af databeskyttelsesforordningen, udgør en central del af EU-rettens menneskeretlige beskyttelse af personoplysninger[12] og opstiller principper, der sikrer en systematisk beskyttelse eller etablerer hvad man kan kalde en governance- og compliance-struktur i form af bl.a. et ansvarlighedsprincip. Forordningen stiller krav om gennemførelsen af såkaldte konsekvensanalyser og indlejring af databeskyttelse direkte i værktøjernes design. Som noget særligt regulerer forordningen desuden eksplicit visse former for automatiserede afgørelser.

Af samme grund spiller forordningen en central rolle i vores rapport, og en stor del af vores anbefalinger bygger videre på de regler, der allerede gælder efter forordningen. Navnlig forordningens grundprincipper har stor betydning. Forordningen opstiller desuden regler, der skal sikre overholdelsen af alle rettigheder beskyttet af EU-retten og ikke kun beskyttelsen af personoplysninger. Forordningens systemiske beskyttelse forventes også at få indflydelse på kommende europæisk regulering af kunstig intelligens (se nedenfor).

Databeskyttelsesforordningen beskytter "personoplysninger", bredt defineret som alle former for oplysninger, der relaterer sig til en identificeret eller identificerbar fysisk person. Det betyder, at også oplysninger, som skal bearbejdes, før de kan henføres til en bestemt fysisk person (såkaldt personhenførbare oplysninger), er omfattet af forordningen.[13] Med til begrebet hører både oplysninger om en person, oplysninger der bruges i relation til en person, og oplysninger der vil have en indvirkning på en person.[14] Personoplysninger skal beskyttes både, når de behandles som led i udviklingen af en algoritmisk profileringsmodel (til at træne modellen) og når modellen behandler oplysninger om en borger (f.eks. ved registersamkøring) for at træffe eller støtte en afgørelse rettet mod borgeren.

EU'S DATABESKYTTELSESFORORDNING: GRUNDPRINCIPPER FOR BEHANDLING AF PERSONOPLYSNINGER

Grundprincipperne følger af forordningens artikel 5 og indbefatter:

- **Lovlighed, rimelighed og gennemsigtighed** i forbindelse med behandling.
- **Formålsbestemthed** – Indsamling til udtrykkeligt angivne og legitime formål - senere behandling må ikke være uforenelig hermed
- **Dataminimering** – Oplysningerne skal være tilstrækkelige og relevante og ikke omfatte mere end nødvendigt (særlig proportionalitetsregel).
- **Rigtighed** – Oplysninger skal være korrekte og ajourførte og urigetige oplysninger skal slettes eller berigtiges.
- **Opbevaringsbegrænsning** – Oplysninger skal slettes eller anonymiseres når de ikke længere er nødvendige.
- **Integritet og fortrolighed** – Beskyttelse mod uautoriseret eller ulovlig behandling.
- **Ansvarlighed** - Ansvar i forhold til at kunne påvise, at de øvrige principper opfyldes.

Endelig regulerer databeskyttelsesforordningens artikel 22 udtrykkeligt afgørelser og profilering (se boks). Artikel 22 finder kun anvendelse, hvis en afgørelse er truffet fuldautomatiseret. Er afgørelsen ikke fuldautomatiseret gælder i stedet de almindelige regler i forordningen. Sondringen mellem fuldautomatiserede afgørelser og automatiseret beslutningsstøtte drejer sig om, hvordan en myndighed helt konkret bruger de resultater og vurderinger, som en algoritmisk profileringsmodel producerer. Det er en central, men ikke entydig sondring, som vil blive udfoldet i kapitel 6.

EU'S DATABESKYTTELSESFORORDNING: ARTIKEL 22 OM AUTOMATISEREDE AFGØRELSER OG PROFILERING

"Den registrerede har ret til ikke at være genstand for en afgørelse, der alene er baseret på automatisk behandling, herunder profilering, som har retsvirkning eller på tilsvarende vis betydeligt påvirker den pågældende.

2. Stk. 1 finder ikke anvendelse, hvis afgørelsen:
a) er nødvendig for indgåelse eller opfyldelse af en kontrakt mellem den registrerede og en dataansvarlig
b) er hjemlet i EU-ret eller medlemsstaternes nationale ret, som den dataansvarlige er underlagt, og som også fastsætter passende foranstaltninger til beskyttelse af den registreredes rettigheder og frihedsrettigheder samt legitime interesser eller
c) er baseret på den registreredes udtrykkelige samtykke.

3. I de tilfælde, der er omhandlet i stk. 2, litra a) og c), gennemfører den dataansvarlige passende foranstaltninger til at beskytte den registreredes rettigheder og frihedsrettigheder samt legitime interesser, i det mindste den registreredes ret til menneskelig indgriben fra den dataansvarliges side, til at fremkomme med sine synspunkter og til at bestride afgørelsen.

4. De afgørelser, der er omhandlet i stk. 2, må ikke baseres på særlige kategorier af personoplysninger, jf. artikel 9, stk. 1, medmindre artikel 9, stk. 2, litra a) eller g), finder anvendelse, og der er indført passende foranstaltninger til beskyttelse af den registreredes rettigheder og frihedsrettigheder samt legitime interesser."

Databeskyttelsesforordningen gælder både for private og offentlige aktører, og derfor bruger forordningen begreber som "databehandler", "dataansvarlig" og "den registrerede". I det følgende bruger vi imidlertid begreberne "myndigheden" og "borgeren" for at tydeliggøre den kontekst, som vi bruger reglerne i.

Databeskyttelsesforordningen bygger på regler og principper, som henover flere år er blevet udviklet i EU-regler og -domstolspraksis. Et af de organer, som har bidraget til fortolkning af og vejledning om regler og praksis er den såkaldte Artikel 29-gruppe, som frem til 2018 var en uafhængige europæiske arbejdsgruppe, der beskæftigede sig med databeskyttelsesretten. Gruppen blev erstattet af det Europæiske Databeskyttelsesråd (Databeskyttelsesrådet) i 2018, hvor databeskyttelsesforordningen trådte i kraft. Rådet er ligeledes et uafhængigt europæisk organ og har godkendt en række af Artikel 29-gruppens vejledninger mv. Disse vejledninger mv. udgør et vigtigt fortolkningsbidrag for databeskyttelsesforordningen og vi henviser derfor løbende til dem i vores analyse.[15]

DISKRIMINATIONSFORBUDDET

Det tredje retsområde, vi beskæftiger os med, er forbuddet mod diskrimination. Det står centralt i menneskeretten og følger af en lang række internationale konventioner, EU-retten og dansk ret (se boks).

Vi har i rapporten taget udgangspunkt i den EU-retlige beskyttelse mod diskrimination snarere end i de internationale konventioner. Dette valg har vi taget fordi domstolspraksis og direktiver i EU-retten er illustrative for netop de problemstillinger, som vi beskæftiger os med i rapporten. EU-retten tjener derfor til at udfolde og forklare nogle af de mere generelle udfordringer, som profileringsmodellerne rejser i forhold til diskriminationsforbuddet selvom EU-retten kun finder anvendelse på en afgrænset del af det offentliges mange sagsområder.

I takt med, at offentlige myndigheder tager profileringsmodeller i brug, opstår der nye risici for diskrimination. Det sker, fordi en allerede diskriminerende praksis risikerer at blive videreført eller forstærket i modellen, og fordi modellen kan have svært ved at beskytte mod alle former for diskrimination på samme tid. Beskyttelse mod diskrimination er derfor omfattet af Del III om særlige udfordringer, og vi bruger kapitel 8 på at gå i dybden med emnet.

DISKRIMINATIONSFORBUDDET

Der gælder et generelt forbud mod diskrimination i FN's Konvention om Borgerlige og Politiske rettigheder og FN's Konvention om Økonomiske og Sociale Rettigheder. FN's tre konventioner, Racediskriminationskonventionen, Kvindekonventionen og Handicapkonventionen indeholder alle særlige forbud mod diskrimination.

Den Europæiske Menneskerettighedskonventions artikel 14 indeholder et såkaldt accessorisk diskriminationsforbud, hvorefter de rettigheder, konventionen beskytter, skal sikres uden forskel på grund af bl.a. race, farve, sprog, religion, national eller social oprindelse eller andre forhold. Bestemmelsen er som nævnt accessorisk og finder kun anvendelse i sammenhæng med en rettighed beskyttet i konventionen eller i én af tillægsprotokollerne.

EU-retten indeholder i traktaterne og EU's Charter om Grundlæggende Rettigheder et ligebehandlingsprincip og et diskriminationsforbud. Ved siden af de generelle regler og principper er der vedtaget en række direktiver, der er implementeret i dansk ret, der beskytter imod diskrimination i særlige situationer. Direktiverne beskytter imod diskrimination på baggrund af race og etnisk oprindelse, religion eller tro, handicap, alder og seksuel orientering og køn.

Diskriminationsforbuddet i de internationale konventioner og EU-retten suppleres af den forvaltningsretlige lighedsgrundsætning, som indebærer, at offentlige myndigheder er forpligtede til at anvende en ensartet fortolkning af loven og til at behandle lige forhold lige, når de træffer skønsmæssige afgørelser.[16]

KOMMENDE EUROPÆISK REGULERING AF KUNSTIG INTELLIGENS

I det foregående har vi beskrevet de tre retsområder, som vores rapport bygger på. Her til slut skitserer vi kort de planer, der er på europæisk plan om kommende reguleringer af brugen af kunstig intelligens. Her forventes nogle af de mest omfattende strukturelle krav at blive fastsat i EU's kommende forordning om kunstig intelligens. EU-Kommissionen har fremsat et udkast til forordningen i foråret 2021, som i skrivende stund er under behandling i Parlamentet og Rådet.

Kommissionens udkast er udarbejdet ud fra en risikotilgang, hvor reguleringen afhænger af, hvor stor risiko der er ved anvendelsen af kunstig intelligens for menneskers sundhed, sikkerhed og grundlæggende rettigheder. Udkastet har et særligt fokus på højrisiko AI-systemer, som bl.a. omfatter offentlige myndigheders brug af profileringsmodeller når de er: "beregnet til at blive anvendt af offentlige myndigheder eller på vegne af offentlige myndigheder til at vurdere fysiske personers berettigelse til offentlige sociale ydelser og tjenester samt til at tildele, reducere, annullere eller tilbagekalde sådanne ydelser og tjenester".[17]

Såfremt et AI-system er i højrisikogruppen indebærer det særlige krav til bl.a. datakvalitet, teknisk dokumentation og registrering, gennemsigtighed og formidling af oplysninger, menneskeligt tilsyn og krav til systemets robusthed, nøjagtighed og cybersikkerhed.

Hovedformålet med EU-Kommissionens udkast er at forbedre det indre marked ved at fastsætte ensartede regler for især markedsføring, salg og brug af kunstig intelligens inden for EU. Beskyttelsen af rettigheder er en integreret del af udspillet, omend det ikke er hovedsigtet.[18]

Vi vil undervejs i rapporten inddrage dette udkast for at illustrere forskellige problematikker og de løsninger, der er under overvejelse i EU. Udkastet er selvsagt ikke endeligt og tjener i vores rapport blot til at fremhæve udvalgte pointer i rapportens forskellige tematikker.

I regi af Europarådet er der overvejelser om at gennemføre et regelsæt for regulering af kunstig intelligens funderet i menneskeret, demokrati og retsstatsprincippet og rådet har oprettet en komite for kunstig intelligens (CAHAI). CAHAI har i sit arbejde med at udvikle et regelsæt for kunstig intelligens bl.a. fokuseret på offentlige myndigheders brug af kunstig intelligens og tager i arbejdet udgangspunkt i gennemførelsen af konsekvensanalyser, og anlægger ligesom i EU-udspillet en risikobaseret tilgang.[19]

NOTER

1 Se Institut for Menneskerettigheder (2017), Retssikkerhed i kommunerne s. 18, tilgængelig på: https://menneskeret.dk/sites/menneskeret.dk/files/media/dokumenter/udgivelser/forskning_2017/rapport_om_retssikkerhed_i_kommunerne_15dec2017.pdf.

2 Se bl.a. EMD, T.P. og K.M. mod Storbritannien, dom af 10. maj 2001, app. nr.: 28945/95, pr. 72

3 Christoffersen (2009), Fair balance: Proportionality, Subsidiarity and Primarity in the European Convention on Human Rights, Martinus Nijhoff Publishers, s. 521

4 Fenger (2018), Forvaltningsret, Jurist- og Økonomforbundets Forlag, s. 299

5 Se f.eks. Krunke, (2020). Forfatningens Principper. In I. Nguyen Duy, S. Bragdø-Ellenes, I. Lorange Backer, S. Eng , & B. E. Rasch (Eds.), Uten sammenligning – festskrift til Eivind Smith på 70-årsdagen (pp. 333-348). Fagbokforlaget og Motzfeldt, (2020) Machine Learning og forvaltningens skønsudøvelse. Juristen, 102(4), 140-147

6 Se ombudsmandens notat (2. juli 2018), forvaltningsretlige krav til det offentliges IT-løsninger, tilgængeligt på: https://www.ombudsmanden.dk/myndighedsguiden/specifikke_sagsomraader/generelle_forvaltningsretlige_krav_til_offentlige_it-systemer/2019/

7 Ombudsmanden (2. juli 2018), Ombudsmandens notat, Forvaltningsretlige krav til det offentliges IT-løsninger tilgængeligt på: https://www.ombudsmanden.dk/myndighedsguiden/specifikke_sagsomraader/generelle_forvaltningsretlige_krav_til_offentlige_it-systemer/2019/

8 Ombudsmanden (24. juni 2020), Ombudsmandens Myndighedsguide, overblik # 13 tilgængeligt på: https://www.ombudsmanden.dk/myndighedsguiden/specifikke_sagsomraader/generelle_forvaltningsretlige_krav_til_offentlige_it-systemer/

9 Motzfeldt, (2020) Machine Learning og forvaltningens skønsudøvelse. Juristen, 102(4), 140-147

10 Digitaliseringsstyrelsens vejledning om digitaliseringsklar lovgivning (maj 2018) tilgængeligt på: https://digst.dk/media/16953/vejledning_om_digitaliseringsklar_-lovgivning_maj_2018_tg.pdf og Justitsministeriets notat af 2015 om forvaltningsretlige krav til offentlige digitale løsninger tilgængeligt på: https://digst.dk/media/12721/notat-om-forvaltningsretlige-krav-til-det-offentliges-itlsninger.pdf

11 Regeringen (2018), Aftale om digitaliseringsklar lovgivning, tilgængelig på: https://www.regeringen.dk/media/4690/digitaliseringsklar-lovgivning.pdf

12 Docksey (januar 2020), The EU approach to the protection of rights in the digital environment: today and tomorrow – State obligations and responsibilities of private parties – GDPR rules on data protection, and what to expect from the upcoming ePrivacy regulation, Human Rights Challenges in the Digital Age: Judicial Perspectives (CoE)
13 Databeskyttelsesforordningen 2016/679 af 27. april 2016, artikel 4, nr. 1
14 Artikel 29-Gruppens retningslinjer om automatiske individuelle afgørelser og profilering i henhold til forordning 2016/679 (2017/2018) s. 12
15 Et overblik over vejledninger mv. godkendt af Det Europæiske Databeskyttelsesråd er tilgængeligt på Datatilsynet hjemmeside: https://www.datatilsynet.dk/presse-og-nyheder/nyhedsarkiv/2018/maj/edpb-godkender-tidligere-udstedte-vejledninger
16 Fenger, N. (2018), Forvaltningsret, Jurist- og Økonomforbundets Forlag, s. 347
17 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final, bilag III, Højrisiko-AI-systemer, jf. artikel 6, stk. 2, punkt 5 (a). Se også uddybning i udkastet betragtning 37
18 For en kritisk gennemgang af kommissionens udkast se Smuha, Ahmed-Rengers, Harkens, Li, MacLaren, Piselli & Yeung (august 2021), How the EU can achieve legally trustworthy AI: A response to the European Commission's proposal for an Artificial Intelligence Act
19 Ad Hoc Committee on Artificial Intelligence, Policy Development Group (CAHAI-PDG) Artificial Intelligence in the Public Sector, CAHAI-PDG(2021)06provisional

KAPITEL 3

DEN TEKNOLOGISKE RAMME: KORT OM PROFILERINGSMODELLER

Borgernes rettigheder skal være i centrum også i en stadig mere digitaliseret forvaltning, hvor der bruges kunstig intelligens i form af f.eks. algoritmiske profileringsmodeller. Det kræver kendskab til rettigheder og den retlige kompleksitet på området, som vi gennemgik i kapitel 2 – men det kræver også kendskab til den teknologi, der bruges i sagsbehandlingen og de særegne udfordringer, der følger med den. Dette kendskab er nødvendigt både for dem, der træffer beslutninger om at tage profileringsmodeller i brug, dem som arbejder med deres resultater og dem, som skal føre tilsyn med brugen af modellerne.

I dette kapitel redegør vi derfor for de centrale forhold om profileringsmodeller, som det efter vores vurdering er helt nødvendigt at kende og forstå for bl.a. beslutningstagere, sagsbehandlere og tilsyn. For at sikre forståelse for såvel rettigheder som teknologi er en af denne rapports anbefalinger, at man styrker træningen og efteruddannelsen af myndighedspersoner (se kapitel 4).

Som nævnt er brugen af algoritmiske profileringsmodeller allerede udbredt blandt offentlige myndigheder, og i dette kapitel nævner vi en række eksempler – både på modeller i drift og på modeller, der har været overvejet i udviklingsfasen. Tilsammen illustrerer de nogle af de anvendelsesmuligheder og udfordringer, der er forbundet med modellerne. Eksemplerne tager udgangspunkt i fire cases beskrevet i detaljer i rapportens bilag 2 (se boks).

Beskrivelserne i dette kapitel kan synes komplekse for den uindviede læser. Det er dog uomgængeligt, at forståelsen af profileringsmodeller og deres brug i det offentlige kræver kendskab til visse tekniske fagtermer, som løbende vil blive brugt i rapportens analyse. Vi har forsøgt at lette læsningen ved at udarbejde en terminologiliste i rapportens bilag 1 (se boks).

FIRE CASES

Vi trækker gennem rapporten på fire eksempler på udvikling og/eller brug af algoritmiske profileringsmodeller i offentlige myndigheder. De fire cases er beskrevet i detaljer i rapportens bilag 2 og er:

- Gladsaxe Kommune: Opsporing af udsatte børn
- Udbetaling Danmark: Kontrol med udbetaling af offentlige ydelser
- Styrelsen for Arbejdsmarked og Rekruttering (STAR): Profilafklaring for nyledige
- Horsens Kommune: Prioritering af kontrolsager om social svindel

TERMINOLOGILISTE
I terminologilisten i rapportens bilag 1 finder du en alfabetiseret liste med definitioner af de tekniske termer, vi bruger i vi bruger denne rapport, og som er markeret med **fed** første gang de anvendes i hvert kapitel.

KORT OM ALGORITMER

En **algoritme** er et computerprogram, som udfører en bestemt serie af matematiske eller logiske operationer på et **datasæt**. En profileringsmodel er mere præcist et computerprogram, som på baggrund af forskellige data om en borger vurderer, om borgeren har en bestemt egenskab, kaldet **målegenskaben** og dermed f.eks. er berettiget til et tilskud eller en ydelse. En brugbar metafor kunne være, at algoritmen er en madopskrift, og data er de ingredienser, der skal til for at lave måltidet, dvs. for at modellen kan foretage en vurdering.[1]

Profileringsmodeller er en blandt mange former for **kunstig intelligens**. Vi bruger som udgangspunkt ikke begrebet kunstig intelligens i denne rapport til at beskrive de konkrete problemer og udfordringer, da det er en bred betegnelse for mange forskellige computerdrevne teknologier, der kan repræsentere og simulere aspekter af menneskelig tænkning – også teknologier, der ligger meget langt fra dem, vi beskæftiger os med her. I det følgende bruger vi de mere specifikke begreber (algoritmisk) profilering og (algoritmiske) profileringsmodeller.

De data, som en algoritmisk profileringsmodel arbejder med, vil typisk optræde som et sæt af **objekter** – oftest borgere – som hver indgår med en række **variable** – dvs. oplysninger om f.eks. køn, bopæl, alder eller hvad der måtte være relevant. I praksis kan objekterne i profileringsmodeller også være f.eks. husstande, virksomheder eller boligområder, men for at tydeliggøre, at modellens vurdering i sidste ende berører borgeren, vil vi i denne rapport tale om variable om borgere.

MASKINLÆRING OG LÆRINGSALGORITMER

Profileringsmodeller udvikles i dag i stigende grad ved **maskinlæring**. Hvor manuel udvikling af en profileringsmodel indebærer, at en person specificerer, hvordan modellen skal fungere, indebærer maskinlæring, at der bruges en **læringsalgoritme** i udviklingen af modellen. Vi behandler i denne rapport udelukkende profileringsmodeller baseret på maskinlæring.

Maskinlæring stiller store krav til myndighederne om udviklingsprocessen. Der skal træffes en række afgørende valg om modellens design herunder om, hvilken type model man vil anvende og hvordan den skal fungere, ligesom det også skal besluttes, hvordan modellen overhovedet skal bruges i sagsbehandlingen (se boks). I kapitel 4 diskuterer vi de rammer, der bør være for denne type valg og beslutninger.

Herudover skal det besluttes, hvilke data der skal indgå i **træningsdatasættet**. Formålet med en læringsalgoritme er at "træne" modellen til at give de bedst mulige resultater. Det gøres ved, at læringsalgoritmen analyserer et træningsdatasæt, som typisk kan bestå af oplysninger om allerede afgjorte sager på det område, som modellen skal arbejde med. Baseret på disse sager og øvrige oplysninger bestemmer læringsalgoritmen, hvordan modellen skal fungere. Læringsalgoritmen skal typisk bruge mange data, både i form af oplysninger om mange personer og i form af mange variable om hver enkelt person. Vi diskuterer, hvilke krav der bør stilles til denne data i kapitel 5.

Det stigende fokus på udvikling af modeller ved maskinlæring skyldes, at disse modeller ofte kan finde sammenhænge som mennesker overser, især meget komplekse sammenhænge, som er vanskelige for mennesker at indbygge i modellen. Ydermere er læringsalgoritmen i stand til matematisk at tilpasse modellen, så den er statistisk optimeret, med netop den kombination af **vægte** – dvs. værdier for de enkelte variable – som bedst muligt lader modellen vurdere målegenskaben. Læringsalgoritmen er imidlertid helt afhængig af, at man matematisk har defineret, hvad det vil sige, at modellen "bedst muligt" vurderer målegenskaben. Det stiller særlige krav til udviklerne, særligt i tilfælde, hvor modellens resultater skal indgå i skønsmæssige afgørelser. Vi diskuterer sådanne krav for anvendelse af modellerne i kapitel 6.

De modeller, som udvikles ved maskinlæring er som nævnt ofte mere komplekse end manuelt udviklede modeller. Det betyder på den ene side, at de kan være bedre til løse opgaven med at vurdere den egenskab ved borgeren, som de er udviklet til. Den øgede kompleksitet betyder på den anden side også, at det i mange tilfælde kan være meget vanskeligt selv for specialister at forstå, præcist hvordan modellen fungerer. Dette er et centralt tema, som vi vender tilbage til i kapitel 7 om transparens.

Herudover kan risikoen for diskrimination blive videreført og forstærket af en profileringsmodel. Det kan bl.a. ske ved, at modellen overtager og reproducerer skævheder i datasættet, f.eks. en underrepræsentation af bestemte grupper eller et diskriminerende element i den hidtidige afgørelsespraksis. Disse problemstillinger ser vi nærmere på i kapitel 8.

FULDAUTOMATISEREDE AFGØRELSER OG AUTOMATISERET BESLUTNINGSSTØTTE

Offentlige myndigheder kan bruge algoritmiske profileringsmodeller på forskellige måder. På et generelt niveau kan man skelne mellem fuldautomatiserede afgørelser og automatiseret beslutningsstøtte.

Fuldautomatiserede afgørelser vil sige, at modellens resultat i sig selv fører til en forvaltningsretlig afgørelse, f.eks. hvis en myndighed bruger modellen til automatisk at godkende eller afvise ansøgninger om tilskud.

Automatiseret beslutningsstøtte vil sige, at modellens rolle er enten at foreslå, hvilke borgere, der skal oprettes sager om, eller at prioritere eller at oplyse allerede eksisterende sager, hvor der i sidste ende skal træffes afgørelser af mennesker, f.eks. hvis en myndighed bruger modellens resultat som en af flere input i en sagsbehandlers vurdering af, hvorvidt en familie skal tilbydes forebyggende foranstaltninger.

Der er imidlertid en væsentlig gråzone imellem de to typer funktioner. Sagsbehandleren kan få afgørende indflydelse på en proces, som oprindeligt var tænkt som en fuldautomatiseret proces ved f.eks. at kunne underkende modellens afgørelse eller godkende den, før den er endelig. Omvendt kan modellens vurdering tillægges stor vægt som beslutningsstøtte, f.eks. hvis sagsbehandleren tager udgangspunkt i eller sjældent afviger fra modellens vurdering.

Et væsentligt spørgsmål i vurderingen af, om der er tale om automatiseret beslutningsstøtte eller fuldautomatisering er derfor, hvordan den menneskelige kontrol sikres. Dette vender vi tilbage til i kapitel 6.

FEJLRATER OG FEJLTYPER

Profileringsmodeller er fejlbarlige. En models **fejlrate** er den mest grundlæggende målestok for modellens kvalitet og dækker over, hvor ofte modellen vurderer personen forkert. Hvis modellen f.eks. klassificerer 100 personer og vurderer fem personer forkert, er fejlraten 5%. Modellens fejlrate påvirkes af **bias** og **varians** (se nedenfor).

Modellers vurderinger kan være forkerte på forskellige måder – man taler om forskellige **fejltyper**. En model, som har til opgave at klassificere borgere, kan f.eks. begå to typer fejl: Den kan lave et falsk positivt resultat, hvor den vurderer, at en person besidder målegenskaben, selvom dette ikke er tilfældet og omvendt et falsk negativt resultat, hvor modellen vurderer, at en person ikke besidder målegenskaben, selvom det faktisk er tilfældet.

Balancen mellem fejltyperne afhænger af, hvor man sætter **tærskelværdien**, dvs. grænsen for, hvornår en borger vurderes positivt eller negativt i forhold til målegenskaben. Typisk kan man kun reducere antallet af falske positive på bekostning af flere falske negative eller omvendt, og fastsættelsen af beslutnings-tærsklen er derfor et vigtigt designvalg med retssikkerhedsmæssige implikationer, som vi vender tilbage til i kapitel 6.

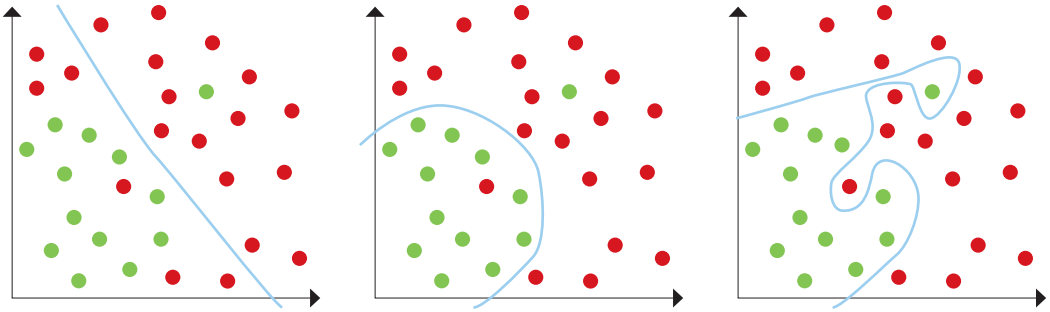
BIAS OG VARIANS

Den rette balance mellem modellens bias og varians er en af de væsentligste opgaver i udviklingen af en model – bl.a. fordi det er en central metode til at opnå en lav fejlrate.

Bias vil sige modellens iboende tendens til at vurdere personer på bestemte måder, som kan få modellen til at overse eller misrepræsentere sammenhænge mellem variable og målegenskaben. En høj bias giver en for "grovkornet" sortering af borgere med en høj fejlrate til følge. Man taler også om, at modellen kan være **undertilpasset** til træningsdatasættet.

Varians vil sige modellens følsomhed overfor variationer i træningsdatasættet. Er modellens varians for høj, vil den repræsentere sammenhænge, som kun optræder tilfældigt i træningsdatasættet, og som ikke kan overføres til nye borgere eller nye sager. En høj varians giver altså en for "finkornet" sortering af borgere, hvilket også fører til en høj fejlrate, når modellen sættes i drift. Man taler også om, at modellen kan være **overtilpasset** til træningsdatasættet.

BIAS OG VARIANS



Høj bias
(undertilpasning)

Balance mellem bias og varians
(korrekt tilpasning)

Høj varians
(overtilpasning)

Bias og varians har betydning for modellens **generaliserbarhed,** dvs. den evne til også at sortere nye data korrekt, når den bliver sat i drift. En model med høj bias har høj generaliserbarhed, men til gengæld også en høj fejlrate. En model med høj varians har en lav fejlrate i træningsdatasættet, men vil have lav generaliserbarhed og dermed høj fejlrate, når modellen præsenteres for nye data. Den optimale model skal ramme den rette balance mellem bias og varians, for at opnå lav fejlrate og høj generaliserbarhed (se figur). Også dette kommer vi ind på i kapitel 6.

EKSEMPLER PÅ ALGORITMISKE PROFILERINGSMODELLER

I det følgende beskriver vi en række almindelige typer af profileringsmodeller. Moderne maskinlæring arbejder også med en mange andre modeltyper, men vi har valgt disse, da de illustrerer nogle principper, som også findes i mange beslægtede modeller. Profileringsmodeller i det offentlige vil desuden ofte bygge på en af disse grundmodeller, og de egner sig derfor godt til at belyse de udfordringer, som vi behandler i denne rapport.

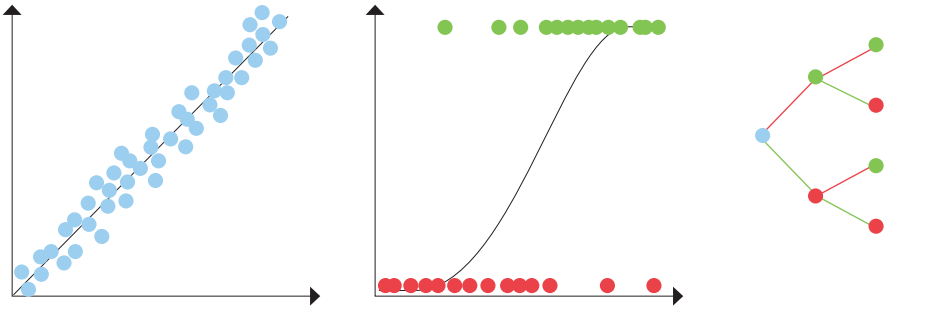
Beskrivelserne er på ingen måde udtømmende, eftersom der bag hver model ligger et komplekst stykke teknologisk udviklingsarbejde. For de særligt interesserede findes der righoldig teknisk litteratur om dette, men i denne rapport er formålet blot at forklare de grundlæggende mekanismer i de forskellige modeltyper, rettet mod den ikke-teknikkyndige læser.

Overordnet set kan en profileringsmodel levere to forskellige typer resultater, baseret på, hvordan modellen behandler den målegenskab, den er sat til at vurdere:

Den ene type vil afgøre, på hvilket niveau borgeren besidder målegenskaben, f.eks. en beregning af den forventede levetid for en given gruppe patienter. Her placerer modellen borgeren på en skala baseret på de oplysninger, den har til rådighed. En sådan model kaldes en **regressionsmode**l. I det følgende beskriver vi en **lineær regressionsmodel** eksempel på denne modeltype.

Den anden type model vil afgøre, hvorvidt borgeren besidder målegenskaben eller ej. Et eksempel kan være, om borgeren er berettiget til en ydelse eller ej. Her er der tale om en enten/eller-afgørelse, hvor modellen på baggrund af en række oplysninger om borgeren enten vender tilbage med et positivt eller negativt resultat. En sådan model kaldes en **klassificeringsmodel**. I det følgende beskriver vi **logistisk regression** og prædiktive **beslutningstræer**, som begge er meget brugte klassificeringsmodeller (se figur).

TRE MODELTYPER



Lineær regression | Logistisk regression | Beslutningstræ

LINEÆRE OG LOGISTISKE REGRESSIONER

Lineære regressioner og logistiske regressioner er to meget enkle og relativt almindelige modeltyper, som kan bruges til profileringsmodeller. Vi beskriver dem her samlet, da de på mange måder minder om hinanden, men ret beset er kun lineær regression en regressionsmodel, mens logistisk regression er en klassificeringsmodel.

En lineær regressionsmodel f.eks. vil alene fordele borgerne på en skala alt efter målegenskabens værdi – f.eks. ved at vurdere patienters forventede levetid baseret på en række sundhedsoplysninger. En logistisk regressionsmodel vil derimod klassificere borgerne i to grupper, baseret på en vurdering af sandsynligheden for, at de besidder målegenskaben – f.eks. hvorvidt en straffet person vil begå ny kriminalitet. Her vil modellen etablere en risikoskala, og baseret på en **tærskelværdi** vil den herefter inddele borgerne i to grupper, alt efter om de vurderes i risiko eller ej. En tærskelværdi er værdien for, hvilket niveau af sandsynlighed, der placerer borgeren i den ene eller anden gruppe.

TABSFUNKTION

Et centralt valg i udviklingen af begge modeller – og flere af de øvrige, vi omtaler i dette kapitel – er bestemmelsen af **tabsfunktionen.** En tabsfunktion er en matematisk funktion, som fortæller læringsalgoritmen, hvad der tæller som henholdsvis en god og en dårlig vurdering. Læringsalgoritmen bruger tabsfunktionen til at evaluere modellens kvalitet, og den har derfor afgørende betydning for den endelige model.

I en lineær regressionsmodel er en simpel og ofte anvendt tabsfunktion gennemsnittet af kvadrerede fejl (eng. "mean squared error"). En model, der f.eks. udvikles til at anslå, hvilken pris huse kan sælges til (baseret på bl.a. størrelse, beliggenhed og stand) vil blive trænet via et datasæt med allerede solgte huse. Læringsalgoritmen måler modellens kvalitet ved at finde forskellen mellem

modellens vurderede og den faktiske pris, huset blev solgt for. Hver af disse afvigelser ganges med sig selv, så store afvigelser får langt større betydning end små afvigelser. Til sidst udregnes gennemsnittet af de kvadrerede afvigelser, og denne værdi vil læringsalgoritme så forsøge at minimere.

I en logistisk regressionsmodel er den simpleste tabsfunktion at måle summen af forkerte klassificeringer. Hvis modellen trænes til f.eks. at vurdere risikoen for fornyet kriminalitet hos allerede straffede, vil den blive trænet via et datasæt af borgere, hvoraf nogle allerede har begået ny kriminalitet. Tabsfunktionen vil så definere modellen som god, hvis dens vurdering matcher den reelle fordeling i datasættet, og algoritmen vil søge at minimere fejlraten.

I praksis er bestemmelsen af tabsfunktion altid betydeligt mere kompleks, end vi her har beskrevet, og der kan indgå mange typer tabsfunktioner i udviklingen af en model. Der findes typisk ikke én korrekt tabsfunktion for en ønsket type vurdering, og bestemmelsen af tabsfunktionen involverer uundgåeligt en afvejning af forskellige hensyn op mod hinanden og udgør et vigtigt led i modellens udvikling og design.

POLYNOMISKE VARIABLE OG REGULARISERINGSFAKTORER

Alle algoritmiske modeller arbejder som nævnt med en række oplysninger om borgerne, såkaldt variable. Ofte vil der være interaktioner mellem de forskellige variable, hvor værdien for en variabel påvirker, hvilken indflydelse en anden variabel har på målegenskaben. Det kan f.eks. være en medicinsk diagnose, hvor alder selvstændigt har indflydelse på om en patient skal diagnosticeres med en given sygdom, men hvor alder også er relevant for modellens vurdering af sammenhængen mellem personens vægt og sygdommen, fordi kropsvægt har stor betydning for diagnosen for unge, men lille betydning for ældre.

Det kan være vanskeligt for nogle modeller at repræsentere denne type interaktioner. Derfor kan udvikleren vælge at modellere interaktionerne ved at indsætte såkaldt **polynomiske variable** i modellen. En polynomisk variabel er en ny variabel, som er baseret på en eller flere eksisterende variable, men som giver en mere kompleks repræsentation af dem i modellen. Det vil ofte være meget vanskeligt for udvikleren at finde og manuelt tilføje netop de relevante polynomiske variable, og en af styrkerne ved maskinlæring er, at udvikleren kan indsætte et stort sæt af polynomiske variable og lade læringsalgoritmen vurdere, hvilke der er relevante.

Der er imidlertid to ulemper ved at indsætte store sæt af polynomiske variable. De følger begge af, at man på den måde dramatisk øger antallet af variable, som modellen arbejder med.

Den første ulempe er risikoen for, at modellen overtilpasses – også kaldet høj varians som nævnt ovenfor. Overtilpasning betyder, at modellen tilpasses **træningssættet** så nøje, at den inddrager tilfældigheder, som kun findes i træningssættet. Det sænker modellens generaliserbarhed, fordi den så vil få en tendens til at lave flere fejl, når den anvendes på nye data når den sættes i drift.

Den anden ulempe er, at jo mere man modellerer interaktionerne mellem variable – f.eks. ved at indsætte polynomiske variable – jo højere bliver modellens kompleksitet. Og jo højere kompleksiteten er, jo sværere er det for mennesker at forstå modellens funktion og hvilken betydning, den har tillagt de enkelte oplysninger om borgeren, og dermed baggrunden for modellens resultat. Det fører til problemer med den algoritmiske transparens; en problematik, som vi udfolder mere i kapitel 7.

Udvikleren vil ofte forsøge at reducere overtilpasning og kompleksitet ved at introducere en **regulariseringsfaktor** i tabsfunktionen. Regulariseringsfaktoren definerer en omkostning for hver vægt, der stiger med vægtens størrelse. Det tvinger læringsalgoritmen til at afveje værdien ved bedre vurderinger på **testsættet** mod værdien ved at minimere vægtene, og giver den derved en tendens til at træne en mere enkel, mindre tilpasset model. Valget af en regulariseringsfaktor kræver imidlertid omhyggelig kalibrering, for jo højere regulariseringfaktoren er, jo enklere bliver modellen og dermed øges typisk dens fejlrate, mens en for lav regulariseringfaktor på den anden side kan føre til overtilpasning. Valg af regulariseringsfaktor er derfor endnu et centralt designvalg, som skal træffes under udviklingen af modellen.

PRÆDIKTIVE BESLUTNINGSTRÆER

En anden almindeligt benyttet modeltype er et såkaldt **prædiktivt beslutningstræ** (eng. "predictive decision tree"). Beslutningstræer er både i udvikling og funktion væsentligt anderledes end lineær regression og logistisk regression, som vi har set på ovenfor. Modeltypen har flere fordele, bl.a. evnen til at modellere ikke-lineære sammenhænge og interaktioner, samt relativ høj transparens. Den har imidlertid også den udfordring, at den kan have svært ved at undgå overtilpasning.

Et prædiktivt beslutningstræ består af en forgrenet serie af beslutningspunkter (eng. "decision nodes"). Træets grene ender i slutpunkter (eng. "end nodes"). I hvert beslutningspunkt vurderer modellen en given variabel – oplysning om borgeren – og deler på den baggrund alle borgere op i to grupper, som sendes videre i enten den ene eller den anden forgrening af træet. Denne proces fortsætter med nye beslutningspunkter om nye variable, indtil processen når et slutpunkt, hvor hver borger klassificeres i forhold til den målegenskab, algoritmen er sat til at vurdere.

I udviklingsarbejdet med et prædiktivt beslutningstræ – dvs. når man træner en læringsalgoritme for denne modeltype – skal der derfor træffes mindst to afgørende designvalg.

Det første er at bestemme **splitbetingelsen**, som fortæller læringsalgoritmen, hvilke oplysninger – eller variable – den skal vurdere, og ud fra hvilke kriterier, i hvert beslutningspunkt, eller sagt i tekniske termer, hvilke test algoritmen skal udføre på data, og i hvilken rækkefølge. Et beslutningstræ opererer ikke med en tabsfunktion i samme forstand som lineær og logistisk regression, men splitbetingelsen spiller en lignende rolle.

Det andet designvalg er at bestemme **slutbetingelsen**, der fortæller lærings-algoritmen, hvornår den skal definere et slutpunkt i processen, fremfor at definere endnu et beslutningspunkt. Hvis læringsalgoritmen fik lov til at blive ved med at definere nye beslutningspunkter, ville træet i sidste ende bestå af en række slutpunkter, som hver indeholdt én person i træningssættet. Modellen ville med andre ord blive ved med at finde små forskelle mellem borgerne, som ikke spiller en rolle for den målegenskab, algoritmen er sat til at måle, og ville dermed få en ekstremt høj varians og overtilpasning (se ovenfor).

Slutbetingelsen sikrer altså, at modellen er generaliserbar, dvs. at den også fungerer på nye data med nye borgere. En slutbetingelse kan f.eks. være, at gruppen er helt "ren", dvs. kun indeholder personer som besidder (eller ikke besidder) målegenskaben, at der er færre end et bestemt antal personer i gruppen, eller at personerne har været testet i et bestemt antal beslutningspunkter.

Det sidste skridt i træningen af modellen er at bestemme, hvilke slutpunkter, som skal klassificere personer positivt eller negativt i forhold til målegenskaben. For hvert slutpunkt sammenligner læringsalgoritmen antallet af personer i slutpunktet, som har målegenskaben, med antallet af personer i slutpunktet, som ikke har målegenskaben. Denne rate kan tolkes som sandsynligheden for, at en person i slutpunktet vil have målegenskaben. Her introducerer udvikleren, ligesom i logistisk regression, en tærskel for denne sandsynlighed, som modellen bruger til at vurdere, om dette slutpunkt klassificerer borgeren positivt eller negativt. Ligesom for logistisk regression er valget af denne tærskel derfor et vigtigt designvalg under træningen af et beslutningstræ.

GLADSAXE KOMMUNE: OPSPORING AF UDSATTE BØRN

Gladsaxe kommune gik i midten af 2017 i gang med at udvikle en "dataunderstøttet opsporingsmodel" som led i kommunens projekt "Tidlig opsporing". Formålet med modellen var at identificere børn med særligt høj risiko for udsathed og mistrivsel tidligt i deres liv og dermed styrke kommunens mulighed for at støtte børnene. Kommunen valgte at udvikle et prædiktivt beslutningstræ, som regelmæssigt skulle kunne analysere data for samtlige børnefamilier i kommunen og vurdere den statistiske sandsynlighed for mistrivsel for hver enkelt familie. I de tilfælde, hvor den vurderede sandsynlighed oversteg en bestemt tærskel, ville modellen gøre kommunale sagsbehandlere opmærksom på dette. Disse sagsbehandlere ville derpå kunne beslutte at foretage en individuel, manuel vurdering af de relevante familier.

På baggrund af en indledende analyse udvalgte kommunen i alt 44 variable, som vurderedes at kunne være relevante for modellen. Det drejede sig om bl.a. forældres beskæftigelsesstatus og -historik, statsborgerskab, faderskabsforhold, og forældres bopæl, samt om barnets brug af tandpleje, underretninger, pasningsforhold, og sprog.

Gladsaxe kommune besluttede imidlertid at afbryde udviklingen af modellen, før kommunens udviklere lagde sig fast på betingelser for splitbetingelser og slutbetingelser, fastsatte tærsklen for sandsynligheder, eller testede og validerede modellen. Det skyldtes tests, som viste, at modellens fejlrate ville blive for høj, først for fremmest fordi antallet af historiske sager var begrænset. Kommunen havde kun 117 sager med udsatte børn i alderen 0-6 år, som kunne fungere som positive eksempler i datasættet. Kommunen overvejede to mulige løsninger på denne udfordring. Den første var at træne modellen på et større datasæt, som dels kunne inkludere data om større børn, og dels data om sager fra andre kommuner. Den anden løsning var at inkludere visse variable, som kommunen vurderede kunne have stærk statistisk signifikans, herunder især sundhedsdata fra den kommunale sundhedspleje og det kommunale rusmiddelcenter.

TILFÆLDIGE SKOVE

Prædiktive beslutningstræer har en række styrker, som forklarer deres popularitet, men de har som nævnt også svagheder. De er således meget følsomme overfor variationer i det sæt af data, som modellen trænes på. Selv meget små forskelle i træningssættet kan føre til meget store forskelle i det resulterende beslutningstræ. Det skyldes bl.a., at ændringen af et beslutningspunkt påvirker de beslutningspunkter, som findes længere ude på den samme gren af beslutningstræet. En lille ændring i træningssættet, som fører til en ændring af et beslutningspunkt tidligt i beslutningstræet, kan derfor have en sneboldeffekt, som påvirker hele modellen.

En udbredt løsning på disse udfordringer er at bruge en såkaldt tilfældig skov-model (eng. "random forest"). En **tilfældig skov** er en sammensat model, som kombinerer et sæt beslutningstræer. Fordi modellen trækker på mange forskellige beslutningstræer, har en tilfældig skov ofte både lavere følsomhed overfor variationer i træningssættet og lavere fejlrate end et enkelt beslutningstræ. En tilfældig skov består af en hel række beslutningstræer, der hver for sig vurderer personen. Profileringsmodellen foretager den endelige klassificering baseret på resultatet af de enkelte træers klassificeringer, f.eks. ved at vælge den klassificering som der er flertal for blandt de enkelte træer.

En læringsalgoritme træner en tilfældig skov ved at udvikle en række forskellige træer på det samme sæt af træningsdata. Træerne gøres forskellige ved tilfældigt at begrænse udvalget af variable, som læringsalgoritmen kan anvende i hvert enkelt beslutningspunkt, og dermed kan læringsalgoritmen skabe en serie af meget forskellige beslutningstræer. Den tilfældige skov løser dermed udfordringen med varians og overtilpasning ved at lade tilfældighederne i de forskellige træer udligne hinanden, og den er derfor typisk mere robust og har en lavere fejlrate.

Imidlertid er en tilfældig skov-model også langt mindre transparent end et almindeligt beslutningstræ. Det er ofte meget svært at overskue, hvordan det store sæt af beslutningstræer samlet set behandler en person, f.eks. hvilken rolle en bestemt variabel spiller for modellens klassificeringer. Her er det altså nødvendigt at afveje modellens funktion mod behovet for transparens. Dette behandler vi nærmere i kapitel 7.

ØVRIGE MODELLER

Offentlige myndigheder bruger også andre typer af profileringsmodeller, end de, som vi har beskrevet ovenfor. Vi nævner her tre eksempler, nemlig **Support Vector Machine** (SMV), **naiv Bayes-model** samt **klyngeanalyse**.

En SVM er ligesom logistiske regressioner og beslutningstræer i udgangspunktet en klassificeringsalgoritme, dvs. at den forsøger at vurdere, om en person har eller ikke har en målegenskab. En SVM-model minder i mange henseender om en logistisk regression: Modellen består af en funktion, som giver hver variabel en vægt, og tager summen af de vægtede variable, hvorefter personen klassificeres enten positivt eller negativt i forhold til målegenskaben, baseret på en tærskelværdi. Men mens en logistisk regressionsmodel lader alle vurderinger få indflydelse på, hvordan læringsalgoritmen opdaterer modellens vægte, så prioriterer SVM-modellen de vurderinger, som ligger tættest på tærskelværdien og der derfor er mest usikkerhed om, og det gør den mere generaliserbar.

Vi så ovenfor, at man i en logistisk regressionsmodel kan modellere interaktioner mellem variable ved at benytte sig af polynomiske variable. SVM-modellen har udbredt anvendelsen af en anden teknik til formålet, det såkaldte **kernetrick** (eng. "kernel trick"). Kernetricket fungerer ved at erstatte de eksisterende variable

med et nyt sæt variable, som hver er defineret som ligheden mellem en allerede klassificeret person i datasættet og den person, som skal klassificeres. Billedligt kan man sige, at kernetricket gør hver person i træningssættet til et punkt i et rum, der består af lige så mange dimensioner, som der er variable. Når SVM-modellen skal vurdere en ny person, så undersøger den hvilke personer, der befinder sig tæt på og langt fra i dette rum, og klassificerer personen ligesom de, der ligger nærmest.

Ligesom med justering af tabsfunktion, påvirker valg og justering af kernefunktion afgørende modellen, og er et centralt designvalg. En stejl kernefunktion, som kun tilskriver værdi til personer som er meget tæt på hinanden, vil give modellen en meget lav fejlrate i klassificeringen af træningssættet, men reducere modellens generaliserbarhed når den er i drift. Omvendt vil en mere flad kernefunktion tilskrive værdi til personer, som ligger længere fra hinanden, hvilket typisk øger fejlraten for træningssættet, men reducerer sandsynligheden for overtilpasning.

Også i kernetricket opstår der vanskeligheder med at sikre transparens. En SVM-model med kernetrick tilskriver ikke variable i det oprindelige datasæt en vægt, som kan aflæses som udtryk for modellens vurdering af den statistiske sammenhæng mellem den pågældende variabel og målegenskaben. Det kan derfor være svært, selv for eksperter, at gennemskue hvorfor en SVM klassificerer en konkret person på den ene eller den anden måde.

En naiv Bayes klassifikationsmodel (eng. "naive Bayes classification") minder også om en logistisk regressionsmodel. Også her forsøger modellen at bestemme, hvordan hver variabel påvirker sandsynligheden for, at personen besidder målegenskaben. Læringsalgoritmen behandler hver af de variable individuelt, og beregner, hvad den betingede sandsynlighed er, for at en person besidder målegenskaben, givet værdien for netop denne variabel. Når modellen skal klassificere, beregnes den samlede betingede sandsynlighed for, at personen besidder eller ikke besidder målegenskaben, givet de enkelte variables værdier, og personen klassificeres efter den højeste sandsynlighed.

En klyngeanalyse (eng. "clustering") er ret beset ikke en egentlig modeltype, men snarere en teknik på linje med f.eks. klassifikation. Klyngeanalysen adskiller sig ved, at den ikke er styret af kendskab til målegenskaben i historiske sager, men alene af mønstre i det datasæt, som den behandler (eng. "unsupervised learning"). Analysen er designet til at finde klynger af f.eks. sager i datasættet, som minder om hinanden. Groft sagt vurderes to sager at være del af en fælles klynge, hvis tilstrækkelig mange af værdierne for de variable ligger tilstrækkelig tæt. Når analysen har fundet klynger i sættet af sager, vil der ofte være enkelte sager, som ikke er med i nogen klynge (eng. "outliers"). Modellen er derfor velegnet til at fremhæve sager, som på denne måde skiller sig ud og derfor bør udtages til menneskelig kontrol.

UDBETALING DANMARK: KONTROL MED UDBETALING AF YDELSER

I 2014 oprettede Udbetaling Danmark "Den Fælles Dataenhed", som bl.a. fik til opgave at udvikle datadrevne modeller, der kunne styrke kontrollen med udbetalinger.

Dataenheden arbejder med at samkøre registre og bruge profileringsmodeller på disse data til at identificere personer, som har en forhøjet risiko for uberettiget udbetalte ydelser. De brugte modeller varierer betragteligt i kompleksitet og karakter. Langt de fleste og især de første modeller, som man udviklede, er relativt simple, manuelt konfigurerede "udsøgningsmønstre" – en form for meget enkelt beslutningstræ, der udfører en serie af tests på hver person, som fører til at personen klassificeres enten positivt eller negativt. I de seneste år har man i stigende grad taget mere komplekse modeller baseret på maskinlæring i anvendelse, herunder naiv Bayes-klassifikation og klyngeanalyse. Dataenheden arbejder løbende med at udvælge og udvikle nye modeller baseret på erfaringer hos Udbetaling Danmark og kommunerne. I 2019 brugte man ca. 60 forskellige modeller til løbende analyse af registersamkørte data.

Når en model er udviklet og implementeret, foregår kontrollen ved, at modellen regelmæssigt – typisk en gang om ugen – analyserer registersamkørte data fra de relevante personer, f.eks. modtagere af en konkret ydelse. Modellen genererer herefter en "undringsliste" med de borgere, som statistisk set har høj risiko for uberettiget at modtage ydelsen. Sagerne optræder i anonymiseret form, da listen kun indeholder sagsnumre samt en overordnet beskrivelse af hvilke variable, modellen har behandlet.

Næste skridt tages af sagsbehandlere i kontrolafdelinger ved kommunerne og Udbetaling Danmark. Disse kan ved at logge på Udbetaling Danmarks system få adgang til de relevante dele af listen, f.eks. beboere i den pågældende kommune. Sagsbehandlerne kan vælge at trække sager fra undringslisten ud til kontrol og herved få adgang til den fulde sag i ikke-anonymiseret form. Sagsbehandleren er i så fald forpligtet til at oplyse borgeren om, at deres sag er blevet udtrukket til kontrol. Sagsbehandleren foretager derpå en manuel og individuel vurdering af, om der er grund til at gå videre med sagen. Hvis det ikke er tilfældet, lukkes sagen igen, og borgeren gøres opmærksom på at sagen er lukket. I modsat fald forfølges sagen, typisk ved at sagsbehandleren som første skridt beder borgeren om at indsende supplerende oplysninger.

Som led i samarbejdet mellem dataenheden og kommunens kontrolenheder modtager Udbetaling Danmark løbende tilbagemeldinger på sagsbehandlingen af de udvalgte sager, og benytter resultaterne til at opdatere deres modeller, især med henblik på at reducere antallet af falsk positive, dvs. sager som af modellen vurderes til at have høj risiko, men hvor der ikke kan konstateres uberettiget udbetaling.

Udbetaling Danmark fremhæver selv, at brugen af modeller på registersamkørte data som automatiseret beslutningsstøtte har flere fordele. For det første fører det til mere ensartet sagsbehandling og effektiviserer sagsbehandlingen ved at lade sagsbehandlere fokusere på de sager som har høj risiko. For det andet gør modellerne det muligt at opdage sager om uberettiget udbetaling, som ellers ville være meget vanskelige at spore, og at opdage sagerne tidligere, især når der er tale om fejl. Endelig reducerer det antallet af falsk positive, dvs. sager hvor borgeren forstyrres af en ubegrundet henvendelse.

NOTER

1 Se generelt for dette kapitel: Freitas, (2014), Comprehensible classification models: a position paper, SIGKDD Explorations Newsletter 15(1): 1–10, Ribeiro, Singh og Guestrin (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Storkey (2009) When Training and Test Sets Are Different: Characterizing Learning Transfer, 10.7551/mitpress/9780262170055.003.0001

DEL II

FORUDSÆTNINGER FOR EN RETTIGHEDSBASERET TILGANG

I denne del af rapporten giver vi vores bud på de grundlæggende forudsætninger for en rettighedsbaseret tilgang til profileringsmodeller.

Brugen af profileringsmodellerne rejser en række rettigheds- og retssikkerhedsmæssige udfordringer, som gør det nødvendigt at sætte nogle styringsmæssige rammer for beslutningen om at bruge en profileringsmodel, og den efterfølgende udviklings- og anvendelsesfase. Det er afgørende, at udfordringerne bliver adresseret tidligt i beslutnings- og udviklingsfasen og kontinuerligt evalueres under hele modellens livscyklus. Analysen i denne del af rapporten fører frem til en række anbefalinger, der alle har til formål at sikre struktur om og styring og kontrol med brugen af profileringsmodeller.

I kapitel 4 ridser vi kravene til beslutninger op og introducerer AI-konsekvens-analyser som et af de vigtigste redskaber for styring. Dette redskab går igen i alle rapportens tematiske dele.

I kapitel 5 ser vi på kravene til inputdata, dvs. data, som profileringsmodellen trænes på, og i kapitel 6 fokuserer vi på kravene til brugen af modellens output dvs. den vurdering, modellen foretager.

KAPITEL 4

BESLUTNING OG STYRING

I dette kapitel ser vi for det første på behovet for regler og rammer for myndighedernes beslutning om at anvende profileringsmodeller. For det andet ser vi på behovet for en række konkrete styringsværktøjer, der skal sikre, at rettigheds- og retssikkerhedsmæssige spørgsmål bliver adresseret tidligt i beslutnings- og udviklingsfasen, og at de kontinuerligt evalueres under hele modellens livscyklus. Disse konkrete styringsværktøjer tager form af en vejledning, der ensretter myndighedernes brug af modellerne; udarbejdelse af såkaldte Artificial Intelligence (AI) konsekvensanalyser, hvori bl.a. modellens formål og risici beskrives; et styrket tilsyn med modellerne efter de er sat i drift, og endelig tiltag rettet mod at styrke myndighedernes tekniske og rettighedsmæssige forståelse.

I praksis vil offentlige myndigheders brug af profileringsmodeller stort set altid indebære et samarbejde med private aktører, der udvikler og udbyder teknologiske løsninger. Denne form for samarbejde rejser en række alvorlige udfordringer, som imidlertid ligger uden for rammerne af denne rapport. Der kan bl.a. opstå en risiko for magtforskydning, hvis myndigheder bliver afhængig af bestemte udbydere eller udviklere, der får indflydelse på de dele af forvaltningen, der bruger deres teknologi.[1]

Myndighederne bør stille krav til udbyderne tidligt i forløbet og som en integreret del af udbudsfasen, ligesom betingelserne for samarbejdet bør indgå i myndighedens overvejelser og i de AI-konsekvensanalyser, som vi anbefaler, at der gennemføres. Forud for offentlige udbud af profileringsmodeller bør der som fast praksis foretages en risikovurdering, hvor teknologiens mulige indvirkning på rettigheder afklares og hvor der foretages en screening af potentielle udbydere af teknologien. Dernæst bør myndighederne sikre, at rettigheder indgår som betingelser i kontrakten.[2]

Vi anerkender vigtigheden af disse udfordringer, men vil ikke - indenfor rammerne af denne rapport - beskæftige os yderligere med dem, og henviser i stedet til vores øvrige arbejde på området.[3]

HJEMMEL VED FULDAUTOMATISEREDE AFGØRELSER

Er der – eller bør der være – krav om udtrykkelig lovhjemmel for myndighedernes brug af profileringsmodeller?

Myndighederne er styret af et legalitetsprincip, der indebærer, at de kun må træffe afgørelser, når de har hjemmel til det, og i øvrigt ikke må handle i strid med

loven eller andre bindende regler og principper. Kravet om hjemmel er derfor fundamentalt for forvaltningen.

Udgangspunktet er, at de samme regler gælder, uanset hvordan en myndighed træffer sine afgørelser. Hvis myndigheden har hjemmel til at træffe en afgørelse, f.eks. om udbetaling af tilskud, kræves der ikke særskilt hjemmel at bruge en profileringsmodel til at træffe afgørelsen. Som vi kommer ind på nedenfor, bør myndigheden imidlertid tidligt i beslutningsfasen forholde sig til, hvorfor den ønsker at anvende modellen, modellens egnethed til opgaven og oplysninger om eventuelle alternative modeller eller løsninger.

Når det gælder fuldautomatiserede afgørelser – dvs. hvor modellens resultat i sig selv fører til en forvaltningsretlig afgørelse – stiller EU-databeskyttelses-forordningen krav om, at disse skal være "hjemlet i EU-ret eller medlemsstaternes nationale ret" og at borgerens rettigheder i øvrigt sikres.[4] Det er imidlertid uafklaret, om bestemmelsen i forordningen stiller krav om udtrykkelig hjemmel for selve automatiseringen (processen), eller blot om, at der er en hjemmel til at træffe (den materielle) afgørelse.[5] I sidstnævnte tilfælde indebærer forordningens regler ikke yderligere for offentlige myndigheder, end hvad der almindeligvis følger af hjemmelskravet.

Det er blevet diskuteret i litteraturen, om forfatningsretlige betragtninger om hensynet til at sikre demokratisk legitimitet og lovgivers eksplicitte stillingtagen kan tale for et krav om lovhjemmel for brugen af modellen, i hvert fald, når brugen af modellen fører til "omfattende ændringer i forvaltningens organisation, styrings- og ansvarsforhold og arbejdsgange, samt for samspillet mellem myndigheder og borgere."[6]

Er der tale om fuldautomatiserede afgørelser, som har indgribende virkning på borgeren, kan også retssikkerhedsmæssige betragtninger tale for, at der skal være hjemmel i lov for brugen af modellen. Dette gælder især, hvis selve modellens brug er med til at øge de retssikkerhedsmæssige risici, hvilket vi vurderer er tilfældet både for modellens output (se kapitel 6), dens manglende transparens (se kapitel 7) og dens risiko for diskrimination (se kapitel 8)

Ud fra disse betragtninger er det vores vurdering, at der bør gælde et krav om lovhjemmel for fuldautomatiserede afgørelser. De kommende kapitler tjener til at udfolde denne anbefaling yderligere. Dette er desuden ikke det eneste krav vi stiller til brugen af fuldautomatiserede afgørelser. I kapitel 6 anbefaler vi, at fuldautomatiseret brug kun anvendes i afgørelser, der er baseret på regler med entydige kriterier.

Vi anbefaler, at Justitsministeriet tager initiativ til at indføre regler i forvaltningsloven med krav om udtrykkelig lovhjemmel for offentlige myndigheders brug af profileringsmodeller til fuldautomatiserede afgørelser.

De fleste af de tilfælde vi kender til, hvor myndighederne bruger profilerings-modeller, vedrører slet ikke fuldautomatisering, mens derimod beslutningsstøtte. Det må da også forventes, at det i sager af særlig indgribende karakter eller i sager, der kræver komplekse juridiske og fagkyndige afvejninger oftest vil være beslutningsstøtte og ikke fuldautomatisering, som modellerne anvendes til (se boks.

EKSEMPLER PÅ BESLUTNINGSSTØTTE

I kapitel 3 introducerede vi en skelnen mellem fuldautomatiserede afgørelser og automatiseret beslutningsstøtte. I fuldautomatiserede afgørelser fører modellens resultat i sig selv til en forvaltningsretlig afgørelse, hvorimod automatiseret beslutningsstøtte vil sige, at modellens rolle er enten at udvælge, hvilke borgere, der skal oprettes sager om, eller at prioritere eller at oplyse allerede eksisterende sager, hvor der i sidste ende skal træffes afgørelser af mennesker.

En profileringsmodel kan bruges til udvælgelse, som når Udbetaling Danmark anvender modeller til at udarbejde en "undringsliste" med de personer, som statistisk set har høj risiko for uberettiget at modtage en ydelse. Denne "undringsliste" sendes til kommunernes kontrolafdelinger, der beslutter, om der skal oprettes en sag på grundlag af de modtagne oplysninger.

En profileringsmodel kan bruges til at prioritere sager som i Horsens kommune forsøgsvise arbejde med en model til prioritering af kontrolsager om social svindel. Kommunens model skulle angive, om der i eksisterende sager var høj eller lav sandsynlighed for social svindel. På baggrund af modellens vurdering ville sagsbehandlere i kommunens kontrolafdeling kunne vælge at afsætte mere eller mindre tid til at vurdere sagen.

En profileringsmodel kan bruges til at oplyse sagsbehandlingen, når den giver sagsbehandleren adgang til ny information om en sag, som kan have indflydelse på, hvordan sagen bør behandles. I disse tilfælde er formålet med modellens brug at kvalificere sagsbehandlingen ved at skabe et bedre beslutningsgrundlag. Den profileringsmodel, der bruges af Styrelsen for Arbejdsmarked og Rekruttering for at vurdere risiko for langtidsledighed, er et eksempel på dette.

Se bilag 2 for udførlige beskrivelser af myndigheders brug af profileringsmodeller.

I de tilfælde, hvor en myndighed planlægger at bruge en profileringsmodel til beslutningsstøtte, er det vores vurdering, at legalitetsprincippet værner tilstrækkeligt om retssikkerheden (og lovgivers involvering), og at der derfor ikke er behov for lovhjemmel for at bruge modellen.

Derimod er der både for fuldautomatiserede afgørelser og for brugen af modellerne til beslutningsstøtte behov for styringsværktøjer, når det gælder spørgsmålet om, ikke hvorvidt, men hvordan modellen kan bruges af myndigheden. Disse styringsværktøjer udgør selve kernen af den rettighedsbaserede tilgang, som vi præsenterer i denne rapport. I resten af dette kapitel kommer vi med vores bud på sådanne styringsværktøjer i form af en vejledning for ensartet brug, AI-konsekvensanalyser, tilsyn og styrket teknologisk og rettighedsmæssig forståelse. De efterfølgende kapitler vil udfolde styringsværktøjernes formål og indhold yderligere.

VEJLEDNING FOR ENSARTET BRUG

Med ombudsmandens ord gælder "de almindelige regler, som myndigheden skal overholde" også, "når computer afløser papir".[7] Lovgivningen er med andre ord teknologineutral.

Princippet om teknologineutralitet kan virke relativt ligetil, men er det ikke nødvendigvis. Det er blevet fremhævet, at et princip om teknologineutralitet kan føre til uensartet praksis og manglende forudsigelighed om, hvordan digitalisering gennemføres hos de enkelte myndigheder. Samtidig skaber det også et stort ansvar for den enkelte myndighed, som f.eks. beslutter at anvende profileringsmodeller i sagsbehandlingen, da myndigheden selv skal fortolke og vurdere, hvordan den vil sikre overholdelsen af regler og principper inden for en teknisk kompleks ramme. Myndigheden bærer ansvaret for, at der er truffet korrekte valg i udformningen af modellen og at rettigheder og principper er fortolket korrekt, uden at få vejledning om spørgsmålet.[8]

I fraværet af udtrykkelige regler om hvornår og hvordan myndigheden må bruge bl.a. profileringsmodeller, kan beslutningen om at tage en model i brug efter vores vurdering risikere at blive ressourcemæssigt tung for den enkelte forvaltning. Myndigheden bør derfor være opmærksom på, at det øgede ansvar for komplekse retssikkerhedsmæssige spørgsmål kan øge ressourceforbrug og omkostninger.

Uden klar vejledning om, hvordan profileringsmodeller kan og ikke kan indgå i sagsbehandlingen på en retssikkerhedsmæssigt forsvarlig måde, stiger risikoen desuden for, at myndighederne begår fejl eller træffer forkerte valg.

Det er derfor vores vurdering, at der er behov for at ensrette myndighedernes praksis om brugen af profileringsmodeller. En vejledning vil efter vores vurdering sikre ensartethed, transparens og forudsigelighed i myndighedernes brug af modellerne.

Vi anbefaler, at Justitsministeriet med inddragelse af Datatilsynet og Digitaliseringsstyrelsen udsteder en vejledning om offentlige myndigheders brug af profileringsmodeller med fokus på de rettigheds- og retssikkerhedsmæssige udfordringer ved modellerne.

Anbefalingen om at udstede en vejledning fungerer som en overordnet anbefaling for mange af rapportens øvrige anbefalinger, der følger i resten af dette og de følgende kapitler. Når vi i de følgende anbefalinger henviser til Justitsministeriets vejledning, er det således en henvisning til denne anbefaling.

KONSEKVENSANALYSER SOM STYRINGSVÆRKTØJ

Et af de vigtigste styringsværktøjer for at sikre en rettighedsbaseret tilgang til profileringsmodeller er såkaldte konsekvensanalyser. Vi anbefaler, at der i vejledningen stilles krav om AI-konsekvensanalyser, som myndigheden skal gennemføre tidligt i beslutningsfasen og periodisk herefter. Vores anbefaling herom tager udgangspunkt i eksisterende tiltag og regler på området, som vi giver et kort oprids af i det følgende.

EKSISTERENDE TILTAG OG REGLER

Konsekvensanalyser er ikke et entydigt begreb og kan forstås på mange forskellige måder. I kontekst af rettigheder og profileringsmodeller er der over de sidste mange år blevet fremsat en række forskellige bud på en strukturel kontrol i form af konsekvensanalyser.

Algoritmiske konsekvensanalyser (algorithmic impact assessment) er et felt i hurtig udvikling inden for litteraturen, hvor der med afsæt i forskellige sociotekniske, juridiske og etiske fagområder beskrives særlige kriterier, som en konsekvensanalyse bør indeholde, for at kunne adressere "algoritmiske" udfordringer.[9] Kategorien rummer mange forskellige forslag, herunder både nogle, der omfatter alle former for kunstig intelligens, og andre, der er målrettet bestemte former for teknologi, ligesom nogle er målrettet alle aktører, der anvender teknologien og andre specifikt vedrører enten private aktører eller offentlige myndigheder.

Menneskeretlige konsekvensanalyser (human rights impact assessments) er en af de metoder, der bruges som led i private virksomheders iagttagelse af "nødvendig omhu" (due diligence) for at sikre respekt for menneskerettigheder, med afsæt i FN's vejledende principper for menneskerettigheder og erhvervsliv.[10] Til forskel fra algoritmiske konsekvensanalyser er fokus her på menneskeretlige forpligtelser og på at beskytte og forbedre borgernes rettigheder, men sondringen er ikke skarp og visse eksempler omhandler både dataetiske overvejelser og menneskeretlige forpligtelser.[11] Menneskeretlige konsekvensanalyser kan hjælpe med at identificere og vurdere omfanget af udfordringer, og Institut for Menneskerettigheder har udarbejdet en detaljeret værktøjskasse for private virksomheders menneskeretlige

konsekvensanalyser med et specifikt fokus på digitale aktiviteter.[12] Da det sjældent vil være offentlige myndigheder selv, der udvikler profileringsmodellen, bør myndigheder stille krav til de private virksomheder, der udvikler profileringsmodeller.

Nogle af de mest omfattende strukturelle krav til brugen af kunstig intelligens forventes at blive fastsat i EU's kommende forordning om kunstig intelligens (se kapitel 2). EU-Kommissionen har fremsat et udkast til forordningen i foråret 2021, som indeholder en række forskellige styringsværktøjer (se boks).

KOMMENDE EU-FORORDNING OM KUNSTIG INTELLIGENS: KRAV TIL GOVERNANCE

Risikostyringssystemer: Ifølge udkastet skal der bl.a. udarbejdes risikostyringssystemer[13], hvor den ansvarlig myndighed skal identificere, analysere og evaluere de risici, som kan opstå ved anvendelsen af et AI-system (som f.eks. en profileringsmodel). Myndigheden skal fastsætte en hensigtsmæssig risikostyringsforanstaltning ved bl.a. at fjerne eller begrænse risici i udviklingen af modellen og gennemføre foranstaltninger for at kontrollere de risici, som ikke kan fjernes og sikre gennemsigtighed og tilstrækkelige oplysninger herom. Om nødvendigt skal myndigheden desuden uddanne de sagsbehandlere, der må forventes at bruge systemet.

Kvalitetsstyringssystemer: Myndigheden er endvidere forpligtet til at udarbejde kvalitetsstyringssystemer[14], som fastsætter detaljerede procedurer for mange af udkastets krav.

Overensstemmelsesvurdering: Myndigheder skal også gennemføre en intern "overensstemmelsesvurdering".[15] Heri skal det sikres, at der er etableret et kvalitetsstyringssystem, og at den fornødne tekniske dokumentation er til stede for at kunne vurdere, om AI-systemet er i overensstemmelse med udkastets regler om bl.a. datakvalitet, teknisk dokumentation og registrering, gennemsigtighed og formidling af oplysninger, menneskeligt tilsyn og krav til systemets robusthed, nøjagtighed og cybersikkerhed. Det skal også sikres, at der er overensstemmelse mellem den tekniske dokumentation og design og udvikling samt efterfølgende monitorering af modellen.

Det er også i den forvaltningsretlige litteratur blevet fremhævet, at der i udviklingen af profileringsmodellen skal gennemføres undersøgelser om overholdelsen af rettigheder og retssikkerhed, udledt af grundlæggende forvaltningsretlige principper og ombudsmandspraksis.[16] Det er en central forudsætning inden for forvaltningsretten, at myndighederne overholder lovgivningen og indretter deres organisation og arbejdsgange for at sikre dette. Forudsætningen er dels baseret på legalitetsprincippet (se ovenfor). Dels er forudsætningen baseret på kravet om god

forvaltningsskik, hvorefter myndigheder skal indrette deres organisation forsvarligt og have effektive arbejdsgange, der understøtter lovlig og korrekt forvaltning.[17]

Ombudsmanden har i en central udtalelse fra 2014 forholdt sig til overholdelsen af forvaltningsretlige krav under udviklingen af IT-systemer. I udtalelsen hedder det bl.a.:

"En forsvarlig tilrettelæggelse forudsætter [...] bl.a., at man fra starten skaber sig et overblik over de sagstyper og processer, som det nye IT-system skal omfatte, at man gør sig klart, hvilke formelle og materielle regler der gælder for behandlingen af de pågældende sager, og at man er meget omhyggelig med at tage stilling til, hvorledes det nye IT-system konkret skal udformes for at kunne overholde disse regler i de forskellige forløb, som sagerne kan tænkes at ville få."[18]

De væsentligste regler om konsekvensanalyser findes i EU's databeskyttelses-forordning, som forpligter myndigheden til at foretage "konsekvensanalyser vedrørende databeskyttelse".[19]

Disse konsekvensanalyser skal som minimum indeholde en systematisk beskrivelse af den planlagte behandling af personoplysninger og formålene med behandlingen, herunder de legitime interesser, der forfølges; en vurdering af, om behandlingen af personoplysninger er nødvendige og står i rimeligt forhold til formålene; en vurdering af risiciene for borgernes rettigheder og frihedsrettigheder, og endelig de foranstaltninger, som myndigheden påtænker for at imødegå disse risici. Til de sidste hører garantier, sikkerhedsforanstaltninger og mekanismer, som kan sikre beskyttelse af personoplysninger og påvise overholdelse af forordningen, under hensyntagen til borgeres rettigheder og legitime interesser.[20]

Efter databeskyttelsesforordningen skal der gennemføres konsekvensanalyser ved brug af nye teknologier, der indebærer "høj risiko" for personers rettigheder og frihedsrettigheder. Kravet om konsekvensanalyser vil derfor efter vores vurdering blive udløst ved brugen af profileringsmodeller, både til fuldautomatiserede afgørelser og til automatiseret beslutningsstøtte. Konsekvensanalysen skal gennemføres før behandlingen af personoplysninger, men er i øvrigt en løbende proces, og ajourføring af analysen gennem hele værktøjets levetid er derfor et krav.[21]

Konsekvensanalyserne foretages som led i behandlingen af personoplysninger, men skal vedrøre alle rettigheder og frihedsrettigheder sikret i EU-retten. Myndigheder kan være undtaget fra kravet om konsekvensanalyser, hvis en generel analyse er blevet gennemført i forbindelse med udformningen af et lovforslag eller eventuelt en bekendtgørelse.[22]

Hvis konsekvensanalysen viser, at behandlingen vil føre til en høj risiko, og denne ikke kan begrænses ved passende foranstaltninger, er det et krav, at Datatilsynet høres inden behandlingen påbegyndes.[23] Desuden skal Datatilsynet høres som

led i udarbejdelsen af lovforslag mv., som skaber hjemmelsgrundlaget for en myndigheds højrisikobehandling af oplysninger.[24]

Databeskyttelsesforordningens regler om konsekvensanalyser skal ses i lyset af, at forordningen bygger på et paradigme der indebærer, at myndighederne skal kende risikoen forbundet med behandlingen af personoplysninger. Dette indebærer, at myndighederne, også når kravet om konsekvensanalyser ikke udløses – og ved siden af de krav, der følger af konsekvensanalysen – skal kende den risiko, som et nyt tiltag fører til.[25]

Som vi har fremhævet i kapitel 2 udgør databeskyttelsesforordningens regler, herunder reglerne om obligatoriske konsekvensanalyser, et vigtigt grundlag for beskyttelsen af rettigheder, hvor myndighederne bruger profileringsmodeller.[26]

Der er imidlertid visse begrænsninger i databeskyttelsesforordningens konsekvensanalyser: for det første stilles der ikke særlige krav til indholdet af analyserne, når myndighederne bruger en profileringsmodel, og forordningen adresserer derfor heller ikke modellernes særegne udfordringer. For det andet er der ikke krav om, at konsekvensanalyserne eller deres væsentligste indhold offentliggøres. For det tredje er de forvaltningsretlige regler og principper, som vi har beskrevet i kapitel 2 ikke – eller kun delvist – omfattet af forordningens beskyttelse af "rettigheder og frihedsrettigheder" efter EU-retten. For det fjerde kan konsekvensanalyserne i praksis komme til at fokusere alene på databeskyttelse selvom de skal omfatte alle rettigheder omfattet af EU-retten.[27]

RETTIGHEDSBASEREDE AI-KONSEKVENSANALYSER

De forskellige typer af konsekvensanalyser, som vi har beskrevet her, omfatter hver især og i kombination med hinanden mange nødvendige aspekter for at sikre en rettighedsbaseret tilgang til profileringsmodeller.

Det er efter vores vurdering problematisk, at der ikke er udtrykkelige regler om, at offentlige myndigheder skal gennemføre særlige rettighedsbaserede konsekvensanalyser forud for beslutningen om at udvikle en profileringsmodel og periodisk herefter under anvendelsen af modellen. Selvom forskellige af ovennævnte tiltag eller regler dækker aspekter af sådanne konsekvensanalyser, er det retssikkerhedsmæssigt problematisk, at der ikke findes et ensartet styringsværktøj for myndighederne, som eksplicit forholder sig til udfordringerne ved profileringsmodeller. Vi anbefaler derfor, at der vedtages regler om, hvad vi i rapporten benævner rettighedsbaserede AI-konsekvensanalyser.

En rettighedsbaseret AI-konsekvensanalyse bør efter vores vurdering især bygge videre på gældende regler i databeskyttelsesforordningen og kan i øvrigt suppleres af fremtidige regler som f.eks. den kommende EU-forordning om kunstig intelligens. Ligesom konsekvensanalyserne i databeskyttelsesforordningen[28], bør AI-konsekvensanalyserne udgøre en proces, der bidrager til både at skabe og påvise beskyttelse af rettigheder og retssikkerhed.

Vi anbefaler desuden, at konsekvensanalyserne eller deres væsentligste indhold offentliggøres. Vi kommer nærmere ind på behovet for en offentlig tilgængelig oversigt over offentlige myndigheders brug af profileringsmodeller i kapitel 7.

Vi anbefaler, at der i vejledningen stilles krav om konsekvensanalyser for kunstig intelligens (AI-konsekvensanalyser) tidligt i beslutningsfasen og periodisk herefter under modellens udvikling og anvendelse til både beslutningsstøtte og fuldautomatisering.
Konsekvensanalyserne bør som minimum omfatte vurdering af:

- **Beslutning: Hjemmel og begrundelse for anvendelse af modellen, herunder om modellen anvendes til beslutningsstøtte eller fuldautomatiseret afgørelse, modellens egnethed og oplysninger om eventuelle alternative modeller eller løsninger.**
- **Risici: Om modellen rejser risici for usaglige og forkerte (ulovlige) afgørelser samt øvrige utilsigtede konsekvenser for borgerens rettigheder, om modellen overholder databeskyttelsesforordningens krav om dataminimering og sikrer brug af korrekte data, hvilke diskriminationsrisici modellen rejser, samt hvorledes samtlige risici vil blive imødegået**
- **Transparens: Om modellen lever op til krav om algoritmisk transparens.**
- **Tilsyn og kontrol: Hvordan tilsyn og kontrol med modellen sikres.**

I lighed med anbefalingen om en vejledning for ensartet brug, er også anbefalingen om AI-konsekvensanalyser en overordnet og tværgående anbefaling som vi udbygger og begrunder i rapportens øvrige kapitler.

TILSYN

Det er en væsentlig retsgaranti, at myndighedernes sagsbehandling og afgørelser underlægges kontrol i form af tilsyn og prøvelse ved uafhængige instanser. Denne retsgaranti omfatter naturligvis også myndighedernes brug af profileringsmodeller. Tilsyn og kontrol med modellen er med til at sikre beskyttelse af borgerens rettigheder gennem hele modellens livscyklus.

For at sikre, at profileringsmodeller ikke fører til en forringelse af borgernes rettigheder og retssikkerhed, ser vi i dette afsnit nærmere på, hvordan forudgående og løbende generelt tilsyn bedst gennemføres på området.

Myndigheders sagsbehandling og afgørelser kan efter gældende regler efterprøves enten ved rekurs eller ved domstolsprøvelse. Administrativ rekurs er en adgang for borgeren til at klage over en konkret afgørelse til en anden forvaltningsmyndighed. F.eks. er Ankestyrelsen rekursinstans for kommunerne på det sociale område. Uanset, om der er adgang til administrativ rekurs eller ej, vil forvaltningens sagsbehandling og afgørelser desuden kunne prøves af domstolene efter grundlovens § 63.

Herudover fører Folketingets Ombudsmand kontrol med myndighedernes overholdelse af forvaltningsretten og kan rejse kritik eller komme med anbefalinger til myndighederne. Dette kan både omfatte forhold, som vedrører den enkelte afgørelse (uden at ombudsmanden dog kan træffe bindende afgørelser) eller mere generelle forhold, som f.eks. myndighedernes sagsgange.

Foruden disse institutioner, har Datatilsynet og Ligebehandlingsnævnet særlige beføjelser for overholdelsen af henholdsvis databeskyttelsesretten og lovgivning om ligebehandling, ligestilling og ikke-diskrimination, og kan træffe afgørelser inden for deres tilsynsområder.

Der findes således mange forskellige tilsyns- og kontrolorganer, som regulerer forskellige områder i myndighedernes sagsbehandling.

Der er imidlertid på nuværende tidspunkt ingen regler om et tilsyn med myndighedernes brug af profileringsmodeller eller anden kunstig intelligens.

Et forudgående og løbende tilsyn udgør en uafhængig kontrol med udformning af profileringsmodellen og dens anvendelse uden, at der nødvendigvis rejses en konkret sag baseret på modellens afgørelse. Det kan være med til at sikre, at rettigheder er tænkt ind under udviklingsfasen, hvor der skal træffes mange og vigtige valg om modellens udformning. Et forudgående og løbende tilsyn bør efter vores vurdering både dække de helt tidlige stadier i modellens livscyklus (dvs. beslutnings- og udviklingsfasen), hvor der foretages vigtige designvalg og testning, og modellens efterfølgende brug i sagsbehandlingen (dvs. anvendelsesfasen). Der er vigtigt, at tilsynet er løbende og foretages på ny ved forskellige stadier i modellens levetid, herunder for at se, hvordan modellen fungerer ude i "den virkelige verden".

Såvel internationalt og på europæisk plan er der kommet mange bud på, hvordan et effektivt tilsyn med myndigheders brug af profileringsmodeller kan og bør se ud.

Det er blevet fremhævet, at der er behov for et nyt organ, der foretager "algoritmisk tilsyn" eller "audit" med særlig fagkundskab og særlige beføjelser. Dette er også blevet fremhævet som nødvendigt for at sikre, at konsekvensanalyser, som f.eks. vores AI-konsekvensanalyser, underlægges kontrol.

Bl.a. har Europarådet anbefalet, at staterne fastsætter bindende regler om et uafhængigt tilsyn med kunstig intelligens, som dækker både udvikling og anvendelse hos såvel offentlige som private aktører. Ifølge Europarådet kan sådanne bindende regler være med til at styrke eller etablere et samarbejde mellem administrative, judicielle og parlamentariske tilsynsmekanismer og det bør derfor sikres, at institutionerne udstyres med nødvendig ekspertise, kompetencer og ressourcer. Tilsynet bør efter Europarådets vurdering være uafhængigt fra de, som det fører tilsyn med og bør regelmæssigt rapportere til staternes parlamenter.[29]

Det er også blevet foreslået, at et nyt tilsyn bør etableres med inspiration i databeskyttelsesforordningens regler om nationale datatilsyn[30], og stille krav om forudgående rapportering og godkendelse af profileringsmodeller, før de tages i brug. Den såkaldte Artikel 29-gruppe (se kapitel 2) har foreslået, at der for fuldautomatiserede afgørelser bør være en "best practice" om algoritmisk audit, dvs. testning af modellen for at påvise, at den rent faktisk fungerer efter hensigten og ikke giver diskriminerende eller fejlagtige resultater. I den forbindelse bør der også ifølge gruppen sikres en uafhængig tredjemandsaudit, hvis profileringen har stor indvirkning på borgeren.[31]

Endnu andre har foreslået tilsyn, der kan opbygges efter samme krav, som gælder i menneskeretten for staternes forpligtelse til at sikre tilsyn med efterretningstjenesterne.[32] Det er også blevet foreslået, at man enten opretter eller udvider eksisterende ombudsmandsinstitutioner.[33]

Mere konkret er der i EU-Kommissionens udspil til en forordning for kunstig intelligens fastsat regler om tilsynsmyndigheder i medlemsstaterne, som enten skal oprettes, eller hvor eksisterende myndigheder skal varetage tilsynet. Der skal være tale om en tilsynsmyndighed, som skal varetage tilsynet bredt for alle "AI-systemer", der er omfattet af udkastet og som udbydes såvel af private aktører som offentlige myndigheder. Den nationale tilsynsmyndighed skal agere uafhængigt og varetage en lang række opgaver fastsat i udkastet.[34] Tilsynsmyndigheden skal besidde ekspertise inden for bl.a. AI-teknologier, data og databehandling, grundlæggende rettigheder, sundheds- og sikkerhedsrisici og viden om gældende standarder og retlige krav.

Kommissionens udspil vil ikke omfatte tilsyn med myndighedens overholdelse af forvaltningsretlige regler og principper, og i udkastet er der heller ingen direkte klageadgang eller anden adgang til håndhævelse af rettigheder for den enkelte borger. Selvom udkastet opstiller regler for tilsynsmyndigheder på tværs af EU, lægger det således ikke op til, at borgeren skal have ret til at få behandlet en klage hos dem. Der er heller ikke lagt op til regler, der skal sikre, at borgerne kan bistås med overblik over eller vejledning om, hvordan teknologien indvirker på deres rettigheder. For offentlige myndigheders brug af profileringsmodeller til at træffe afgørelser gælder desuden, at de underlægges et mindre skærpet tilsyn.[35]

Det Europæiske Databeskyttelsesråd og Den Europæiske Tilsynsførende for Databeskyttelse har anbefalet, at kommende regler i forordningen om kunstig intelligens i højere grad skal tænkes sammen med eksisterende regler i databeskyttelsesforordningen, samt at de nationale datatilsyn bør varetage kommende tilsynsopgaver i den nye forordning.[36] Dette kan imidlertid efter vores vurdering hurtigt komme til at indebære, at Datatilsynet også skal føre tilsyn med øvrige rettigheder end beskyttelsen af personoplysninger på et komplekst og hastigt voksende område.

Et effektivt og velfungerende tilsyn med offentlige myndigheders brug af profileringsmodeller kræver efter vores vurdering et nært samarbejde mellem eksisterende tilsynsinstanser. Når vi i denne rapport anbefaler en rettighedsbaseret tilgang til profileringsmodeller, indebærer det, at samtlige berørte rettigheder skal beskyttes. Det vil kunne indebære øget samarbejde mellem bl.a. Datatilsynet og Ligebehandlingsnævnet, og i nogle tilfælde Ankenævnet, som også behandler spørgsmål inden for ligebehandling.

I forhold til det mere generelle og overordnede tilsyn har især ombudsmanden en central rolle i at sikre rettigheder og retssikkerhed for borgerne, når danske myndigheder anvender profileringsmodeller. I modsætning til f.eks. domstolene kan ombudsmanden udstikke generelle retningslinjer for sagsbehandlingen. Ombudsmanden har også selv fremhævet området som vigtigt[37] og det er vores vurdering, at der – med behørige ressourcer og teknisk ekspertise – vil være klare fordele ved at lade den eksisterende ombudsmandsinstitution stå for dele af et generelt og løbende tilsyn med myndighedernes brug af profileringsmodeller.

Der er også mulighed for at etablere indrapporteringsforpligtelser til diverse rekursinstanser, som inden for hvert deres område vurderer, om profileringsmodellen skaber udfordringer. Som vi vil komme ind på senere i rapporten, vil der være vigtige designvalg, der skal træffes for at sikre mod f.eks. usaglighed eller diskrimination (se kapitel 6 og 8). Disse valg kan den enkelte tilsynsmyndighed, der har kendskab til sagsområdet, være bedre egnet til at vurdere end en generel "algoritmisk tilsynsmyndighed". Inden for kommunernes varetagelse af det sociale område kunne en sådan indrapportering f.eks. ske til Ankestyrelsen, før kommunerne gør brug af profileringsmodeller, og hvor Ankestyrelsen fokuserer på de risici for forkerte og ulovlige afgørelser, som sædvanligvis eksisterer på området. Også her bør en udvidelse af tilsynsopgaven naturligvis indebære, at der samtidig sikres fornøden ekspertise.

Samlet set bør der efter vores vurdering sikres et tilsyn i Danmark med offentlige myndigheders brug af profileringsmodeller. Vi anbefaler, at især Europarådets betragtninger er styrende for dette, men at nationale regler også med fordel kan tænkes sammen med kommende EU-regulering på området. Hvad angår sidstnævnte, vurderer vi dog, at regler herhjemme til forskel fra det nuværende EU-udkast bør sikre en direkte klageadgang for borgere.

Efter vores vurdering er det af afgørende betydning, at der hos eksisterende tilsynsinstanser er indgående kendskab til de retssikkerhedsmæssige udfordringer ved profileringsmodeller. Det gælder ikke mindst hos de tilsynsmyndigheder, hvor borgeren har en direkte klageadgang. En øget brug af profileringsmodeller kræver, at samtlige tilsynsinstanser er i stand til at føre et effektivt tilsyn, og at ekspertise og viden om profileringsmodellernes muligheder, begrænsninger og retssikkerhedsmæssige udfordringer ikke centraliseres ét sted.

Vi anbefaler, at Justitsministeriet med inddragelse af Datatilsynet, Digitaliseringsstyrelsen og Folketingets Ombudsmand sikrer et styrket samarbejde på tværs af de tilsyns-, rekurs- og kontrolinstanser, der forventes at kontrollere myndighedernes brug af profileringsmodeller.

at der fastsættes regler om, at AI-konsekvensanalyserne indrapporteres til en relevant myndighed, herunder at valget af myndighed sker under hensyntagen til EU's kommende forordning om kunstig intelligens.

STYRKET TEKNOLOGISK OG RETTIGHEDSMÆSSIG FORSTÅELSE

Som vi har redegjort for i del I af rapporten, er både det retlige grundlag og teknologien bag profileringsmodeller relativt kompleks. Hvis der skal etableres en rettighedsbaseret tilgang til anvendelsen af modellerne, er det derfor helt nødvendigt, at både beslutningstagere, retsanvendere og tilsyns- og kontrolinstanser har den fornødne tekniske og rettighedsmæssige forståelse.

Det er en central pointe, at man ikke kan sikre borgenes rettigheder uden fornødent kendskab til teknologien og dens styrker og udfordringer i forhold til rettighederne. Det gælder ikke mindst for de konkrete styringsværktøjer, vi har foreslået i løbet af dette kapitel. Det er således en forudsætning for at opfylde anbefalingerne i denne rapport, at myndigheden har kendskab til teknologien og de særegne udfordringer, den rejser. Tilsvarende indebærer et effektivt tilsyn og adgang til efterprøvelse naturligvis kendskab til de tekniske indretninger og valg i modellen og de konsekvenser, de har.

Et sted, hvor kendskab til og forståelse for teknologi og rettigheder med fordel kan tilføjes er i Digitaliseringsstyrelsens "Statens Digitaliseringsakademi", der har til formål at "styrke statslige ledere og medarbejderes digitale kompetencer til at forstå, ville og kunne digitalisering".[38]

Vi anbefaler, at behovet for teknisk forståelse for profileringsmodeller og for de rettigheds- og retssikkerhedsmæssige udfordringer, som de rejser hos beslutningstagere og anvendere af profileringsmodeller samt hos tilsyns- og kontrolinstanser adresseres i vejledningen.

at Digitaliseringsstyrelsen inkluderer teknisk kendskab til profileringsmodeller og kendskab til de retlige udfordringer forbundet med profileringsmodeller i "Statens Digitaliseringsakademi".

NOTER

1 Motzfeldt og Waage (2021). Digitale retsstater – pilotprojekt for Nordisk Ministerråd, s. 16, tilgængelig på: https://pub.norden.org/temanord2021-501/#

2 DataEthics (April 2020), White Paper on Data Ethics in Public Procurement of AI-based Services and Solutions tilgængeligt på: https://dataethics.eu/wp-content/uploads/dataethics-whitepaper-april-2020.pdf

3 Institut for Menneskerettigheder (2021), 'Smart Mix' In the Nordics - A Stocktake on Measures to Foster Business Respect for Human Rights. tilgængeligt på: https://www.humanrights.dk/sites/humanrights.dk/files/media/document/Smart Mix in the Nordics_2021.pdf

4 Databeskyttelsesforordningen 2016/679 af 27. april 2016, artikel 22

5 Se på den ene side Korfitz Nielsen og Lotterup (2020) Databeskyttelsesforordningen og databeskyttelsesloven med kommentarer, Djøf forlag, s. 58 og Justitsministeriets betænkning 1565/2017 om Databeskyttelsesforordningen (2016/679) – og de retlige rammer for dansk lovgivnings. 380 tilgængelig på https://www.justitsministeriet.dk/sites/default/files/media/Pressemeddelelser/pdf/2017/bet_1_1.pdf og på den anden side Databeskyttelsesforordningen 2016/679 af 27. april 2016, betragtning 71 og Artikel 29-Gruppens retningslinjer om automatiske individuelle afgørelser og profilering i henhold til forordning 2016/679 (2017/2018) s. 24

6 Motzfeldt, Ullitis og Kjellerup (2020), Fra forvaltningsjurist til udviklingsjurist – introduktion til offentlig digitalisering, Djøf forlag, s. 57 med henvisning til praksis

7 Se ombudsmandens notat (2. juli 2018), forvaltningsretlige krav til det offentliges IT-løsninger, tilgængeligt på: https://www.ombudsmanden.dk/myndighedsguiden/specifikke_sagsomraader/generelle_forvaltningsretlige_krav_til_offentlige_it-systemer/2019/

8 Regeringskansleit (27. marts 2018), "Juridik som stöd för förvaltningens digitalisering" (SOU 2018:25), s. 125f tilgængeligt på (svensk): https://www.regeringen.se/rattsliga-dokument/statens-offentliga-utredningar/2018/03/sou-201825/

9 Se bl.a. Metcalf et al (2021), Algorithmic Impact Assessment and Accountability: The Co-construction of Impacts, Mantelero (August 2018), 'AI and Big Data: A blueprint for a human rights, social and ethical impact assessment', Computer Law & Security Review Vol. 34, Issue 4, s. 754-772og Reisman et al (2018), Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability

10 Global Compact Network Denmark (2020), FN's vejledende principper for menneskerettigheder og erhvervsliv, tilgængelige på: https://globalcompact.dk/wp-content/uploads/2021/06/FNs-Vejledende-Principper-for-Menneskerettigheder-og-Erhvervsliv.pdf
11 Se EU's Agentur for Grundlæggende Rettigheder (FRA), (2020) Getting the Future Right: Artificial Intelligence and Fundamental Rights, s. 87ff. Mantelero (August 2018), 'AI and Big Data: A blueprint for a human rights, social and ethical impact assessment', Computer Law & Security Review Vol. 34, Issue 4, s. 754-772
12 Institut for Menneskerettigheder (2020), Human rights impact assessment of digital activities tilgængeligt på:
Human rights impact assessment of digital activities | The Danish Institute for Human Rights
13 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final, artikel 9
14 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 artikel 17
15 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 artikel 43 og bilag VI
16 Motzfeldt (2017), "'The Danish Principle of Administrative Law by Design", European Public Law, Issue 4, side 752
17 Se f.eks. ombudsmandens udtalelser i FOB 1992.232, FOB 2006.165 og FOB 2008.380. Se også Fenger (2018), Forvaltningsret, Jurist- og Økonomforbundets Forlag, s. 692.
18 Ombudsmandens udtalelse er tilgængelig på: https://www.ombudsmanden.dk/find/udtalelser/beretningssager/alle_bsager/2014-24/pdf2,se også Digital Forvaltning s. 82ff med gennemgang af sagen.
19 Databeskyttelsesforordningen 2016/679 af 27. april 2016, artikel 35
20 Databeskyttelsesforordningen 2016/679 af 27. april 2016 artikel 35, stk. 7, se også betragtning nr. 84 og 90
21 Artikel 29-gruppens retningslinjer for konsekvensanalyse vedrørende databeskyttelse (DPIA) og bestemmelse af, om behandlingen "sandsynligvis indebærer en høj risiko" i henhold til forordning (EU) 2016/679, s. 16-17
22 Databeskyttelsesforordningen 2016/679 af 27. april 2016, artikel 35, stk. 10.
23 Databeskyttelsesforordningen 2016/679 af 27. april 2016, artikel 36, stk. 1
24 Databeskyttelsesforordningen 2016/679 af 27. april 2016, artikel 36, stk. 4
25 Kofod Olsen (2020), Håndbog i Databeskyttelsesret, kapitel 13. Se også kapitel 14 om konsekvensanalyser.
26 Kaminski og Malgieri (2020), Algorithmic impact assessments under the GDPR: producing multi-layered explanations, International Data Privacy Law, Volume 11, Issue 2 tilgængeligt på: https://doi.org/10.1093/idpl/ipaa020
27 Se FRA 2020 rapport, Artikel 29- i sine retningslinjer om forordningens konsekvensanalyser og Mantelero, 'AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment' (2018) 34(4) Computer Law & Security Review 766

28 Artikel 29-gruppens retningslinjer for konsekvensanalyse vedrørende databeskyttelse (DPIA) og bestemmelse af, om behandlingen "sandsynligvis indebærer en høj risiko" i henhold til forordning (EU) 2016/679, s. 4
29 Se Det Europæiske Råds Kommissariat for Menneskerettigheder (2019), Unboxing Artificial Intelligence: 10 steps to protect Human Rights tilgængeligt på: https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64
30 Databeskyttelsesforordningen 2016/679 af 27. april 2016artikel 36
31 Artikel 29-Gruppens retningslinjer om automatiske individuelle afgørelser og profilering i henhold til forordning 2016/679 (2017/2018)
32 Det Europæiske Råds ad hoc Kommitté for Kunstig Intelligens (2020), Feasibility Study tilgængelig på: https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da og McGregor, Murray og Ng (2019), International Human Rights Law as a Framework for Algorithmic Accountability tilgængelig på: http://repository.essex.ac.uk/24505/1/div-class-title-international-human-rights-law-as-a-framework-for-algorithmic-accountability-div.pdf
33 Se f.eks. McGregor, Murray og Ng (2019), International Human Rights Law as a Framework for Algorithmic Accountability [tilgængeligt på: http://repository.essex.ac.uk/24505/1/div-class-title-international-human-rights-law-as-a-framework-for-algorithmic-accountability-div.pdf] og Miller, Ohrvik-Stott, Coldicutt (2018), Regulating for Responsible Technology: Capacity, Evidence and Redress: a new system for a fairer future, Doteveryone [tilgængeligt på: https://doteveryone.org.uk/wp-content/uploads/2018/10/Doteveryone-Regulating-for-Responsible-Tech-Report.pdf].
34 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206, artikel 59
35 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206, artikel 43, stk. 2 og "procedurer for overensstemmelsesvurdering baseret på intern kontrol" jf. bilag VI
36 Det Europæiske Databeskyttelsesråd (21. juni 2021), EDPB & EDPS call for ban on use of AI for automated recognition of human features in publicly accessible spaces, and some other uses of AI that can lead to unfair discrimination tilgængelig på: https://edpb.europa.eu/news/news/2021/edpb-edps-call-ban-use-ai-automated-recognition-human-features-publicly-accessible_en
37 Se f.eks. ombudsmandens udtalelse i FOB 2019 Hvordan digitaliserer vi uden at skade vores retssikkerhed? tilgængelig på: https://www.ombudsmanden.dk/findviden/fob-artikler/hvordan_digitaliserer_vi/ og Fenger (2020), Ombudsmanden – et værn for borgernes retssikkerhed, UfR 2020B37
38 Se nærmere om digitaliseringsakademiet på: https://digst.dk/styring/statens-digitaliseringsakademi/om-akademiet/

KAPITEL 5

MODELLENS INPUT: NØDVENDIGE OG KORREKTE DATA

I dette kapitel ser vi på kravene til modellens input, dvs. de data som modellen er trænet på.

En profileringsmodel er bygget på og anvender data. Jo flere data en model trænes på, jo højere bliver modellens kvalitet typisk. Det skyldes to forhold:

For det første forbedres træningen af en model normalt, jo flere personer der optræder i **datasættet.** Når datasættet indeholder flere personer, så vil de relevante sammenhænge mellem **variable** og **målegenskab** typisk være tydeligere, samtidig med at indflydelsen fra tilfældige og ikke-**generaliserbare** sammenhænge bliver mindre.

For det andet øges kvaliteten, jo flere variable – dvs. oplysninger om borgerens forhold – modellen kan anvende, fordi modellen derved bliver i stand til at foretage mere komplekse og nuancerede vurderinger. Dog er dette kun tilfældet under den væsentlige forudsætning, at antallet af variable står i et relevant forhold til antallet af personer i datasættet. Hvis ikke det er tilfældet, risikerer man, at modellen finder tilfældige eller irrelevante sammenhænge i data.

Denne sammenhæng mellem datamængde og kvalitet kan skabe en stærk tilskyndelse til at lade **læringsalgoritmer** og modeller behandle så store mængder data, som muligt. Det er som tidligere nævnt netop kendetegnende for **maskinlæring**, at den er i stand til at behandle store mængder af data på ganske avanceret vis, og man taler ofte om potentialerne i "big data".

Myndighederne er imidlertid ikke frit stillede på dette område. Brugen af oplysninger om borgere er reguleret i både forvaltningsretten og i databeskyttelsesforordningen, og disse regelsæt sætter rammerne for mængden og kvaliteten af data. Det er disse rammer, vi ser nærmere på i dette kapitel.

DATAKRAV I FORVALTNINGSRETTEN OG DATABESKYTTELSESRETTEN

Både i forvaltningsretten og databeskyttelsesforordningen stilles en række krav, der har til formål at sikre, at data er nødvendige og korrekte, før de må anvendes af myndigheder. For myndigheder er der en tæt sammenhæng i forpligtelserne efter de to regelsæt, da databeskyttelsesretten udfylder og supplerer de forvaltningsretlige principper på området (se kapitel 2).

Inden for forvaltningsretten følger det af det såkaldte officialprincip, at myndighederne skal oplyse en sag i nødvendigt omfang og sikre, at data er korrekt og retvisende. Officialprincippet indebærer, at offentlige myndigheder skal oplyse sagen på en fyldestgørende måde, før de træffer afgørelse. Formålet er at støtte kravet om, at forvaltningen træffer lovlige afgørelser. Officialprincippet er af afgørende betydning for profileringsmodellens lovlighed, da unødvendige eller forkerte data kan få betydning for modellens evne til at træffe korrekte afgørelser. Princippet er desuden en garantiforskrift, hvilket indebærer, at en afgørelse, der er afgjort uden iagttagelse af princippet, kan blive ugyldig.[1]

Databeskyttelsesforordningens principper om dataminimering og formålsbestemthed har til formål at sikre, at brugen af personoplysninger ikke bliver uforholdsmæssig indgribende. Dataminimeringsprincippet indebærer både, at uforholdsmæssige store mængder data ikke må behandles og, at en hvilken som helst oplysning, der ville føre til et uproportionalt indgreb, ikke kan behandles.[2] Samtidig supplerer princippet om rigtighed[3] og retten til berigtigelse af personoplysninger[4] officialprincippet, hvad angår krav til kvaliteten af data.[5]

Fælles for både officialprincippet og de databeskyttelsesretlige principper er, at de rummer særegne former for proportionalitetsvurderinger, dvs. vurderinger af, at indsamlede oplysninger om borgeren er egnede og nødvendige for det (lovlige) formål, som myndigheden forfølger.

I databeskyttelsesforordningen er der endvidere en regel om at sikre databeskyttelse gennem design i udviklingsstadiet. [6] Dette krav betyder, at myndigheden som dataansvarlig skal gennemføre "passende tekniske og organisatoriske foranstaltninger" for at sikre garantierne og rettighederne i forordningen. Foranstaltningerne skal både gennemføres i forberedelsesfasen, på tidspunktet hvor den første behandling af personoplysninger begynder og på tidspunktet for efterfølgende behandlinger.[7] I Justitsministeriets, Datatilsynets og Digitaliseringsstyrelsens vejledning anføres det desuden, at offentlige myndigheder vil skulle stille krav i udbudsfasen om databeskyttelse gennem design.[8]

Brugen af profileringsmodeller gør det nødvendigt tidligst muligt i beslutnings- og udviklingsstadiet at forholde sig til datakrav i forvaltningsretten og databeskyttelsesforordningen. I forhold til udviklingsstadiet udgør forordningens regler om databeskyttelse gennem design et vigtigt redskab hertil.

NØDVENDIGE DATA I PROFILERINGSMODELLER

Hvis myndigheden skal iagttage kravene til data efter forvaltningsretten og databeskyttelsesforordningen, må den som nævnt allerede under beslutningsprocessen og i udviklingsfasen forholde sig til en række spørgsmål.

I beslutningsfasen må myndigheden identificere, hvor meget data der skal indsamles, for at profileringsmodellen kan fungere mest optimalt, og om brugen af denne data er nødvendig for det tilsigtede formål. Myndigheden skal også sikre sig, at den nødvendige data overhovedet er tilgængelig, ligesom den skal sikre sig, at den tilgængelige data er korrekt og retvisende (se herom nedenfor).

I udviklingsfasen skal myndigheden ligeledes træffe en række valg, der skal sikre både nødvendighed og kvalitet.

Valget af variable er et af de vigtigste designvalg under konstruktionen af et datasæt, og det kan være vanskeligt at afgøre, hvilke af mange variable der skal indgå i datasættet. Valget er begrænset af såvel forvaltnings- og databeskyttelsesretten til variable, der udgør lovlige og nødvendige (proportionale) kriterier. Valget er imidlertid også begrænset af, at det normalt er sådan, at jo flere variable et datasæt indeholder, des større sandsynlighed er der for, at der optræder tilfældige og irrelevante sammenhænge mellem variabel og målegenskab. Valget af variable har således betydning for kvaliteten af modellens resultater, men skal foretages af myndighederne med udgangspunkt i proportionalitetsvurderingen. Typisk kan man ikke vide på forhånd, om en variabel vil få betydning i modellen og i givet fald hvor stor betydning, før man har trænet modellen.

Under modellens udvikling skal det også sikres, at den kan opfylde princippet om bl.a. dataminimering.

Et eksempel på manglende overholdelse af princippet om dataminimering findes i en afgørelse, hvor Datatilsynet udtalte alvorlig kritik om TDC's behandling af personoplysninger.[9] Sagen drejede sig om TDC's logning af lokationsdata ved mobildatatrafik. TDC anførte i sagen, at det var nødvendigt at registrere lokaliseringsdata for al mobildatatrafik for at overholde logningsbekendtgørelsens krav om registrering af lokaliseringsdata for MMS-kommunikation, eftersom det i TDC's system ikke var muligt at registrere det sidste uden samtidig at registrere det første. Datatilsynet anførte derimod, at opbygningen af TDC's system ikke kunne bruges som argument for manglende overholdelse af databeskyttelsesreglerne, og det samme gjaldt de eventuelle udgifter forbundet med at etablere nye systemer, der gjorde det muligt alene at registrere de nødvendige oplysninger. Tilsynet lagde bl.a. vægt på, at oplysningerne alene blev registreret på grund af TDC's systemopbygning, og at TDC selv havde oplyst, at TDC ikke har et formål med at registrere de pågældende, overskydende oplysninger. Tilsynet fandt derfor, at TDC's behandling af personoplysninger var i strid med forordningens artikel 5, stk. 1, litra c, om dataminimering.

Datatilsynets sag vedrørte en privat aktør, TDC, men princippet om dataminimering gælder naturligvis også for offentlige myndigheder og sagen giver også indblik i, hvorledes offentlige myndigheder skal indrette deres systemer for at overholde princippet.

Noget andet er, at myndigheder (i modsætning til private) på baggrund af officialprincippet har en vis skønsmargin i forhold til, om en oplysning anses for nødvendig. Myndigheder har ikke pligt til at gennemgå hver enkelt oplysning i sagen i forhold til kravet om nødvendighed.[10] Dette ændrer imidlertid ikke på, at myndigheder i forhold til udvikling af en profileringsmodel skal sikre, at dataminimeringsprincippet overholdes.

Ulovlige oplysninger kan ikke udgøre nødvendige data for sagsoplysningen, hverken, når dette skyldes selve oplysningens natur (oplysninger, som der ikke er hjemmel til at behandle), eller måden, hvorpå oplysningerne er indsamlet (ulovlig indsamling af i øvrigt lovlige og nødvendige oplysninger). En illustrativ sag herom er Datatilsynets afgørelse om SKAT's behandling af oplysninger, som SKAT havde modtaget fra Statsadvokaten for Økonomisk og International Kriminalitet (SØIK).[11] Oplysningerne var blevet indsamlet i strid med retsplejelovens regler og SØIK havde tilintetgjort dem. SKAT gjorde imidlertid brug af oplysningerne med henvisning til officialprincippet. Datatilsynet udtalte kritik om SKAT's behandling af oplysningerne og anførte, at den daværende persondatalovs regler om bl.a. god databehandlingsskik, herunder rimelig og lovlig behandling, som i dag er erstattet af databeskyttelsesforordningens regler begrænsede officialprincippet. Behandling af personoplysninger til sagens oplysning kunne kun ske inden for rammerne af de databeskyttelsesretlige regler.[12]

Er der oprindeligt indsamlet oplysninger om borgeren til andre formål end profileringsmodellens træning, må brugen af oplysningerne i datasættet som udgangspunkt ikke være uforenelig med det oprindelige formål. Ligeledes må brug af borgerens oplysninger, som oprindeligt blev indsamlet til andre formål ikke være uforenelig med brugen af oplysningerne til profilering af borgeren, når modellen tages i brug. I begge tilfælde følger det nemlig af databeskyttelsesforordningens princip om formålsbestemthed, at borgerens oplysninger ikke må bruges til formål, som er uforenelige med de, hvortil oplysningerne oprindeligt blev indsamlet (formålsbestemthedsprincippet). Det er muligt at undtage fra dette centrale princip herunder når oplysningerne anvendes til algoritmisk profilering. Denne undtagelse indebærer, at myndigheden kan anvende langt flere oplysninger, end formålsbestemthedsprincippet fordrer. Dette stiller krav til myndighedens oplysningspligt, som vi ser på i kapitel 7.

Modellerne trænes på historiske sager, dvs. for længst afgjorte sager om borgere, hvis personoplysninger også skal beskyttes. Et centralt tema er derfor beskyttelsen af personoplysninger i datasættet til træning af modellen.

BESKYTTELSE AF PERSONOPLYSNINGER I DATASÆT TIL TRÆNING

En model bliver trænet på historiske data, herunder tidligere praksis fra myndigheden samt en lang række oplysninger, som er samkørt fra en række forskellige kilder. Det stiller krav til beskyttelsen af personoplysninger for de borgere, hvis oplysninger figurerer i datasættet, som modellen trænes på.

Det rejser bl.a. spørgsmål om, hvorvidt der anvendes data for en større eller mindre del af befolkningen; om mængden og typen af data er proportional med formålet samt om der kan der anvendes data, som i mindre grad eksponerer borgernes private forhold.[13]

Efter databeskyttelsesforordningen skal beskyttelsen af personoplysninger i datasæt tænkes direkte ind i modellens design (databeskyttelse gennem design).

En effektiv gennemførelse af forordningens regler om databeskyttelse gennem design kræver kendskab til teknologier og metoder hos udviklerne, projektledere, testteams, databeskyttelsesrådgivere mv.[14] Databeskyttelse gennem design kræver således tværfagligt arbejde og er præget af den væsentlige udfordring, at metoder og rutiner for beskyttelse gennem design endnu ikke er gængse, men skal udvikles løbende ved sammensætningen af forskellige løsninger og teknologier.[15]

Såfremt det er muligt, kan brugen af anonymiserede eller fiktive data løse de databeskyttelsesretlige bekymringer. I disse tilfælde indeholder data slet ikke personoplysninger i forordningens forstand (se boks). Er der derimod tale om personoplysninger, nævner forordningen bl.a. pseudonymisering og dataminimering som to metoder, hvorpå databeskyttelse kan sikres, men der kan også anvendes andre metoder eller teknologier – såkaldte Privacy Enhancing Technologies (PET'er).[16]

En udfordring ved pseudonymiserede datasæt er, at der er en risiko for utilsigtet reidentifikation, når værdierne af variable, som optræder i datasættet, sammenholdes med information om identificerede individer. Studier har vist, at det ofte vil være muligt at reidentificere mange personer i datasæt, selv med adgang til relativt få informationer om dem. En berømt analyse af USA's folketalsopgørelse fra 1990 viste således, at mere end halvdelen af befolkningen (53%) kunne identificeres unikt ved at anvende kombinationen af det område de boede i (f.eks. by), deres køn, og deres fødselsdato.[17] Et senere studie af folketalsopgørelsen fra 2000 demonstrerede, at 63% af befolkningen kunne identificeres unikt ved at anvende kombinationen af køn, fødselsdato og postnummer.[18]

ANONYMITET ELLER PSEUDONYMITET

Databeskyttelse vedrører oplysninger om personer og spørgsmålet om anonymisering eller pseudonymisering af data er derfor helt afgørende. Anonyme oplysninger er oplysninger der ikke kan føres tilbage til en person, og de falder derfor udenfor databeskyttelsesforordningens område. Pseudonymiserede oplysninger er derimod omfattet af reglerne, da oplysningerne kan kobles til personen ved brug af supplerende oplysninger. Medmindre oplysningerne er anonymiserede gælder de krav, der følger databeskyttelsesforordningen, herunder kravet om formål, proportionalitet og dataminimering.

Andre eksempler på privatlivsbeskyttende teknologier eller PET'er er mere avancerede tekniske løsninger. Nogle af disse tekniske løsninger indebærer, at kvaliteten af data i datasættet bliver mindsket, hvilket kan have en negativ virkning på den endelig profileringsmodels kvalitet. Det betyder, at der for disse løsninger skal foretages et valg mellem modellens adgang til at danne præcise resultater på den ene side og beskyttelsen af personoplysninger i datasættet på den anden.[19]

Beskyttelsen af personoplysninger kan også styrkes gennemtiltag rettet mod informering og kontrol, hvor den registrerede borger oplyses om brugen af personoplysninger i modellen og kan kontrollere sine oplysninger.[20]

Overholdelsen af dataminimeringsprincippet herunder også i beskyttelsen af oplysninger i datasæt bør som nævnt i vores anbefaling indgå i den tidligste AI-konsekvensanalyse og periodisk herefter (se kapitel 4).

Af hensyn til beskyttelsen mod diskrimination kan det være nødvendigt at vide, om der er oplysninger om f.eks. etnicitet eller handicap i datasættet og hvorledes oplysningerne indgår i modellen. Dette kan skabe en potentiel udfordring mellem beskyttelsen af personoplysninger i datasættet og adgangen til at opdage og imødegå risikoen for diskrimination.[21]

EU-Kommissionens udkast til en forordning for kunstig intelligens (se kapitel 2) er der udtrykkeligt taget stilling til samspillet mellem på den ene side beskyttelsen af personoplysninger i form af beskyttede kendetegn, hvoraf de fleste som udgangspunkt ikke må behandles efter databeskyttelsesforordningen – og på den anden side forbuddet mod diskrimination, som i kontekst af en profileringsmodel ofte kræver adgang til netop disse oplysninger. Der lægges i udkastet op til, at sådanne beskyttede oplysninger kan behandles, hvis der sikres passende sikkerhedsforanstaltninger, herunder tekniske begrænsninger for videreanvendelse og tekniske metoder for beskyttelse af privatliv og personoplysninger.[22] Vi ser nærmere på samspillet mellem diskriminationsforbuddet og databeskyttelse i kapitel 8.

KORREKTE OG RETVISENDE DATA – I HELE MODELLENS LEVETID

Udvælgelsen af data til træning af modellen, herunder tidligere afgørelser fra myndigheden, er væsentlig, da dette datasæt udgør modellens "grundsandhed". Dette forhold mellem datakvalitet og modellens vurderinger opsummeres ofte i sloganet: "Garbage in, garbage out." Hvis tidligere afgørelser er behæftet med fejl f.eks. fordi der indgår afgørelser, som er usaglige eller diskriminerende, vil der være risiko for, at modellen viderefører disse fejl. Også øvrige fejl i datasættet kan videreføres i modellen. Træner man således modellen på data, der ikke er korrekte eller retvisende for den del af befolkningen, som berøres af modellen, kan læringsmodellen komme til at træne på usaglige eller diskriminerende og dermed ulovlige sammenhænge (se også kapitel 8).

I udviklingsstadiet skal der derfor være fokus på, hvordan data udvælges, så det både er korrekt og retvisende for samtlige befolkningsgrupper.

Hvis der korrigeres eller forsimples i datasættet, skal datasættet fortsat have samme kvalitet og være retvisende. Navnlig alt for mange forsimplinger eller såkaldte datareduktioner kan føre til, at læringsalgoritmen overser sammenhænge, der måtte være relevante. Det er således vigtigt, at AI-konsekvensanalyserne (se kapitel 4) fokuserer på såvel nødvendighed og kvalitet i udformningen af datasættet.

EU-Kommissionens udkast til en forordning for kunstig intelligens (se kapitel 2) er der foreslået regler for datakvalitet, bl.a. når offentlige myndigheder bruger profileringsmodeller til at træffe afgørelser, som udgør højrisikotilfælde efter udkastet. Det følger af udkastet, at datasættet skal være underlagt "datastyrings- og dataforvaltningspraksisser" som bl.a. adresserer dataindsamling og -behandling, eventuelle forudgående antagelser om de oplysninger, som data skal repræsentere, vurdering af datasættets egnethed og nødvendighed, kortlægning af eventuelle datamangler eller utilstrækkeligheder, og hvordan de kan afhjælpes. Datasættet skal desuden efter udkastet være relevante, repræsentative, fejlfri og fuldstændige.[23]

Datakvalitet kan forringes som følge af, at data gradvis bliver forældede (eng.: "dataset shift"). Udfordringen skyldes, at de faktiske omstændigheder forandrer sig, så de sammenhænge, som det oprindelige datasæt indeholder, ikke længere passer med virkeligheden. Opdatering af datasættet kan forbedre datakvalitet, især hvis myndigheden opdaterer data også med fokus på, at forkerte data rettes til eller udgår fra datasættet.

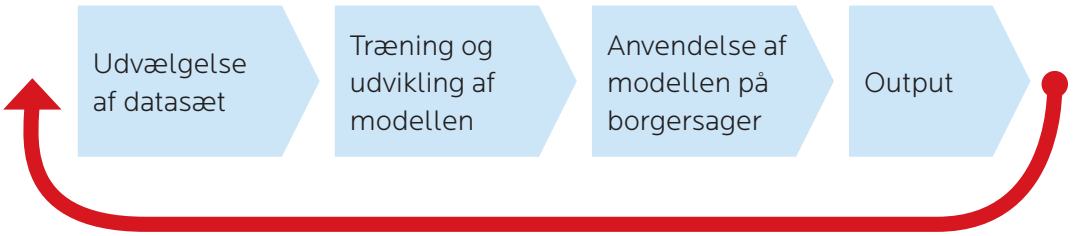
Udfordringen med forældede data kan forsøges løst ved periodisk at gentræne modellen på et datasæt der alene indeholder nyere oplysninger, eller ved at få **læringsalgoritmen** til at give mindre vægt til data, jo ældre de er. I sidstnævnte tilfælde løses udfordringen ved at hensynet indgår direkte i modellens design.

En anden problemstilling er, at modellens output efter omstændighederne skal suppleres af oplysninger, som hverken kan eller skal indgå i modellens input. Dette er tilfældet for oplysninger, som myndigheden får kendskab til ved partshøring. For at opfylde kravene om en tilstrækkelig og fyldestgørende sagsoplysning efter officialprincippet skal myndigheden sikre, at oplysninger erfaret ved partshøring indgår i den samlede vurdering af borgerens sag. Modellens output skal derfor efter omstændighederne suppleres med oplysninger, som borgeren sidder inde med. Ellers risikerer myndigheden af træffe afgørelse på et fejlagtigt og ufuldstændigt grundlag (se kapitel 6).

Vi anbefaler, at der i vejledningen (se kapitel 4) stilles krav om periodisk gentræning af modellen på nye data eller, at modellen udvikles således, at den som udgangspunkt giver mindre vægt til data, jo ældre de er.

Såfremt myndigheden ikke opdager fejl i datasættet, kan det føre til såkaldte negative feedbacksløjfer der efter omstændighederne kan forstærke risikoen for usaglighed (se kapitel 6) og diskrimination (se kapitel 8). I en negativ feedbacksløjfe trænes modellen på data, der enten fører til usaglighed eller diskrimination og som modellen viderefører i sine egne vurderinger. Disse vurderinger indgår i en myndigheds afgørelser, som derfor også kommer til at indebære usaglighed eller diskrimination. Når nyere afgørelser herefter inkluderes i datasætter for træning af en opdateret model, får modellen tendens til endnu stærkere forskelsbehandling. (se figur).

SIKRING AF DATAKVALITET I HELE MODELLENS LEVETID



En feedbacksløjfe er kun negativ, når anvendelsen af en model fører til, at myndigheden begår flere fejl. Hvis den øgede fejlrate registreres, så vil udfordringerne gradvist forsvinde fra data, og en opdateret model trænet på de nye data vil have tendens til færre fejl eller mindre diskrimination.

En negativ feedbacksløjfe vil netop opstå når det er vanskeligt at måle eller opdage forkerte resultater. Anvendes en profileringsmodel f.eks. til tidlig opsporing af børns mistrivsel, skal modellen vurderes ikke blot i forhold til, hvor mange sager, modellen opdagede, som var korrekte, men også, om der var tilfælde, som burde være opdaget, men som modellen vurderede, at der ikke var grund til at fokusere på. Det giver sig selv, at det vil være meget vanskeligt at måle, hvor mange sager, myndigheden burde havde opdaget.

En enkelt metode for at forhindre eller begrænse virkningen af negative feedbacksløjfer, er at introducere et element af tilfældighed i afgørelsesprocessen.[24] Den

negative feedbacksløjfe opstår, fordi modellen påvirker afgørelsesprocessen, og opdateres med data om resultaterne af beslutningerne. Ved at introducere tilfældighed i processen sikres, at modellen også opdateres med data, som er skabt uafhængigt af modellens vurderinger. Derved får den løbende træning af modellen mulighed for at korrigere for de effekter, som en negativ feedbacksløjfe skaber. Et eksempel kunne være en model, som anbefaler selvangivelser til SKAT, som skal udtages til manuel kontrol. SKAT kunne i den situation forsøge at forebygge en negativ feedbacksløjfe ved også at udtage en vis kvote af tilfældigt udvalgte selvangivelser til manuel kontrol.

Udfordringen for metoden med at introducere tilfældighed i beslutningsprocessen er, at dette i nogle situationer vil være retssikkerhedsmæssigt problematisk.

Et element af tilfældighed er relativt uproblematisk i eksemplet med udtagning af selvangivelser til manuel kontrol. Som eksempel på det modsatte kan nævnes, at mange stater i USA benytter modeller til at vurdere risici i forbindelse med beslutninger om varetægtsfængsling og prøveløsladelse.[25] Introduktionen af tilfældighed i en sådan beslutning, således at visse sigtede eller dømte blev løsladt eller fængslet fordi de var tilfældigt udtrukket, ville være indlysende uacceptabelt. Det samme kunne være tilfældet i særligt indgribende forvaltningsafgørelser.

I Kommissionens udkast til en forordning om kunstig intelligens (se kapitel 2) foreslås det, at AI-systemer skal udvikles således, at virkningerne af eventuelle negative feedback loops er behørigt adresseret med foranstaltninger for at mindske deres virkning i systemet.[26]

Dette bør efter vores vurdering ske som led i de periodiske AI-konsekvensanalyser og tilsynet med disse (se kapitel 4).

NOTER

1 Ombudsmanden (29. juni 2021), Ombudsmandens myndighedsguide, Officialprincippet, afsnit 2 tilgængeligt på: https://www.ombudsmanden.dk/myndighedsguiden/generel_forvaltningsret/officialprincippet/#cp-title
2 Kuner, Bygrave, Docksey og Drechsler (2020), The EU General Data Protection Regulation (GDPR): A Commentary, Oxford University Press, s. 317
3 Databeskyttelsesforordningen 2016/679 af 27. april 2016, artikel 5, stk. 1, litra d
4 Databeskyttelsesforordningen 2016/679 af 27. april 2016, artikel 16
5 Fenger (2018), Forvaltningsret, Jurist- og Økonomforbundets Forlag, s. 489
6 Databeskyttelsesforordningen, (EU) 2016/679 af 27. april 2016, artikel 25
7 Justitsministeriets, Datatilsynets og Digitaliseringsstyrelsens (juni 2018), Vejledning om Behandlingssikkerhed og Databeskyttelse gennem design og standardindstillinger, s. 25, tilgængelig på: https://www.datatilsynet.dk/Media/7/C/Behandlingssikkerhed%20og%20databeskyttelse%20gennem%20design%20og%20standardindstillinger%20(2).pdf
8 Justitsministeriets, Datatilsynets og Digitaliseringsstyrelsens (juni 2018), Vejledning om Behandlingssikkerhed og Databeskyttelse gennem design og standardindstillinger, s. 29f, tilgængelig på: https://www.datatilsynet.dk/Media/7/C/Behandlingssikkerhed%20og%20databeskyttelse%20gennem%20design%20og%20standardindstillinger%20(2).pdf
9 Datatilsynet, afgørelse af 11. februar 2019 (j.nr. 2018-31-0070), tilgængelig på: https://www.datatilsynet.dk/afgoerelser/afgoerelser/2019/feb/klage-over-tdcs-registrering-af-trafik-og-lokaliseringsdata
10 Se f.eks. Korfitz Nielsen og Lotterup (2020) Databeskyttelsesforordningen og databeskyttelsesloven med kommentarer, Djøf forlag s. 329f.
11 Datatilsynet, afgørelse af 17. maj 2018 (j.nr. 2018-32-0065), [tilgængelig på: https://www.datatilsynet.dk/afgoerelser/afgoerelser/2018/maj/skats-brug-af-ulovligt-fremskaffede-oplysninger].
12 Sagen er gengivet hos i Korfitz Nielsen og Lotterup (2020) Databeskyttelsesforordningen og databeskyttelsesloven med kommentarer, Djøf forlag s. 551
13 Se Kofod Olsen, (2019) Håndbog i Dataansvarlighed, Djøf, s. 463 med disse og øvrige spørgsmål til dataansvarlige.
14 Se Kofod Olsen, (2019) Håndbog i Dataansvarlighed, Djøf, s. 453f
15 Se Kofod Olsen, (2019) Håndbog i Dataansvarlighed, Djøf, s. 454.
16 Se Kofod Olsen, (2019) Håndbog i Dataansvarlighed, Djøf, s. 461ff.
17 Sweeney, (2000), Simple Demographics Often Identify People Uniquely. Data Privacy Working Paper, Carnegie Mellon University.

18 Golle, (2006). Revisiting the uniqueness of simple demographics in the US population. Proceedings of the 5th ACM workshop on Privacy in electronic society. Alexandria, Virginia, USA, Association for Computing Machinery: 77–80
19 Se bl.a. om såkaldt k-anonymitet gennem datareduktion: Ji, ZI, Lipton og Elkan (2014) Differential Privacy and Machine Learning: a Survey and Review. arXiv e-prints, arXiv:1412.7584. Om differentieret privacy gennem dataløring: Dwork, og Roth (2014). The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science 9(3-4): 211-407. Om multi-party computation: Kilbertus, Gascón, Kusner, Veale, Gummadi, et al. (2018). Blind Justice: Fairness with Encrypted Sensitive Attributes. arXiv:1806.03281. Se også Veale og Binns (2017) Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society 4(2): 205395171774353. Se også mere generelt: Colesky, M., J.-H. Hoepman and C. Hillen (2016). A Critical Analysis of Privacy Design Strategies.
20 Se Kofod Olsen, (2019), Håndbog i Dataansvarlighed, Djøf s. 469ff. Se også Europarådets anbefalinger Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems af 8. April 2020, pkt. 2.1.
21 Se EU's Agentur for Grundlæggende Rettigheder (FRA) (2018), #BigData: Discrimination in data-supported decision making s. 9 tilgængelig på: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-focus-big-data_en.pdf og Borgesius (2018), Discrimination, artificial intelligence, and algorithmic decision-making, s. 25 tilgængelig på: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73
22 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final, artikel 10, stk. 5
23 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final artikel 10
24 Kroll, Huey, Barocas, Felten, Reidenberg, et al. (2017). "Accountable Algorithms." University of Pennsylvania Law Review 165(3): 633
25 Berk, Heidari, Jabbari, Kearns and Roth (2018). "Fairness in Criminal Justice Risk Assessments: The State of the Art." Sociological Methods & Research
26 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final artikel 15

KAPITEL 6

MODELLENS OUTPUT: RAMMER FOR ANVENDELSE

I dette kapitel fokuserer vi på de rettigheds- og retssikkerhedsmæssige udfordringer, der knytter sig til en profileringsmodels output, dvs. den profilering eller vurdering af borgeren, som modellen genererer.

I faglitteraturen om maskinlæring fremhæves ofte, at der er behov for "fairness" og beskyttelse mod **bias** i brugen af profileringsmodeller.[1]

Begreberne er ikke juridiske, men dækker typisk over hensyn, der er beskyttet retligt. Det juridiske modstykke til fairness og beskyttelse mod bias er saglighedskravet, som vi beskæftiger os med i dette kapitel, og diskriminationsforbuddet, som vi ser på i kapitel 8.

Der er imidlertid efter vores vurdering et stort behov for at klarlægge, hvordan forvaltningsretlige regler og principper skal – og bør – finde anvendelse på brugen af profileringsmodeller, og spørgsmålet er endnu kun meget sparsomt adresseret. I den forbindelse tjener faglitteraturens beskrivelser af fairness og bias til at illustrere problemstillinger, som myndighederne må forventes at blive mødt med, når de anvender profileringsmodeller og skal overholde rettigheds- og retssikkerhedsmæssige krav.

Nogle af disse udfordringer opstår i udviklingsfasen, hvor modellen trænes og dens kvalitet skal sikres. I denne proces opstår risiko nogle retssikkerhedsmæssigt problematiske designvalg eller algoritmiske dilemmaer. Det fører efter vores vurdering til, at der bør sættes grænser for myndighedernes brug af profileringsmodeller til fuldautomatiseret brug, ligesom det fører til, at problemerne bør adresseres i myndighedernes AI-konsekvensanalyser (se kapitel 4).

Andre udfordringer opstår i anvendelsesfasen. Disse vedrører samspillet mellem profileringsmodeller og sagsbehandleren når modellen bruges til beslutningsstøtte. Her er det vigtigt at sikre, at modellen får den tiltænkte indflydelse på den endelige afgørelse, samt at afgørelsen træffes på et tilstrækkeligt oplyst grundlag.

Vi ser i det følgende på begge disse typer af udfordringer og kommer med en række anbefalinger, der sætter rammerne for brugen af modellens output.

PROFILERINGSMODELLER OG FORVALTNINGSRETLIGE AFGØRELSER

Som vi har nævnt, er myndighederne bundet af de samme forpligtelser og regler, uanset om der anvendes profileringsmodeller eller ej, og uanset om modellerne bruges til fuldautomatisering af afgørelsen eller til beslutningsstøtte. En afgørelse truffet af eller med hjælp fra en model skal med andre ord leve op til de samme krav, som enhver anden forvaltningsafgørelse.

Alle forvaltningsafgørelser skal overholde saglighedskravet – også kaldet forbuddet mod magtfordrejning. Saglighedskravet indebærer, at forvaltningen ikke må inddrage uvedkommende hensyn, dvs. de er forpligtet til at lade sig styre af de hensyn bag reglerne, som lovgivningsmagten har anerkendt som relevante i det konkrete tilfælde, ligesom der er hensyn, som er usaglige og dermed ulovlige at inddrage, f.eks. hensyn til borgerens partipolitiske tilhørsforhold, køn, religiøse tilhørsforhold eller andre forhold, som i det konkrete tilfælde anses for usaglige.

Saglighedskravet er en grundlæggende forudsætning for, at myndighedens afgørelser er lovlige og selvom saglighedskravet indgår i alle myndighedens afgørelser, aktualiseres det især, når afgørelserne baseres på regler, der giver grundlag for myndighedens fortolkning eller skøn.

Det forvaltningsretlige skøn opstår når der anvendes regler, som fra lovgivers side tilsigtet er ufuldstændige i beskrivelsen af betingelser, som skal være opfyldt, eller kriterier, der skal tillægges vægt, når der træffes en afgørelse.[2] Modstykket til det frie skøn er en afgørelse baseret på entydige regler. F.eks. vil en entydig regel kunne opstille krav om "mere end 250 kalenderdage om året", mens en skønsmæssig og skønspræget regel vil kunne opstille krav om "størstedelen af året".[3]

De fleste regler befinder sig oftest et sted mellem de to yderpunkter og der er en glidende overgang fra skønsregler til regler, der rummer fortolkningstvivl og har en vaghed eller elasticitet i deres indhold.

Ombudsmanden anfører om det forvaltningsretlige skøn i automatiserede processer:

"Hvis regelgrundlaget på et område forudsætter, at en afgørelse eller elementer i afgørelsen beror på et skøn, skal sagsprocessen indrettes, så det er muligt at udøve et sådant skøn. Det kan eventuelt indebære, at sagsbehandlingen ikke fuldt ud kan være digital. [...] En forudsætning for en hel eller delvis automatisering af sagsbehandlingen på et område er således som udgangspunkt, at afgørelserne eller de automatiserede dele af afgørelserne kan træffes efter rent objektive kriterier – det vil sige efter på forhånd fastsatte faktuelle kriterier og en fast givet retsvirkning af et givet faktum. Man kan også sige, at hvis det ikke er muligt klart at svare ja eller nej på et givet spørgsmål, kan behandlingen af det som udgangspunkt ikke automatiseres."[4]

Et lignende synspunkt er afspejlet i vejledningen om digitaliseringsklar lovgivning, hvori det anføres, at objektive kriterier i lovgivningen som udgangspunkt er en forudsætning for, at sagsbehandlingen automatiseres, og det anføres videre:

"I tilfælde, hvor lovgivningen på et område forudsætter, at afgørelsen eller elementer i afgørelsen beror på et skøn, vil det kunne begrænse, hvor langt automatiseringen af sagsbehandlingen kan række. Det kan løses ved, at skønnet foretages manuelt, mens øvrig sagsbehandling understøttes digitalt."[5]

I medfør af en politiske aftale fra 2018 har Folketinget pålagt sig selv at lovgive "digitaliseringsklart". Kravet om digitaliseringsklar lovgivning og udfordringerne ved at automatisere skønnet har ført til, at man fra lovgivers side i videst muligt omfang fastsætter regler med objektive kriterier, frem for skønsregler. I vejledningen om digitaliseringsklar lovgivning anføres det, at skønnet fortsat skal være til stede i sager, hvor der skal foretages en helhedsvurdering.[6] En fjernelse af skønsregler i lovgivningen kan imidlertid rejse retssikkerhedsmæssige betænkeligheder. For eksempel er der risiko for, at borgerens konkrete forhold ikke i tilstrækkelig grad bliver inddraget af forvaltningen, hvis reglerne ikke tillader en skønsafvejning.[7]

Afgørelser baseret på skøn eller regler med fortolkningsrum er kendetegnet ved, at de i høj grad lægger vægt på konkrete og individuelle omstændigheder i borgerens sag. Reglernes vage formuleringer varetager på den måde et retssikkerhedsmæssigt hensyn og et nærhedshensyn, hvorefter beslutninger, der har indflydelse på borgeren, skal afgøres tættest muligt på borgeren og tage hensyn til dennes konkrete situation, ligesom myndighedens fagkundskab skal være styrende for udfaldet af sagen.

UDFORDRINGER I UDVIKLINGSFASEN

En række forhold ved profileringsmodellerne indebærer efter vores vurdering, at de ikke bør bruges til fuldautomatiserede afgørelser, når afgørelsen baseres på skøn eller fortolkning.

Begrænsninger i brugen af fuldautomatiserede afgørelser indebærer, at skønsvurderingen eller fortolkningen forbliver "manuel", som også ombudsmanden og Digitaliseringsstyrelsen lægger op til.

Der er efter vores vurdering en række grunde til at begrænse brugen af fuldautomatisering. Nogle af grundene følger af de risici for diskrimination, som vi behandler i kapitel 8. Andre følger af forholdene, som vi gennemgår i det følgende.

Myndighederne er forpligtede til at indlejre rettighederne direkte i modellens design, også kaldet "rettigheder gennem design", der både følger af EU's databeskyttelsesforordning[8] og er blevet fremhævet i litteraturen som et forvaltningsretligt krav med udgangspunkt i ombudsmandspraksis.[9]

Rettigheder kan imidlertid sjældent alle indlejres på en sådan måde, at man kan skabe en model, der fuldkommen beskytter borgerens rettigheder og retssikkerhed. Dette kan enten skyldes, at beskyttelseshensynet ikke kan sættes på en matematisk formel, som modellen kan anvende, eller at man i udviklingen af modellen kan blive tvunget til at vælge mellem et eller flere lige vigtige beskyttelseshensyn. Dette benævner vi designvalg eller algoritmiske dilemmaer.

Disse valg eller dilemmaer taler efter vores vurdering for at begrænse brugen af fuldautomatisering, ligesom valgene eller dilemmaerne efter vores opfattelse udtrykkeligt bør adresseres i myndighedernes AI-konsekvensanalyser (se kapitel 4), når myndighederne anvender modellerne til beslutningsstøtte.

Det må forventes, at myndighederne oftest i sager af særligt indgribende karakter netop vil anvende modellerne til beslutningsstøtte, Her er problemstillingerne fortsat til stede på trods af den begrænsede brug af modellerne – blot kan de retssikkerhedsmæssige betænkeligheder imødegås og mindskes i højere grad end ved fuldautomatisering.

I det følgende gennemgår vi de problemstillinger, som bør sætte grænser for brugen af fuldautomatisering og krav til AI-konsekvensanalyserne i tilfælde af beslutningsstøtte.

Vi anbefaler, at Justitsministeriet tager initiativ til at indføre regler i forvaltningsloven om, at myndighederne alene må anvende profileringsmodeller til fuldautomatisering, når afgørelsen er baseret på regler med entydige kriterier.

GRÆNSER FOR AUTOMATISERING AF SAGLIGHEDSKRAVET

Selvom alle afgørelser rettet mod borgeren skal være individuelle, tillader retsregler med entydige kriterier ofte, at der træffes standardafgørelser. Sådanne må derimod ikke træffes, hvor der er tale om en skønsvurdering. Skønnet er pligtmæssigt og forbuddet mod at standardisere en skønsafgørelse omtales som forbuddet mod at sætte skøn under regel.[10] Også for regler, der efterlader rum for fortolkning, skal myndigheden sikre, at afgørelsen er truffet ud fra individuelle og konkrete omstændigheder. Til forskel fra skønsvurderingen, kan myndigheden imidlertid i større grad standardisere sin praksis ved anvendelsen af sådanne regler. Kun for helt entydige regler er det muligt at træffe en ren "standardafgørelse". I praksis er det imidlertid de færreste regler, der fuldstændigt og entydigt fastlægger regelanvendelsen.

Vanskeligheden ved automatiseringen af en beslutning og ved at sætte myndighedens vurdering på en formel i modellen øges, jo mere vag eller åben for skøn og fortolkning den pågældende retsregel er (se figur). Hvis den målegenskab, modellen skal vurdere, ikke baserer sig på entydige kriterier, er det vanskeligt eller umuligt at definere, hvordan modellen opnår den "bedst mulige" vurdering og hvordan fejl skal måles. Hvis afgørelsen derimod skal træffes ud fra entydige kriterier i en retsregel, kan en model sandsynligvis trænes til at danne så gode vurderinger som muligt, f.eks. ved brug af en **tabsfunktion** (se kapitel 3).

Det er væsentligt at fremhæve, at i tilfælde, hvor fuldautomatisering kan anvendes, bør det sikres, at samtlige materielle og formelle forvaltningsretlige krav (se kapitel 2), herunder partshøring fortsat sikres, som hvis sagen blev behandlet manuelt.

PROFILERINGSMODELLER OG FORVALTNINGSRETLIGE AFGØRELSER

Regler med frit skøn	Skønsprægede regler	Fortolkningstvivl i regler	Entydige regler
	Automatiseret beslutningsstøtte er mulig		**Fuldautomatisering er mulig**

En beslægtet udfordring, som blot skal nævnes kort her er, at der kan være forskel på den målegenskab, som en profileringsmodel på baggrund af tilgængelig data er i stand til at vurdere, og den afgørelse, en myndighed i sidste ende skal træffe efter retsreglen. Trænes en model f.eks. på historiske sager om tvangsfjernelse af børn med det formål at sætte den i stand til at vurdere fremtidige sager om risikoen for mistrivsel af børn, skal forskellen mellem det tilgængelige **datasæt** om tvangsfjernelse og den tiltænkte målegenskab om mistrivsel indarbejdes i modellens design.

Tre konkrete problemstillinger fører bl.a. til vores anbefaling om, at fuld-automatiserede afgørelser kun bør træffes på grundlag af entydige regler

AFVEJNING AF FEJLRATER OG FEJLTYPER

Den første problemstilling, som vi vurderer sætter grænser for fuldautomatiseret brug og krav om konsekvensanalyser, er valget mellem at sikre en samlet set lav **fejlrate** for modellen eller optimere modellen i forhold til bestemte **fejltyper**.

I kapitel 3 omtalte vi fejltyper og fejrater som et væsentligt element i udviklingen af profileringsmodeller. Målet er naturligvis, at modellens fejlrate – dvs. antallet af forkert vurderede borgere eller sager – skal være så lav som muligt, men i praksis indebærer det en afvejning af forskellige hensyn.

Forenklet sagt kan en model levere falske positive, hvor modellen vurderer, at personen har målegenskaben, selvom dette ikke er tilfældet (en borger får tilskud, selvom vedkommende ikke er berettiget), og falske negative, hvor modellen omvendt vurdere, at personen ikke har måleegenskaben, selvom det faktisk er tilfældet (en borger får afslag på tilskud, selvom vedkommende er berettiget). Balancen mellem de to typer fejl afhænger af, hvor man sætter **tærskelværdien**, dvs. grænsen for, hvornår en borger vurderes positivt eller negativt i forhold til målegenskaben. Typisk kan man kun reducere antallet af falske positive på bekostning af flere falske negative eller omvendt (se kapitel 3).

Omsætter vi dette til en model, der f.eks. vurderer borgeres berettigelse til et givent tilskud, bliver spørgsmålet, om det er vigtigst, at en model sikrer, at der ikke er for mange uberettigede, der får tilskud end at sikre, at der ikke er for mange berettigede, der fejlagtigt får afslag.

Ud fra almindelige proportionalitetsbetragtninger kan falske positive virke særligt indgribende for borgeren ved bebyrdende eller sanktionerende afgørelser, mens falske negative må antages at have særlig stor betydning for borgeren i sager, hvor det vurderes, om en borger er tilskudsberettiget eller ej.

FASTSÆTTELSE AF TÆRSKELVÆRDIEN I STAR'S PROFILERINGSMODEL FOR NYLEDIGE

Styrelsen for Arbejdsmarked og Rekruttering (STAR) benytter en profileringsmodel til at vurdere nyledige risiko for at blive langtidsledige (se bilag 2). I modellen har man fastsat en tærskelværd, hvor der skal være 60% sandsynlighed for at personer har målegenskaben (dvs. vurderes i risiko for at blive langtidsledige) før de klassificeres positivt. Dette skyldes en bevidst beslutning fra STAR's side om, at det er vigtigere at undgå falsk positive end falsk negative. Myndigheden har med andre ord vurderet, at det vigtigst at undgå, at borgere uberettiget bliver klassificeret, som værende i risiko for langtidsledighed, også selvom det måtte betyde, at der kan være tilfælde, hvor borgere i risikogruppen ikke bliver opdaget af modellen.

Men ikke alle sager er lige lette at dele op. Er det f.eks. vigtigst, at en model finder frem til flest mulige tilfælde af mistrivsel i børnefamilier eller at modellen ikke fører til ressourcetunge observationer og mistænkeliggørelse af flere familier end nødvendigt?

Sikring mod den ene type af fejl frem for den anden vil ofte være kontekstbaseret og det er derfor ikke altid muligt at operere ud fra en standardregel for fordelingen af fejltyper på bebyrdende og begunstigende afgørelser.

Udover et valg mellem fejltyperne indbyrdes skal der således også træffes et valg overfor den samlede fejlrate. Definitionen af tærskelværdien er nemlig også vigtig, fordi den kan (og typisk vil) påvirke den samlede fejlrate.

Dertil kommer spørgsmålet om, hvor fejlbarlig en model må være. Skal modellen vurderes ud fra lovens ideal eller menneskelige (og dermed fejlbarlige) sagsbehandlere? Selvom legalitetsprincippet (se kapitel 2) skal sikre, at forvaltningen når frem til en lovlig og saglig afgørelse, begår myndigheder fejl i den virkelige verden, og der er altid en risiko for, at der træffes en forkert afgørelse. Menneskelige fejl i forvaltningen er utilsigtede og ikke kalkulerede, og myndigheden er forpligtet til at undgå tilsvarende fejl i fremtiden. Der er en pligt til at indrette forvaltningen således, at fejl så vidt muligt undgås.

Omvendt kan det blotte forhold, at der for modeller – ligesom for mennesker – vil være en risiko for fejl næppe i sig selv føre til, at profileringsmodeller slet ikke kan bruges i sagsbehandlingen. Det kan endda overvejes, om fejlmarginen for en profileringsmodel efter omstændighederne kan være mere forudsigelig (f.eks., at myndigheden på forhånd ved, at den største risiko for fejl vedrører bestemte fejltyper, fordi man har truffet et designvalg herom under udviklingen) end for sagsbehandlere. Således kan der for profileringsmodeller være større viden om, i hvilke tilfælde, fejl er mest nærliggende. Profileringsmodeller kan imidlertid både videreføre og forstærke eksisterende udfordringer og skabe nye rettighedsmæssige udfordringer.

Af EU-Kommissionens udkast til en forordning om kunstig intelligens (se kapitel 2) følger, at modellerne skal designes og udvikles på en sådan måde, at de opfylder deres tiltænkte formål og bl.a. sikrer et "passende niveau af nøjagtighed" i hele modellens livscyklus.[11] Spørgsmålet bliver imidlertid, hvad et "passende niveau" skal omsættes til i valget mellem afvejningen af fejlrater og fejltyper.

Modellens fejlrate kan være en misvisende målestok for kvalitet i situationer, hvor målegenskaben er enten meget almindelig eller meget ualmindelig i datasættet. Hvis målegenskaben f.eks. kun optræder i 0,1% af personerne, så kan en model, som har en fejlrate på 1% være dårligere til korrekt at vurdere målegenskaben, end en model, som helt konsekvent afviser at personerne har målegenskaben. Til gengæld kan en sammenligning af de to modeller, som ikke kigger på den overordnede fejlrate, men på deres evne til at ramme rigtigt i netop de tilfælde, hvor en person besidder målegenskaben, vise at den første model i denne henseende er bedre end den anden.

Disse valg kan indebære, at retssikkerhed for en gruppe af borgere forringes på bekostning af en eller flere andre grupper og indgår derfor som et væsentligt moment til støtte for vores anbefalinger ovenfor.

TILPASNING AF MODELLEN

Den anden problemstilling, der fører til vores anbefaling om grænser for fuldautomatisering og krav til AI-konsekvensanalyser, vedrører designvalget mellem **bias** og **varians**.

I kapitel 3 omtalte vi balancen mellem bias og varians som et væsentligt element i udviklingen af modellerne. Bias og varians er udtryk for, hvordan modellen sorterer de sager, den bliver præsenteret for. Høj bias vil sige, at modellen foretager en for "grovkornet" sortering af borgere i træningsdatasættet – man taler også om, at den er **undertilpasset** datasættet. Høj varians vil omvendt sige, at modellen foretager en for "finkornet" sortering af borgere i træningsdatasættet – her taler man om, at modellen er **overtilpasset** træningsdatasættet. Begge dele vil have indflydelse på modellens fejlrate.

Den konkrete tilpasning af modellen – dvs. den rette balance mellem bias og varians – er derfor en væsentlig udviklingsopgave. Skønsregler eller regler, der giver et vist fortolkningsrum, gør det i særlig grad vanskeligt at tilpasse modellen korrekt. Der er nemlig grænser for, hvor mange undtagelser eller variationer modellen kan tage hensyn til, uden at det får betydning for fejlraten. Hvis modellen tilpasses for meget i forhold til varians, kan det føre til en overtilpasning, hvor undtagelsesvise eller tilfældige forhold indgår i den generelle "regeldannelse" for modellen og derfor fylder for meget. På et generelt niveau skyldes det, at en læringsalgoritme reducerer bias ved at lave en mere finkornet repræsentation af sammenhænge mellem variable og målegenskab. Dette inkluderer uundgåeligt både sammenhænge som kan generaliseres, og sammenhænge som kun optræder tilfældigt. Derved øges modellens varians. Omvendt reduceres varians ved at få modellen til at lave en mere grovkornet repræsentation, hvor kun de væsentligste sammenhænge mellem variable og målegenskaben anvendes af modellen.

I den ene ekstrem kan modellen komme til at lave en sortering, der behandler hver eneste person som unik givet netop den pågældende persons kvaliteter. Men en sådan overtilpasning betyder, at modellen vil have meget vanskeligt ved at generalisere til andre datasæt, fordi den er baseret på sammenhænge, som alene optræder i sættet af træningsdata. Den vil derfor fejlklassificere mange personer, når den tages i brug.

I den anden ekstrem kan en model sortere på baggrund af ganske få oplysninger om borgeren. Det er overordentlig sandsynligt, at denne sammenhæng kan generaliseres, men en sådan grov undertilpasning betyder, at modellen i praksis fejlklassificerer mange personer, fordi den ikke tager højde for de mange andre relevante sammenhænge.

Som vi nævnte indledningsvist, vil myndighederne ofte skulle foretage individuelle og konkrete vurderinger af borgernes forhold. Dette vil indebære, at der fra sag til

sag kan indgå forskellige omstændigheder med forskellig vægt i myndighedernes afgørelse.

F.eks. skal der i en regel som servicelovens § 50 om børnefaglige undersøgelser foretages en vurdering af både hvilke forhold, der skal inddrages i undersøgelsen (retsfaktum), og hvilke foranstaltninger der eventuelt skal iværksættes (retsfølge). Hvad angår det første (retsfaktum), skal kommunalbestyrelsen som udgangspunkt anlægge en helhedsbetragtning af, hvad der i det konkrete tilfælde anses for relevante forhold som udvikling og adfærd, familie, skole, sundhed, fritid, venskaber og andre relevante forhold. Hvad angår det andet (retsfølge), skal undersøgelsen resultere i en begrundet stillingtagen til, om der er grundlag for at iværksætte foranstaltninger, og i bekræftende fald af hvilken art.

Saglighedskravet er tæt beslægtet med den forvaltningsretlige ligheds-grundsætning, der lidt forenklet indebærer, at offentlige myndigheder er forpligtede til at anvende en ensartet fortolkning af loven og til at sikre, at lige tilfælde behandles lige og forskellige tilfælde forskelligt (lighedsgrundsætningen har et overlap med diskriminationsforbuddet behandlet i kapitel 8).[12]

Denne balance kan være noget nær umulig at tilpasse i modellen når myndigheden er forpligtet til at foretage en vurdering som i servicelovens bestemmelse.

Alle modeller vil desuden have en tendens til at begå flere fejl i atypiske sager. Atypiciteten kan både have betydning i forhold til hvilke variable, der indgår i modellen og i vægtningen af variable. Atypiske sager kan vedrøre variable, som modellen ikke vægter nok eller korrekt i forhold til deres reelle betydning i den konkrete sag, ligesom det per definition kan være svært eller umuligt på forhånd at vide, hvilken sag, der er atypisk og hvorfor.

Også disse forhold bør derfor efter vores vurdering sætte grænser for, hvorledes profileringsmodellerne bruges af myndigheden.

ULOGISKE SAMMENHÆNGE

Den tredie problemstilling, som fører til vores anbefalinger om begrænsningen i fuldautomatiseret brug og krav til AI-konsekvensanalyser, vedrører **maskinlæringens** ulogiske karakter. Et af de værdifulde elementer ved maskinlæring er, at maskinlærte modeller er i stand til at foretage avanceret statistisk analyse. De behandler alle sammenhænge ens, uanset om det er en sammenhæng, der giver mening for mennesker eller ikke, og dermed kan de også finde statistisk pålidelige sammenhænge, som for mennesker kan virke tilfældige og ulogiske.[13]

De ulogiske sammenhænge skaber udfordringer for centrale forvaltningsretlige regler og principper:

Hensynet til forudsigelighed forstået som borgerens mulighed for at indrette sig efter lovens regler og forstå, hvorledes vedkommende vil blive vurderet og bedømt efter loven mindskes, når modellernes sammenhænge ikke er forståelige. Ligeledes udfordres adgangen til en begrundelse for afgørelsen og en efterfølgende adgang til at efterprøve afgørelsen. Afgørelserne mister således en vis "kontrollerbarhed".

Som nævnt vil myndighedens afgørelser ganske ofte basere sig på skøn eller fortolkning. Det er kendetegnende for disse typer af afgørelser, at der ikke nødvendigvis er ét rigtigt resultat. Retssikkerheden i disse tilfælde ligger ikke nødvendigvis kun i hvad afgørelsens resultat er men også i måden, hvorpå man nåede frem til resultatet.[14]

Overvejelserne her har tæt tilknytning til spørgsmålet om transparens i kapitel 7, herunder borgerens krav på en begrundelse for en bebyrdende afgørelse. Det adskiller sig imidlertid derfra ved, at ulogiske sammenhænge ikke nødvendigvis er uigennemsigtige i dén forstand, som vi behandler i kapitlet om transparens – sammenhængene kan efter omstændighederne sagtens identificeres, det er bare svært for mennesker at forstå, hvorfor der overhovedet er en sammenhæng.[15]

Er – og bør – sådanne sammenhænge anses for lovlige og saglige? Der kan næppe gives et entydigt svar på dette. En væsentlig fordel ved maskinlæring er netop, at den på avanceret vis kan finde interaktioner mellem variable, som mennesker ikke er i stand til at finde.

Så længe modellen behandler personoplysninger indsamlet i overensstemmelse med databeskyttelsesforordningens principper om dataminimering, formålsbegrænsning og rigtighed (se kapitel 5), og så længe myndighederne i sager baseret på fortolkning eller skøn udelukkende anvender modellerne til beslutningsstøtte, kan modellerne tjene til at oplyse og støtte sager også gennem resultater, der baserer sig på ulogiske sammenhænge. De retssikkerhedsmæssige betænkeligheder sætter naturligvis høje krav til myndighederne, både, når de skal identificere nødvendigt inputdata, som vi har beskrevet i kapitel 5 og når de skal bruge modellens output, som vi beskæftiger os med i dette kapitel.

Vores anbefalinger om konsekvensanalyser (se kapitel 4), om krav til inputdata (se kapitel 5), om uigennemsigtighed (se kapitel 7) og om rammerne for brugen af outputtet (dette kapitel) tjener således alle til at sikre, at maskinlæringens måske mest kendetegnende egenskab bruges inden for rammerne af, hvad der er retssikkerhedsmæssigt forsvarlig.

UDFORDRINGER I ANVENDELSESFASEN

Vi har frem til nu beskæftiget os med problemstillinger, der vedrører modellens design og udvikling. Der findes imidlertid også en række udfordringer, som sjældent eller slet ikke kan adresseres under udviklingen af modellen, men som

opstår, når modellen sættes i drift. Det er derfor en central pointe, at rettighederne skal sikres i hele modellens livscyklus.

Som de forudgående afsnit illustrer, er der efter vores opfattelse væsentlig forskel på, hvordan en model skal vurderes retssikkerhedsmæssigt alt afhængig af, om den bruges til fuldautomatisering af en afgørelse eller til beslutningsstøtte.

I Danmark bruges modeller overvejende til beslutningsstøtte (se kapitel 4). I denne sammenhæng stiller vi i det følgende to væsentlige krav i anvendelsen af en model til beslutningsstøtte:

Vi ser først på myndighedens pligt til at sikre, at der i den enkelte borgers sag er indgået relevante oplysninger om vedkommendes forhold før der træffes en afgørelse ved at sikre partshøring af borgeren. Dernæst ser vi på, hvorledes myndigheden skal sikre sig mod såkaldt automatiseringsbias, hvor modellen får en utilsigtet stor indflydelse på den endelige afgørelse og bliver beslutningsstyrende frem for beslutningsstøttende.

TILSTRÆKKELIG SAGSOPLYSNING OG REEL PARTSHØRING

En myndighed må kun træffe en afgørelse, efter at myndigheden har sikret, at sagen er tilstrækkeligt oplyst (det såkaldte officialprincip). Et væsentligt led i vurderingen af dette er, om borgeren er blevet inddraget i arbejdet med sagsoplysning. Dette er reguleret gennem de forvaltningsretlige regler om partshøring, der indebærer, at borgeren aktivt involveres i oplysningen af sagen. Partshøring udgør en nødvendig og vigtig del af myndighedens pligt til at oplyse sagen og sikre, at forvaltningens afgørelser træffes på et korrekt faktisk grundlag.

Kravet er nært forbundet med reglerne om aktindsigt og oplysningspligt, som vi beskriver i kapitel 7, men adskiller sig fra disse ved, at regler om partshøring også omhandler borgerens ret til aktiv deltagelse i sin egen sag (retten til kontradiktion) og som nævnt myndighedens forpligtelse til at oplyse sagen (officialprincippet).[16] Det handler således ikke kun om at give borgeren indsigt, men også om at sikre, at borgeren har indflydelse på sin sag og at myndigheden inddrager alle relevante kriterier i sagen.

Ombudsmanden har fremhævet, at der er en tendens til, at myndigheder ikke er "opmærksomme på retten til partshøring, hvor en IT-løsning indebærer, at der som led i behandlingen af en afgørelsessag automatisk indhentes oplysninger fra registre og tidligere sager." [17]

Ombudsmanden har videre anført, at:

> "Automatiserede afgørelsessystemer forudsætter, at myndighederne på forhånd kan afgrænse, hvilket faktum der måtte være relevant i fremtidige sager. Når sagsbehandlingen automatiseres, bliver den faktiske del af beslutningsgrundlaget således ikke længere genstand for myndighedens

> konkrete vurdering. Sagsoplysningen bliver i stedet til en generel model for det typisk forekommende. Modellen danner grundlag for de oplysninger, borgeren selv skal udfylde, og som myndigheden eventuelt supplerer gennem samkøring med databaser. Når myndigheder anvender automatiserede systemer, indebærer det derfor en risiko for, at de ikke fanger tilfælde med atypiske fakta, eller tilfælde, hvor retsgrundlaget forpligter dem til at tage hensyn til forhold, der enten normalt ikke opstår eller normalt ikke er relevante."[18]

Manglende partshøring kan føre til, at der træffes en forkert afgørelse, fordi grundlaget for afgørelsen er fejlagtigt eller utilstrækkeligt. Hvis partshøringen ikke gennemføres, kan det føre til afgørelsens ugyldighed og til, at myndigheden må behandle sagen på ny.

Det er imidlertid vigtigt, at der er tale om en reel og ikke en illusorisk partshøring. Det skal derfor sikres, at partshøringen faktisk kan ændre udfaldet, og at modellens profilering af borgeren samt borgerens egne oplysninger og synspunkter begge indgår i fornøden grad i oplysning af sagen.

Vi anbefaler, at der i vejledningen (se kapitel 4) fastsættes krav, der sikrer en reel partshøring ved anvendelsen af profileringsmodeller til fuldautomatisering og til beslutningsstøtte, herunder, at vejledningen fastsætter, hvornår og hvordan borgeren inddrages og hvordan input fra borgeren indgår i sagen.

SIKRING MOD AUTOMATISERINGSBIAS

Når offentlige myndigheder anvender profileringsmodeller, kan der opstå risiko for automatiseringsbias. Automatiseringsbias er en psykologisk tendens til at have for høj tillid til automatiserede beslutningssystemer, og at foretage hurtigere, mindre grundige vurderinger af en opgave, som et sådant system har behandlet, end man ellers ville.[19]

Det overordnede spørgsmål, som risikoen for automatiseringsbias rejser er, hvilken rolle og vægt modellens resultater skal gives i myndighedernes afgørelser. Udfordringen består konkret i, at modellens resultat kommer til at spille en større rolle, eller få en anden vægt, end den ideelt set burde.

Der er naturligvis intet i vejen med, at en myndighed eller den enkelte sagsbehandler lægger vægt på profileringsmodellens resultat, når formålet netop er at oplyse og støtte den endelige beslutning. Problemet opstår, hvis myndigheden eller den enkelte sagsbehandler tillægger modellen mere vægt end tilsigtet, og afgørelsen derfor går fra at anvende modellen til automatiseret beslutningsstøtte til fuldautomatisering, hvor den endelige beslutning overlades til modellen.

Det afgørende for sondringen mellem disse to typer af brug af modellen – beslutningsstøtte eller fuldautomatisering – er netop, om det er sagsbehandleren eller modellen, der træffer den endelige forvaltningsretlige afgørelse. Risikoen for automatiseringsbias viser imidlertid, at sondringen ikke altid er klar.

I databeskyttelsesforordningen er sondringen central, da forordningen fastsætter regler om afgørelser, "der alene er baseret på automatisk behandling"[20] (dvs. fuldautomatiserede afgørelser). Sondringen i forordningen må forventes at have betydning for, hvordan myndighederne også i øvrige dele af forvaltningens arbejde vil sondre mellem automatiseret beslutningsstøtte og fuldautomatisering.

Den såkaldte Artikel 29-gruppe, hvis vejledninger udgør et fortolkningsbidrag til forordningen (se kapitel 2) forklarer, at i tilfælde, hvor en person træffer en afgørelse på baggrund af et automatiseret hjælpeværktøj, vil behandlingen ikke være fuldautomatiseret. I de tilfælde, hvor modellen beslutter, hvad retsstillingen for den registreredes skal være "uden nogen forudgående og meningsfuld menneskelig vurdering" vil afgørelsen være fuldautomatiseret.[21] Artikel 29-gruppen tilføjer:

> "For at kunne betragtes som menneskelig indgriben skal den dataansvarlige sikre, at et tilsyn med afgørelsen er meningsfuldt og ikke blot en tom gestus. Det bør foretages af en person, der har den fornødne kompetence og mulighed for at ændre afgørelsen. Der bør som led i analysen tages hensyn til alle relevante oplysninger"[22]

Det er ikke afklaret i praksis eller litteratur, hvad der skal anses for en "meningsfuld" menneskelig vurdering.

Som noget særligt er der i databeskyttelsesforordningens en adgang til efterfølgende "menneskelig indgriben".[23] Denne efterfølgende menneskelige indgriben ændrer ikke ved om en afgørelse er fuldautomatiseret, men er en efterfølgende klageadgang.[24] Ligeledes indebærer et løbende tilsyn (se kapitel 4) ikke, at en fuldautomatiseret afgørelse ændrer karakter til beslutningsstøtte (se boks).

SAMSPIL MELLEM MENNESKER OG MASKINER: TRE KATEGORIER

Samspillet mellem mennesker og maskiner er af afgørende betydning for at sikre, at rettigheder ikke forringes i den faktiske anvendelse af en profileringsmodel.[25] Menneskeligt kontrol har til formål at sikre tilsyn med modellen og dens output og kan opdeles i tre kategorier[26]:

- **Human-in-the-loop:** Menneskelig indgriben i hver af modellens afgørelser. Dette svarer efter vores vurdering til modellens brug som beslutningsstøtte.
- **Human-on-the-loop:** Menneskelig indgriben i modellens designcyklus og overvågning af modellens anvendelse uden indgriben i hver enkelt afgørelse. Dette kan f.eks. være intern kvalitetskontrol og svarer efter vores vurdering til fuldautomatiseret brug af modellen.
- **Human-in-command**: Mulighed for at kontrollere modellens overordnede aktivitet herunder dens bredere økonomiske, samfundsmæssige, juridiske og etiske indvirkning og muligheden for at afgøre, hvornår og hvordan modellen skal anvendes i en bestemt situation. Dette svarer efter vores vurdering til en fuldautomatiseret afgørelse med enten intern eller ekstern kontrol f.eks. i form af et generelt tilsyn (se kapitel 4)

Automatiseringsbias er altså en måde, hvorpå en myndighed eller sagsbehandler kan tillægge modellens resultater for stor vægt og dermed foretage mindre grundige vurderinger af de sager, modellen behandler.

To underkategorier for automatiseringsbias er konfirmationsbias og ankereffekter. De knytter sig til det tidspunkt i processen, hvor modellens vurdering inddrages i myndighedens arbejde og rejser potentielt forskellige udfordringer.

Konfirmationsbias kan opstå, hvis sagsbehandleren vurderer sagen og først efterfølgende modtager modellens vurdering, og herefter justerer sin første vurdering i lyset af denne. Ankereffekter kan opstå, hvis sagsbehandleren modtager modellens vurdering inden vedkommende har vurderet sagen, og herefter behandler modellens vurdering på linje med andre forhold.

Det modsatte af automatiseringsbias er såkaldt aversionsbias, hvor modellens vurdering tillægges for lidt vægt.[27]

Automatiseringsbias kan mindskes ved systematisk test for bias (f.eks. ved at lave kontrolgrupper), og ved at etablere procedurer for anvendelse af vurderingen som minimerer den potentielle indflydelse af bias.

Myndigheden skal desuden gøre det klart, hvordan sagsbehandleren skal forstå og anvende modellens output, og skal sikre mod utilsigtet eller misforstået brug af modellens resultat

Det er desuden vigtigt, at sagsbehandleren har de nødvendige informationer om modellens vurdering af borgerens konkrete sag og et overblik over, hvordan modellen har vurderet andre sager, ligesom sagsbehandleren skal have kendskab til de centrale forhold omkring profileringsmodellen og dens generering af et output. Dette stiller krav til transparens (se kapitel 7).

I EU-Kommissionens udkast til en forordning for kunstig intelligens (se kapitel 2) er der en bestemmelse, som skal sikre gennemsigtighed og formidling af oplysninger til bl.a. sagsbehandlere, der anvender profileringsmodeller. Herudover følger af bestemmelsen, at der skal udarbejdes "brugsanvisninger" til sagsbehandleren, der indeholder oplysninger om modellens egenskaber, kapacitet og begrænsninger for modellens ydeevne, informationer om hvorvidt modellen er testet og valideret og oplysninger i forhold til risici for borgeres rettigheder.[28]

I udkastet til forordningen er der tillige en udtrykkelig bestemmelse om menneskeligt tilsyn med systemet.[29] I bestemmelsen anføres, at modellerne skal udformes med "passende menneske-maskine-grænsefladeværktøjer", således at mennesker kan føre effektiv kontrol med modellen. Modellen skal være udformet således, at mennesker, der fører kontrol kan forstå modellens kapacitet og begrænsninger fuldt ud og være opmærksomme på den mulige tendens til automatiseringsbias og kunne beslutte ikke at anvende modellen eller på anden måde se bort fra, tilsidesætte eller omgøre output fra modellen i enhver given situation

I udkastet er der endvidere fastsat regler om, at modellerne skal være modstandsdygtige over for fejl, svigt og uoverensstemmelser, der kan forekomme, herunder også på grund af modellernes interaktion med mennesker eller andre systemer.[30] Graden af nøjagtighed og parametrene som det måles på skal gøres tilgængelig for sagsbehandlerne, der anvender profileringsmodellen.[31]

Der er således forskellige måder, hvorpå automatiseringsbias kan imødegås og vi anbefaler, at Justitsministeriet i vejledningen (se kapitel 4) forholder sig til udfordringen og dens mulige løsninger. Dette kan bl.a. gøres med afsæt i udkastets regler. Herudover tjener vores anbefalinger om teknisk og rettighedsmæssig forståelse hos bl.a. sagsbehandleren (kapitel 4) samt krav til systemisk og algoritmisk transparens (se kapitel 7) som vigtige elementer for at imødegå risikoen for automatiseringsbias.

Vi anbefaler, at vejledningen (se kapitel 4) adresserer risikoen for, at modellen får en uforholdsmæssig stor indflydelse på den endelige afgørelse og bliver beslutningsstyrende frem for beslutningsstøttende.

NOTER

1 Grgić-Hlača, Gummadi, Redmiles og Weller (2018), Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction
2 Fenger (2018), Forvaltningsret, Djøf forlag, s. 316
3 Digitaliseringsstyrelsens vejledning om digitaliseringsklar lovgivning (maj 2018) s. 12 tilgængelig på: https://digst.dk/media/16953/vejledning_om_digitaliseringsklar_-lovgivning_maj_2018_tg.pdf
4 Se Folketingets Ombudsmand, overblik # 12, Partsrettigheder og offentlige IT-systemer, tilgængeligt på: http://www.ombudsmanden.dk/myndighedsguiden/specifikke_sagsomraader/partsrettigheder_og_offentlige_it-systemer_/
5 Digitaliseringsstyrelsens vejledning om digitaliseringsklar lovgivning (maj 2018) s. 11 tilgængelig på: https://digst.dk/media/16953/vejledning_om_digitaliseringsklar_-lovgivning_maj_2018_tg.pdf
6 Digitaliseringsstyrelsens vejledning om digitaliseringsklar lovgivning (maj 2018), s. 12 tilgængelig på: https://digst.dk/media/16953/vejledning_om_digitaliseringsklar_-lovgivning_maj_2018_tg.pdf
7 Se Gøtze, (2018). Digital reform af dansk lovgivningskultur. Juristen, (5-6), 182-190
8 Databeskyttelsesforordningen 2016/679 af 27. april 2016, artikel 25
9 Se Motzfeldt (2017), The Danish Principle of Administrative Law by Design 23, European Public Law, Issue 4, s. 739-754
10 Fenger (2018), Forvaltningsret, Djøf forlag, s. 316f.
11 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final, artikel 15
12 Fenger (2018), Forvaltningsret, Jurist- og Økonomforbundets Forlag, s. 347
13 Selbst og Barocas (2018), The Intuitive Appeal of Explainable Machines 87 Fordham Law Review 1085
14 Binns (2018), Human Judgement in Algorithmic Loops; Individual Justice and Automated Decision-Making
15 Selbst og Barocas (2018), The Intuitive Appeal of Explainable Machines 87 Fordham Law Review 1085
16 Fenger (2018), Forvaltningsret, Jurist- og Økonomforbundets Forlag, s. 605
17 Se ombudsmandens notat af 2. juli 2018 om forvaltningsretlige krav til det offentliges IT-løsninger, som er tilgængeligt på: https://www.ombudsmanden.dk/myndighedsguiden/specifikke_sagsomraader/generelle_forvaltningsretlige_krav_til_offentlige_it-systemer/2019/ Se også Justitsministeriets notat af 2015 om forvaltningsretlige krav til offentlige

digitale løsninger s. 7 tilgængeligt på https://digst.dk/media/12721/notat-om-forvaltningsretlige-krav-til-det-offentliges-itlsninger.pdf

18 Folketingets Ombudsmand, overblik # 12, Partsrettigheder og offentlige IT-systemer tilgængeligt på: http://www.ombudsmanden.dk/myndighedsguiden/specifikke_sagsomraader/partsrettigheder_og_offentlige_it-systemer_/

19 Goddard, Roudsari og Wyatt (2011) Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association 19(1): 121-127 samt Lyell og Coiera (2017) Automation bias and verification complexity: a systematic review. Ibid. 24(2): 423-431

20 Databeskyttelsesforordningen 2016/679 af 27. april 2016 artikel 22

21 Se Artikel 29-gruppens retningslinjer om automatiske individuelle afgørelser og profilering, s. 9 og s. 21f.

22 Se Artikel 29-gruppens retningslinjer om automatiske individuelle afgørelser og profilering, s. 22

23 Databeskyttelsesforordningen 2016/679 af 27. april 2016 artikel 22, stk. 3

24 Se Artikel 29-gruppens retningslinjer om automatiske individuelle afgørelser og profilering, s. 34

25 Cobbe, Lee, Singh (2021) Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems, arXiv:2102.04201

26 Den Uafhængige Ekspertgruppe på Højt Niveau om Kunstig Intelligens (2019), Etiske Retningslinjer for Pålidelig Kunstig Intelligens, s. 19, tilgængelige på: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

27 Se om denne form for bias f.eks. Dietvorst, Simmons og Massey (2015), Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. Journal of Experimental Psychology: General 144(1): 114-126, Prahl og Van Swol (2017), Understanding algorithm aversion: When is advice from automation discounted? Journal of Forecasting 36(6): 691-702, Dietvorst, Simmons og Massey (2018), Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. Management Science 64(3): 1155-1170, Burton, Stein og Jensen (2020). A systematic review of algorithm aversion in augmented decision making, 33(2): 220-239

28 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final, artikel 13

29 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final, artikel 14

30 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final, artikel 15

31 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final betragtning nr. 49

DEL III

SÆRLIGE UDFORDRINGER

I denne del af rapporten ser vi på to særlige udfordringer ved profileringsmodeller.

Den første udfordring omhandler mulighed for at få indsigt i og om modellen. Både forvaltningsretlige og databeskyttelsesretlige regler og principper indebærer, at myndighederne skal være åbne og transparente, men princippet udfordres når sagsbehandlingen helt eller delvist udføres af en profileringsmodel. Dette stiller krav til algoritmisk og systemisk transparens svarende til transparens i modellen og transparens om modellen. Den problematik behandler vi i kapitel 7.

Den anden udfordring omhandler risikoen for, at modellerne viderefører eller forstærker ulovlig forskelsbehandling, fordi data som køn, etnicitet, alder og handicap direkte eller indirekte kommer til at indgå i de avancerede dataanalyser, som modellerne foretager. Fordi der er tale om komplekse analyser, kan det være svært at opdage diskriminationen for både myndigheder, den enkelte borger, en eventuel tilsynsmyndighed og domstolene. Hertil kommer, at modellerne ikke altid er i stand til at varetage beskyttelseshensynet til fulde, eller gør det på bekostning af andre lige så væsentlige beskyttelseshensyn. De forhold ser vi nærmere på i kapitel 8.

KAPITEL 7

TRANSPARENS I OG OM PROFILERINGSMODELLER

Når myndigheder i tiltagende grad bruger profileringsmodeller i sagsbehandlingen, udfordrer det borgerens og offentlighedens mulighed for at få indblik i myndighedens behandling af sager. Det skyldes, at modellerne kan være uigennemsigtige på forskellige måder, rækkevidden af de gældende regler er uklar og kun dækker fragmentarisk, og at de (med en enkelt undtagelse) ikke specifikt regulerer transparens i og om profileringsmodellerne. I dette kapitel præsenterer vi to typer af transparens, algoritmisk og systemisk og gennemgår fire områder, hvor gældende regler i særlig grad er uklare og mangelfulde. Endelig fremsætter vi en række en række anbefalinger, der skal skabe en øget transparens i og om modellerne og deres brug.

Transparens er ikke et entydigt begreb. Det kan forstås på mange måder, og vi vil i dette kapitel beskrive, hvilke former for transparens, der efter vores opfattelse kræves i og om modeller. I det følgende bruger vi ordet "transparens", ikke som et juridisk begreb, men som samlebetegnelse for de mange måder, hvorpå borgeren, offentligheden eller tilsyns- og kontrolinstanser kan få indblik i profileringsmodellen og dens anvendelse i sagsbehandlingen.

HVAD ER TRANSPARENS?

Krav om transparens afspejler sig i en række regler både inden for data-beskyttelsesretten og forvaltningsretten. Fælles for reglerne er, at de udgør kontrolmekanismer, som gør det muligt for en aktør – en borger, et tilsyn, en domstol eller offentligheden mere generelt – at udøve kontrol over myndigheden.[1]

Transparens eller åbenhed som forvaltningsretligt princip indebærer således, at myndighederne stilles til ansvar for deres handlinger og kan kigges efter i sømmene. Princippet udmønter sig bl.a. i regler om aktindsigt og begrundelsespligt for en afgørelse.

Databeskyttelsesforordningen har et centralt princip om "gennemsigtighed", der bl.a. udmønter sig i oplysningsforpligtelser for myndighederne og indsigtsret for borgerne. Oplysningsforpligtelse og indsigtsret er bestemte måder, hvorpå borgerne kan få indsigt i, hvilke personoplysninger, der behandles om dem. Som noget særligt regulerer databeskyttelsesforordningen også fuldautomatiseret brug af profileringsmodeller.

Alle reglerne tjener på hver deres måde og inden for hvert deres anvendelses-område hensynet til transparens.

For at alle regler og principper kan iagttages til fulde, også når myndighederne benytter sig af profileringsmodeller, er der efter vores vurdering behov for regler om transparens på to niveauer. For det første transparens i modellen, dvs. adgang til information om, hvordan modellen er nået frem til et resultat. For det andet transparens om modellen, dvs. adgang til information om dens kvalitet, træning, anvendelse og tilsyn. Vi kalder disse to niveauer for henholdsvis algoritmisk og systemisk transparens (se boks).

TO FORMER FOR TRANSPARENS FOR PROFILERINGSMODELLER

Algoritmisk transparens: Adgang til information om, hvordan modellen træffer sine beslutninger. F.eks.:

- Har modellen lagt et forhold til grund, som i det konkrete tilfælde ikke er sagligt, eller er bestemte kriterier blevet vægtet forkert eller fejlagtigt undladt? Dette kræver adgang til modellens **variable** og deres effekt på modellens resultat. Dette er nødvendigt, hvis man vil have svar på, om modellen overholder saglighedskravet (se kapitel 6)
- Hvilken rolle spiller en bestemt variabel f.eks. køn eller etnicitet eller et andet beskyttet kendetegn, i modellen enten i profileringen af en konkret borger eller for alle borgere? Dette er nødvendigt hvis man vil undersøge, om modellen diskriminerer direkte (se kapitel 8)
- Hvordan virker modellen som et hele for personer med forskellige variabelværdier i forhold til det beskyttede kendetegn f.eks. køn eller etnicitet? Dette er nødvendigt, hvis man vil undersøge, om modellen diskriminerer indirekte (se kapitel 8)

Systemisk transparens: Adgang til informationer om modellens kvalitet, træning og anvendelse. F.eks.:

- Har myndigheden truffet afgørelsen på det rette oplysningsgrundlag? Dette kræver adgang til information om, hvorvidt der er fejl i værdierne for relevante variable (en borger er opført som fraskilt, men er fortsat gift, eller som arbejdsløs, selvom vedkommende er i arbejde) – dvs. information om de konkrete data om borgeren, som modellen har anvendt (kapitel 5).
- Er der automatiseringsbias i myndighedens brug af modellens resultat? Kræver adgang til information om, hvordan modellens resultatet indgår i forvaltningens afgørelse, og om denne anvendelse svarer til den, der var tiltænkt (se kapitel 6).

ALGORITMISK TRANSPARENS

Algoritmisk transparens dækker over adgangen til information om, hvordan en model træffer sine vurderinger. I faglitteraturen opdeler man denne adgang i muligheden for enten at kunne **fortolke** eller **forklare** modellen – henholdsvis modellens "interpretability" eller "explainability".[2]

Modeller, der kan fortolkes, er umiddelbart transparente i deres design, uden at man har behov for tekniske redskaber for at få indblik i modellen. Det indebærer typisk, at man kan forstå årsagerne til og pålideligt forudsige modellens vurderinger.

Modeller kan imidlertid være så komplekse, at selv personer med teknisk fagkundskab ikke kan fortolke dem. I så fald taler man om, at modellen i stedet skal kunne forklares, og metoder til dette er et centralt tema i forskningen i maskinlæring. En forklaring er her en beskrivelse af, hvordan hele eller dele af modellen fungerer, som hjælper personer med at forstå modellen. Forklaringen dannes ud fra forskellige tekniske metoder til at analysere modellen, såkaldt xAI.[3]

I faglitteraturen taler man også om, at et vist niveau af **domænekendskab** er nødvendigt for at forstå visse typer af informationer.[4] Domænekendskab dækker over kendskab til f.eks. maskinlæringsmetoder, modeltyper, modelkvalitet og fejltyper, som kan være nødvendige for at omsætte data om modellen til forståelse af modellen. F.eks. kan det være meget vanskeligt eller endda umuligt for en person at forstå en model, selv hvis myndighederne giver adgang til modellens funktion, inklusive vægte og definitioner af variable, hvis personen ikke ved, hvordan variable, vægte og funktion tilsammen fører til modellens vurderinger.

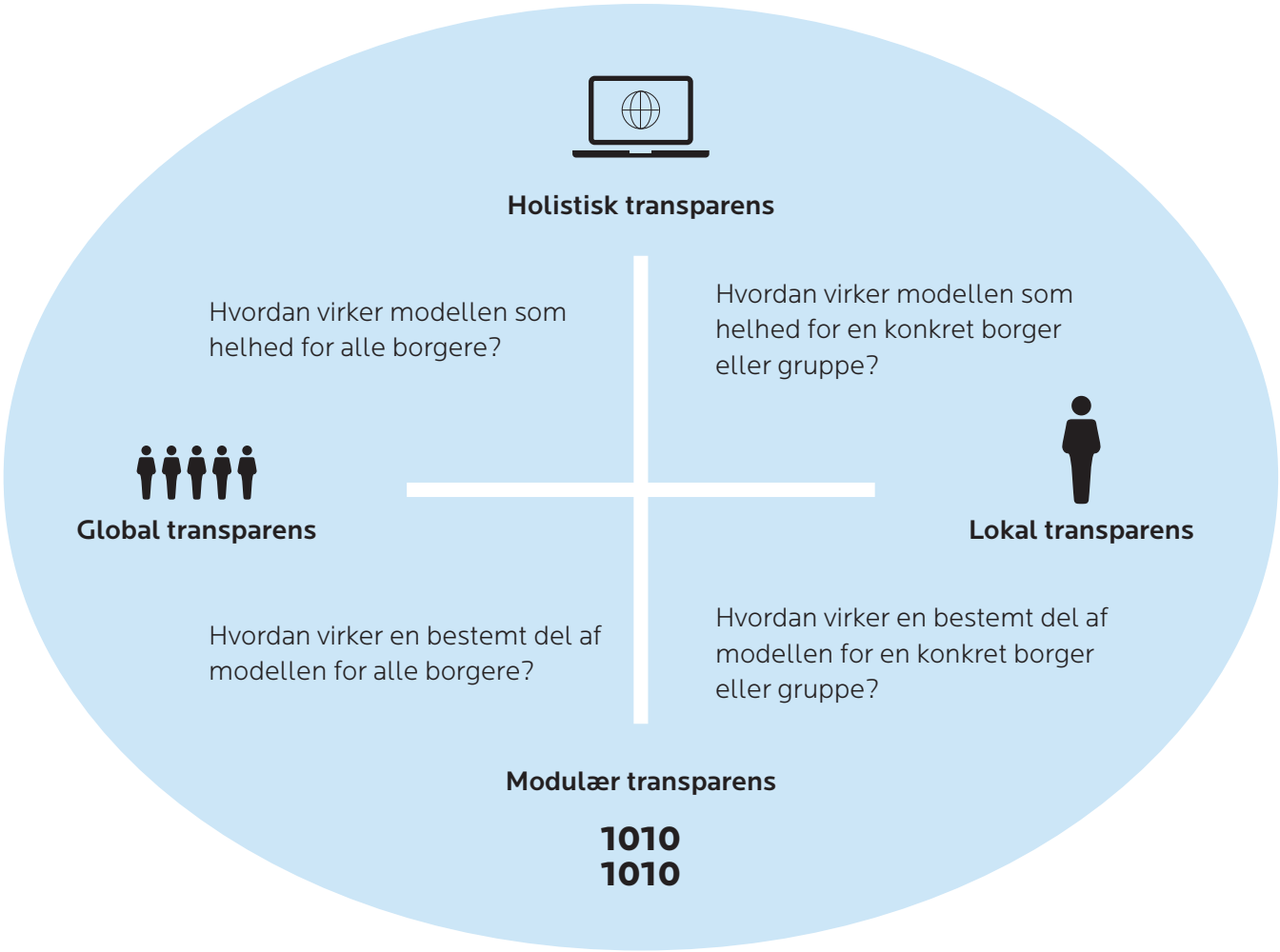
Forklaringer bliver relevante, når modeller stiger i kompleksitet. Lidt forenklet kan man sige, at jo mere kompleks en model er, jo mindre transparent kan den risikere at være, og dermed vanskeligere at fortolke. Mere komplekse modeller vil kunne udføre sofistikerede analyser og dermed oftest have en lavere **fejlrate** end modeller, der er transparente, men simple. Bruger man uigennemsigtige og komplekse modeller, som alt andet lige lave færre fejl, skal man bruge tekniske metoder for at forklare modellerne. Valget mellem modeller, der kan fortolkes og modeller, der kan forklares, ser vi på til slut i kapitlet.

Vi bruger algoritmisk transparens som en samlebetegnelse for modeller, der enten kan fortolkes eller forklares. Det gælder dog for disse begreber – ligesom for vores opdeling i algoritmisk og systemisk transparens – at der (endnu) ikke er enighed om definitionerne. Det engelske begreb "explainable" bruges således både til den særlige type af algoritmisk transparens, som kan opnås gennem tekniske metoder og som også kan benævnes xAI men også som et samlebegreb for flere typer af transparens.[5]

Udover en skelnen mellem forklaring og fortolkning som måder, hvorpå transparens kan opnås, skelner man i faglitteraturen mellem forskellige niveauer af algoritmisk transparens: lokal, global, holistisk og modulær transparens (se figur).[6]

Lokal transparens giver indblik i, hvordan modellen behandler en konkret borger eller en begrænset gruppe borgere, mens global transparens giver indblik i, hvordan en model behandler alle borgere, som modellen anvendes på. På samme måde giver modulær transparens indblik i, hvordan en del af modellen fungerer, mens holistisk transparens giver indblik i hele modellen.

NIVEAUER AF ALGORITMISK TRANSPARENS



Som figuren illustrerer, kan de forskellige niveauer kombineres: Global, holistisk transparens giver typisk information om hele modellen, og forudsætter at man kan forstå, hvordan alle modellens forskellige dele spiller sammen for alle de personer, som modellen vurderer. Hvis man kan opnå denne form for transparens,

er de andre niveauer irrelevante, da man ved global holistisk transparens får et fuldkomment indblik i modellen. Omvendt kræver lokal, modulær transparens blot, at man har adgang til den relevante del af modellen, samt at man kan forstå, hvordan netop denne del af modellen påvirker vurderingen af en konkret person.

Lokal, holistisk transparens giver adgang til informationer om, hvordan modellen behandler netop den enkelte person eller gruppe, mens global, modulær transparens giver informationer om, hvordan en bestemt del af modellen, f.eks. en bestemt variabel virker i modellen for alle personer.

Global transparens kan være nødvendigt for f.eks. at konstatere direkte diskrimination, men det kan også være for at vurdere om modellen tillægger et forhold urimelig vægt (f.eks. "bør uddannelsesniveau virkelig spille så stor en rolle for myndighedens vurdering af, om en familie er i risiko for at børn mistrives?"). Lokal transparens vil oftest være relevant for at den enkelte borger kan få en begrundelse for den konkrete afgørelse.

Det er oftere lettere at få indblik i modellen på lokalt niveau end globalt niveau. Det skyldes, at det ofte ikke er alle dele af modellen, som er relevante for vurderingen af en konkret person og det er derfor ikke nødvendigt at forstå modellens vurderinger i alle sager, hvor den er blevet brugt.

Det kan være svært at opnå transparens på et holistisk niveau. En sådan indsigt i modellen forudsætter nemlig, at en person er i stand til at begribe ikke blot de enkelte dele af modellen, men også hvordan de hænger sammen, påvirker hinanden, og tilsammen fører til modellens vurderinger. Når en models kompleksitet stiger bliver denne type indsigt hurtigt vanskelig eller endda umulig, fordi det er menneskeligt umuligt at overskue alle modellens dele samtidig.[7]

Det er typisk lettere at opnå modulær transparens men også her kan det være udfordrende at få et indblik i modellen. Det skyldes to forhold:

For det første må de enkelte deles funktion ofte forstås i sammenhæng med resten af modellen. F.eks. kan effekten af et beslutningspunkt i et **beslutningstræ** (se kapitel 3), som anvender en bestemt variabel, afhænge af hvor i træet det pågældende punkt optræder, og hvilke variable de andre beslutningspunkter anvender. Tilsvarende kan en variabel med en given vægt i en **lineær regression** (se kapitel 3) have forskellig betydning for, hvordan modellen vurderer personer, afhængigt af hvilke **vægte** de andre variable har, og hvordan værdierne for de forskellige variable normalt fordeles. En variabel med høj vægt kan have relativt lille betydning, hvis værdierne for den variable typisk er lave, eller hvis andre variable har endnu højere vægte.[8]

For det andet kan de relevante dele i sig selv være komplekse. F.eks. kan den samme variabel i et beslutningstræ optræde med forskellige **tærskelværdier** i

flere forskellige beslutningspunkter, eller en logistisk regression kan repræsentere sammenhængen mellem **målegenskaben** og en bagvedliggende egenskab ved flere forskellige variable (såkaldte **polynomiske variable**), der har forskellig vægt (se kapitel 3). I disse tilfælde kræver modulær transparens af en relevant egenskab, at man kigger samlet på de variable, hvor egenskaben optræder.

I nogle tilfælde vil modeller kunne finde statistisk pålidelige sammenhænge, som for mennesker kan virke ulogiske. Ulogiske sammenhænge er ikke nødvendigvis uigennemsigtige i den forstand, som vi behandler i dette kapitel, det er bare svært for mennesker at forstå, hvorfor der er en sammenhæng (se kapitel 6).

SYSTEMISK TRANSPARENS

Mens algoritmisk transparens giver adgang til detaljerede informationer om profileringsmodellens design og funktion, giver systemisk transparens adgang til information om dens træning, kvalitet, risici, anvendelse og tilsyn. Systemisk transparens dækker med andre ord de informationer, der "omkredser" og supplerer algoritmisk transparens. I praksis er en teknisk forklaring af modellens funktion nemlig ikke altid tilstrækkelig for, at borgere og offentlighed kan forstå modellen og dens brug.

Stort set alle bidrag i litteraturen om transparens introducerer begreber for de typer af informationer, som kan vise sig nødvendige udover den algoritmiske transparens.[9] Den manglende ensartede terminologi eller tværfaglige konsensus på tværs af jura og teknik om, hvilke begreber der skal anvendes, og hvad de skal dække over, er formentlig en konsekvens af, at der er tale om et nyt og ureguleret felt. I hvert fald findes der endnu ikke en ensartet, autoritativ terminologi, der omfatter såvel juridiske og tekniske aspekter af, hvad vi her beskriver som systemisk transparens.

Systemisk transparens omfatter fire typer af informationer (se boks).

For det første informationer om modellens træning, dvs. om det **datasæt**, der er brugt til at udvikle modellen, herunder hvor oplysningerne i datasættet stammer fra. Information om træning omfatter også oplysninger om datakvalitet og spørgsmålet om dataminimering (se kapitel 5) samt aggregerede statistikker om hvordan beskyttede kendetegn som f.eks. køn, etnicitet eller handicap anvendes i modellen (se kapitel 8).

For det andet information om modellens kvalitet og risici, dvs. dens **fejlrate** og ydeevne når den anvendes på nye sager. Til denne type hører også, hvorvidt modellen skaber risiko for usaglige og forkerte (ulovlige) afgørelser (se kapitel 6), samt om den i øvrigt har utilsigtede konsekvenser for rettigheder, herunder for beskyttelsen af personoplysninger (se kapitel 5) eller forbuddet mod diskrimination (se kapitel 8).

For det tredje modellens konkrete anvendelse i myndighedens sagsbehandling: Hvordan bruges modellen i sagsbehandlingen og er den egnet til det formål, den er tiltænkt? I forlængelse heraf indeholder denne type informationer også overordnede og letforståelige beskrivelser af modellens funktion, type og design, der ikke har samme detaljegrad, som informationer, der giver algoritmisk transparens.

For det fjerde er der informationer om kontrol og tilsyn med modellen. Dette omfatter spørgsmålet om, hvorvidt der er behov for menneskelig kontrol med modellen og hvilke oplysninger, der er tilgængelige for hvem om f.eks. modellens kildekode, træningsdata, variable, vægte, modeltype og måling af modelkvalitet. Denne type omfatter også informationer om, hvordan modellen forklares for borgeren.

SYSTEMISK TRANSPARENS: FIRE TYPER AF INFORMATIONER

Modellens træning: Informationer om det datasæt, der er brugt til udvikling af modellen

Modellens kvalitet og risici: Informationer om modellens fejlrate og ydeevne, når den anvendes på nye sager og risici i forhold til rettigheder og retssikkerhed

Modellens anvendelse: Information om, hvad modellen skal bruges til og om den er egnet til formålet samt overordnede og letforståelige beskrivelser af modellens funktion, type og design.

Kontrol og tilsyn med modellen: rammer for tilsyn og oplysninger om, hvordan modellen forklares for borgeren

Systemisk transparens omfatter mange af de informationer, vi behandlede i kapitel 4 som en del af anbefalingen om AI-konsekvensanalyser. Dette betyder også, at opfyldelsen af kravene om konsekvensanalyser – herunder offentliggørelse af analyserne eller et uddrag heraf – netop er en måde, hvorpå systemisk transparens delvist kan blive sikret.

En række af de oplysninger, som vi her har opregnet under begrebet systemisk transparens, kan desuden kræves efter gældende regler i databeskyttelses-forordningen, forvaltningsloven og offentlighedsloven. Det er imidlertid uafklaret, hvad reglernes rækkevidde er og præcis hvilke informationer om modellen, reglerne dækker over. Dette er et af vores fire fokusområder

FIRE FOKUSOMRÅDER FOR TRANSPARENS

Som nævnt i starten af dette kapitel er det kendetegnende for gældende regler, at de – med undtagelse af en enkelt regel i databeskyttelsesforordningen – ikke specifikt regulerer brugen af profileringsmodeller.

De forskellige regler omhandler desuden ikke samme typer af transparens. Der hersker derfor usikkerhed om gældende reglers anvendelse på profilerings-modeller, herunder i hvilket omfang reglerne dækker over (dele af) algoritmisk og systemisk transparens.

Som myndighed, borger eller kontrolorgan kan man ikke på forhånd vide, hvilken type transparens, man skal bruge i den konkrete sag. Det er selvsagt ikke muligt på forhånd at vide, om man skal kontrollere modellen for diskriminationsrisici (der kræver algoritmisk transparens) eller for behandlingen af korrekte og nødvendige personoplysninger (der kræver systemisk transparens).

Der er med andre ord behov for samlede regler, der sikrer begge former for transparens. Det indebærer både, at der skal gennemføres nye regler, samt at gældende regler skal præciseres. Vi peger i det følgende på fire væsentlige pointer:

For det første, at brugen af profileringsmodeller skærper behovet for myndighedernes oplysningspligt: hvilke oplysninger om borgeren sidder myndigheden inde med, og hvordan anvender de dem til at profilere borgeren?

For det andet, at retssikkerheden øges, hvis der indføres særregler om systemisk transparens: hvordan anvender myndighederne profileringsmodeller, hvilke rettigheds- og retssikkerhedsmæssige risici rejser de og hvem fører tilsyn?

For det tredje, at et offentligt register over myndighedernes brug af profilerings-modeller ligeledes vil øge retssikkerheden: hvilke myndigheder i Danmark anvender profileringsmodeller og hvordan?

Og endelig, for det fjerde, at myndighedernes pligt til at begrunde en afgørelse stiller krav til algoritmisk transparens: hvorfor vurderes borgeren på en bestemt måde, og hvad ligger til grund for modellens vurderinger?

MYNDIGHEDERS OPLYSNINGSPLIGT

Det første af de fire fokusområder vedrører strengt taget hverken algoritmisk eller systemisk transparens, men det faktum, at borgerens eksisterende rettigheder efter databeskyttelsesforordningen kan risikere at blive forringet ved det blotte forhold, at en profileringsmodel anvendes – uanset, hvor transparent modellen i øvrigt måtte være.

Databeskyttelsesforordningens forpligter myndigheder til at give borgeren oplysning om en række forhold, når de behandler oplysninger om vedkommende.[10] Der findes imidlertid visse begrænsninger i oplysningspligten, som – hvis de fastholdes på trods af brugen profileringsmodeller – kan komme til at forringe borgernes beskyttelse.

Vores fokus i dette afsnit er på, at borgere ikke får at vide, at oplysninger om dem, som er blevet indsamlet tidligere nu indgår i en profileringsmodel, der tjener

et helt andet (og uforeneligt) formål, end hvad oplysningerne oprindeligt blev indsamlet til.[11]

Databeskyttelsesforordningens princip om formålsbestemthed (også kaldes finalité-princippet) opstiller et udgangspunkt om, at oplysninger skal indsamles til udtrykkeligt angivne og legitime formål og ikke må viderebehandles på en måde, der er uforenelig med disse formål.[12] Formålet med princippet er bl.a. at sikre gennemsigtighed og forudsigelighed ved at sætte grænser for brugen af personoplysninger og sikre en rimelig behandling af oplysningerne.[13]

Udgangspunktet er, at personoplysninger gerne må viderebehandles til andre formål end de, hvortil oplysningerne oprindeligt blev indsamlet, blot de nye formål ikke er uforenelige med de oprindelige. Det følger af databeskyttelsesforordningen, at myndigheden skal give borgeren besked om formålet med viderebehandlingen.[14]

Det er imidlertid muligt at fravige princippet om formålsbestemthed (således, at indsamlede oplysninger også kan behandles til nye formål, der er uforenelige med de oprindelige). Denne mulighed har man valgt at gøre brug af i Danmark i databeskyttelsesloven.[15]

Begrænsningen af princippet er bl.a. tiltænkt algoritmisk profiling. Under Folketingets behandling af databeskyttelsesloven i 2017 anførte justitsministeren således, at der med fravigelsen af princippet ville være muligt at udbrede en landsdækkende løsning inspireret af den såkaldte Gladsaxemodel (se bilag 2).[16] Selvom den model, som Gladsaxe kommune udviklede endte med at blev skrinlagt, skaber gældende danske regler således fordelagtige vilkår for, at myndighederne uden krav om formålsbestemthed kan behandle store mængder oplysninger om borgeren til algoritmisk profilering.

Når der undtages fra formålsbestemthed undtages der i den danske data-beskyttelseslov også samtidig for oplysningspligt.[17] Det betyder, at myndigheden ikke giver borgeren besked om den nye behandling af vedkommendes oplysninger til et formål, der kan være uforeneligt med det oprindelige. Oplysningspligten gælder dog fortsat, hvis formålet med viderebehandlingen er kontrol. I disse tilfælde har Folketinget nemlig fundet, at borgeren på forhånd bør være oplyst om myndighedernes fornyede (potentielt uforenelige) brug af oplysningerne.[18] I alle andre tilfælde end kontrol gælder det imidlertid, at borgeren ikke får besked om den nye behandling.

Der er således i databeskyttelsesloven en automatisk begrænsning af oplysningspligten. Når myndigheder bruger profileringsmodeller vurderer vi, at dette er et problem. Begrænsningen af oplysningspligten står i disse tilfælde ikke alene, men skal ses i sammenhæng med de øvrige transparensproblemer, som profileringsmodeller rejser. Ikke nok med, at borgeren ikke ved, hvilke oplysninger, der behandles om vedkommende (herunder oplysninger, som er indsamlet til formål, der er uforenelige med profileringen); udfordringerne med at forstå

modellen (i form af algoritmisk transparens) og få tilstrækkelige oplysninger om modellen og dens brug (i form af systemisk transparens) kan føre til, at borgeren heller ikke ved, hvordan disse oplysninger behandles i modellen, og har betydning for udfaldet af vedkommendes sag. Det samlede billede bliver et, hvor borgeres indsigt i brugen af vedkommendes personoplysninger (og kontrol over disse) svækkes.

Dette skaber efter vores vurdering ikke en tilfredsstillende beskyttelse af borgeren. Hvis det fastholdes, at man i Danmark vil begrænse databeskyttelsesforordningens udgangspunkt om oplysningspligt på trods af brugen af profileringsmodeller – hvor massive mængder data er til rådighed, langt ud over, hvad det kan forventes, at sagsbehandleren ville behandle ved manuel gennemgang af en sag – mindsker det reelt borgerens beskyttelse. Forståelsen af retsgarantier og rettigheder bør stemme overens med de reelle udfordringer. En fastholdelse af den automatiske begrænsning af oplysningspligten, når der bruges profileringsmodeller, indebærer efter vores vurdering en forringelse af borgernes rettigheder.

Vi bemærker, at der efter omstændighederne kan være konkrete grunde til at begrænse oplysningspligten.[19] Nogle af disse behandler vi nedenfor. Pointen her er imidlertid, at en automatisk begrænsning af oplysningspligten, når myndighederne fraviger princippet om formålsbestemthed (især) er problematisk ved brugen af profileringsmodeller.

Vi anbefaler, at Justitsministeriet tager initiativ til at ændre databeskyttelses-loven, således at myndigheden skal give besked om, hvilke oplysninger om borgeren, der bruges i en profileringsmodel til at støtte eller træffe en afgørelse om borgeren.

SÆRREGLER OM SYSTEMISK TRANSPARENS

Det andet fokusområde omhandler borgerens mulighed for at få information om profileringsmodeller og deres brug. Både databeskyttelsesforordningen og offentlighedsloven giver adgang til oplysninger om profileringsmodeller til borgere – og efter offentlighedsloven til den bredere offentlighed.

Der hersker imidlertid stor uklarhed om, hvilke typer informationer om profileringsmodellen, de enkelte regler dækker. Det vil efter vores vurdering øge retssikkerheden, hvis der indføres regler, der udtrykkeligt sikrer en adgang til systemisk transparens. I det følgende gennemgår vi de relevante regler i databeskyttelsesforordningen og offentlighedsloven og de uklarheder, som kan opstå i reglernes anvendelse på profileringsmodeller.

Det er væsentligt at bemærke, at de to regelsæt tjener forskellige formål. Når vi her behandler dem under samme tema, er det for at illustrere, hvordan en særregel om systemisk transparens vil kunne tydeliggøre borgeres og offentlighedens adgang

til informationer om profileringsmodeller og skabe øget åbenhed og offentlighed – og dermed øget retssikkerhed – om myndighedernes brug af modellerne.

DATABESKYTTELSESFORORDNINGENS REGLER

Databeskyttelsesforordningen er som nævnt det eneste regelsæt, der udtrykkeligt regulerer automatiserede beslutninger (se kapitel 2). Forordningen indeholder en særlig form for oplysningspligt (ikke at forveksle med oplysningspligten behandlet ovenfor), når personoplysninger behandles som led i en fuldautomatiseret afgørelse.[20] Den særlige oplysningspligt sikrer, at borgeren som minimum skal have "meningsfulde oplysninger om logikken" i en profileringsmodel, som bruges til fuldautomatiserede afgørelser. Rækkevidden af bestemmelsen er imidlertid uklar.

Den såkaldte Artikel 29-gruppe, hvis retningslinjer udgør et vigtigt fortolkningsbidrag på området (se kapitel 2) anfører, at myndigheden ikke nødvendigvis skal give en kompliceret beskrivelse af modellen eller lægge hele modellen frem. Det anføres også, at kompleksitet ikke er en undskyldning for at undlade at give oplysninger til borgeren.[21]

I den danske betænkning om databeskyttelsesforordningen og de retlige rammer for dansk lovgivning fremhæves det ligeledes, at oplysningspligten ikke kræver en meget detaljeret beskrivelse af behandlingens grundlag. Det afgørende er i stedet, at borgeren kan forstå de overvejelser, som ligger til grund for behandlingen, og hvordan profileringsmodellen mere generelt kommer frem til forskellige afgørelser.[22]

Formålet med reglerne i databeskyttelsesforordningen er at give et overblik over de tilsigtede aktiviteter og dermed give borgeren forståelse for og viden om automatiseringens formål. Det er imidlertid et omdiskuteret spørgsmål, om forordningen giver adgang til oplysninger om, hvorfor en model konkret har truffet en afgørelse, eller om hvordan den har truffet afgørelsen. Betyder "meningsfulde oplysninger om logikken", at der skal gives adgang til en specifik begrundelse for modellens resultat i en konkret sag (hvorfor afgørelsen er truffet) eller at der skal gives adgang til en generel og generisk beskrivelse af modellen (hvordan afgørelsen er truffet)? Det første indebærer algoritmisk transparens på lokalt niveau. Det andet kan formentlig opfyldes ved systemisk transparens.

Til støtte for adgang til en konkret begrundelse for sagens udfald – et "hvorfor" – taler, at databeskyttelsesforordningens indledende betragtninger fremhæver, at den registrerede bør have "ret til en forklaring".[23] Det er også blevet anført, at retten til at bestride den automatiserede afgørelse, som følger af databeskyttelsesforordningen, ville være illusorisk, hvis man ikke kendte den konkrete begrundelse for afgørelsen.[24]

Omvendt fandtes der i et tidligere udkast til forordningen udtrykkeligt en ret til forklaring, som imidlertid ikke kom med i den endelige version. Dette taler imod, at der skulle gælde en ret til at få en konkret begrundelse efter forordningens regler, på trods af forordningens betragtninger, da disse til forskel fra reglerne ikke er juridisk bindende.[25]

Samlet set er det vores vurdering, at databeskyttelsesforordningen næppe giver ret til svar på, hvorfor modellen har truffet den afgørelse, som den har. Forordningen giver snarere overordnet og letforståelig information om, hvordan modellen opererer. Dette er en af informationstyperne inden for den systemiske transparens, som vi har præsenteret ovenfor. Forordningen sikrer med andre ord ikke en ret til begrundelse for en konkret afgørelse. En sådan ret til begrundelse følger imidlertid af forvaltningslovens regler, som vi behandler nedenfor.

Udover den uklare rækkevidde af bestemmelsen er adgangen til "meningsfulde oplysninger om logikken" underlagt en række begrænsninger. For det første gælder reglen kun, når der er tale om fuldautomatiserede afgørelser og ikke ved automatiseret beslutningsstøtte. Det bidrager til at øge uklarheden om reglens anvendelse, at det ikke altid er til at vurdere, om en afgørelse er fuldautomatiseret eller om modellen tjener til beslutningsstøtte (se kapitel 4). For det andet kan en myndighed undtages fra reglerne, hvis f.eks. hensynet til forretningshemmeligheder eller myndighedens kontrolbeføjelser konkret gør det nødvendigt (se nærmere nedenfor).

Det vil efter vores vurdering øge retssikkerheden, hvis forordningens oplysningspligt blev suppleret i dansk ret af en særregel om systemisk transparens, så der er klarhed om retstilstanden, og så borgere får sikret adgang til oplysninger om modellen, også når der er tale om beslutningsstøtte. Derfor anbefaler vi nedenfor, at der tilføjes en særregel i dansk ret, der sikrer systemisk transparens, herunder generelle og letforståelige informationer om profileringsmodellens funktion og design.

AKTINDSIGT EFTER OFFENTLIGHEDSLOVEN

Reglerne om aktindsigt kan ligeledes give adgang til forskellige typer af informationer om modellen. Reglerne om aktindsigt kan opdeles i på den ene side regler, der sikrer borgeren i en sag adgang til sagens oplysninger (partsaktindsigt i medfør af forvaltningsloven), og på den anden side regler, der sikrer offentligheden adgang til myndighedens oplysninger (almindelig dokumentoffentlighed i medfør af offentlighedsloven). De regler og principper, vi omtaler i det følgende, gælder både for offentligheden og for borgeren, der er part i en sag.

Det er ikke afklaret i litteratur eller retspraksis, hvordan offentlighedslovens regler om aktindsigt finder anvendelse på profileringsmodeller. Allerede af lovens første bestemmelse følger imidlertid, at myndighederne skal sørge for, at lovens hensyn til åbenhed i videst muligt omfang varetages ved valg, etablering og udvikling af

nye IT-løsninger.[26] Loven opstiller et udgangspunkt om, at enhver kan forlange at blive gjort bekendt med dokumenter, der er indgået eller oprettet af en myndighed som led i administrativ sagsbehandling.[27] Dette udgangspunkt er suppleret med en række undtagelser, og her er det usikkert, hvornår forskellige informationer om profileringsmodellen er omfattet af udgangspunkt eller undtagelser.

I det følgende fremhæver vi offentlighedslovens regler om tre typer af dokumenter, som kan illustrere uklarheden i spørgsmålet om aktindsigt om profileringsmodeller.

For det første gælder, at såkaldte interne arbejdsdokumenter er undtaget fra aktindsigten.[28] Det kan f.eks. være myndighedens referater, notater og udkast til breve, afgørelser mv. som ikke er givet til udenforstående.[29] Nogle typer af interne dokumenter er "selvstændige dokumenter"[30] og er fortsat omfattet af aktindsigt. Det kan f.eks. være afhøringsrapporter, inspektionsrapporter, interviewskemaer mv.[31] Indeholder dokumenterne imidlertid en internt præget vurderende stillingtagen, er der ingen aktindsigt.[32] Det er ikke afklaret, om oplysninger om en profileringsmodel, der indebærer en vurderende stillingtagen til borgeren vil være omfattet af aktindsigt eller ej efter disse regler.

For det andet gælder, at såkaldte skuffecirkulærer er omfattet af aktindsigt.[33] Dette omfatter f.eks. generelle retningslinjer for behandlingen af en given sagstype. Dog er der ikke aktindsigt i tekniske indretninger, der er af væsentlig økonomisk betydning for myndighedens it-leverandør.[34] Det er også muligt at begrænse aktindsigt af hensyn til bl.a. kontrol- og tilsynsopgaver[35] (se nedenfor). Reglen om skuffecirkulærer giver formentlig adgang til visse informationer om modellen, men det er uafklaret, hvilke informationer og med hvilken detaljegrad.[36]

For det tredje gælder, at såkaldte databeskrivelser er omfattet af aktindsigten.[37] Databeskrivelser er beskrivelser af, hvilke typer af oplysninger, der indgår i en database, hvilket grundlag oplysningerne bygger på, samt hvilke formater databasen anvender. Adgang til databeskrivelser er begrænset bl.a. af hensyn til myndighedens kontrol- og tilsynsopgaver[38], ligesom databeskrivelserne ikke omfatter mere teknisk prægede oplysninger, herunder adgang til kildekoden.[39] Det er ligeledes uafklaret hvordan bestemmelsen finder anvendelse på profileringsmodeller i sagsbehandlingen.

Samlet set regulerer offentlighedslovens således indsigt i mange typer af informationer, men indeholder ingen regler rettet specifikt mod profilerings-modeller og systemisk transparens.

BEGRÆNSNING AF TRANSPARENS

Både i databeskyttelsesforordningen og i reglerne om aktindsigt er der begrænsninger i adgangen til oplysninger ud fra hensynet til, at myndigheder kan foretage effektiv kontrol og tilsyn, samt hensynet til beskyttelsen af forretningshemmeligheder for den virksomhed, der udvikler og sælger modellen.

Hvad angår myndighedernes kontrol- og tilsynsopgaver, er det klare udgangspunkt, at offentlige myndigheder i deres sagsgang skal være styret af åbenhed og offentlighed. Myndighederne har imidlertid mulighed for at hemmeligholde det forhold, at de behandler en sag om kontrol eller tilsyn.

Hensynet til hemmeligholdelse i kontrol- og tilsynsager har til formål at sikre, at oplysningen af sagen ikke hindres af borgeren som følge af kendskabet til det igangværende tilsyn eller kontrol. Dette hensyn indgår efter en konkret vurdering i sagen og gælder uanset, om sagen automatiseres (helt eller delvist) eller ej. Adgangen til informationer om modellen bør derfor ikke begrænses i yderligere omfang, end hvad der ville følge, når sagerne behandles manuelt og bør være en konkret vurdering. Kontrol- og tilsynsopgaver bør efter vores vurdering desuden ikke føre til begrænsninger i borgernes og offentlighedens adgang til systemisk transparens i form af generelle oplysninger om modellen.

Hvad angår beskyttelsen af forretningshemmeligheder, vil en privat virksomhed, som udvikler en profileringsmodel for en myndighed formentlig betragte kildekoden og data om modellens træning og funktion som forretnings-hemmeligheder. Det skyldes, at konkurrenter med disse data muligvis vil kunne kopiere modellen og tilbyde beslægtede produkter, uden at pådrage sig de samme omkostninger til udvikling.[40] De regler, der beskytter forretnings-hemmelighederne stiller imidlertid krav om, at myndigheden foretager en konkret vurdering af behovet for indsigt i modellen over for beskyttelsen af forretningshemmeligheder. Afvejningen mellem disse to hensyn indebærer, at myndigheden har – eller kan få – adgang til oplysninger beskyttet af forretningshemmeligheder, såfremt afvejningen måtte fald ud til fordel for borgerens eller offentlighedens ret til indsigt. Heller ikke beskyttelsen af forretningshemmeligheder indebærer efter vores vurdering, at mere generelle oplysninger om brugen af profileringsmodellen bør begrænses.

OPSAMLING

Som det ses af det forudgående, er anvendelsesområdet for regler i data-beskyttelsesforordningen og offentlighedsloven uklart, og reglerne dækker over både overlappende og særskilte former for informationer. Det er vores vurdering, at der er behov for en særregel om systemisk transparens, når det kommer til offentlige myndigheders brug af profileringsmodeller.

Efter vores vurdering vil det øge retssikkerheden, hvis der bliver fastsat en regel, som udtrykkeligt regulerer borgeres og den bredere offentligheds adgang til informationer om profileringsmodeller, således, at det tydeliggøres, at såvel borgeren, der er part i en sag som offentligheden sikres en vis systemisk transparens.

Konkret bør der efter vores vurdering tilføjes en særregel om systemisk transparens i offentlighedslovens regler om aktindsigt. Reglen kan om nødvendigt underlægges begrænsninger i lighed med øvrige aktindsigtsregler. Reglen bør

dog opstille et udgangspunkt om systemisk transparens og give indsigt i de informationstyper, som er omfattet af dette. Dvs. at reglen bør give information om modellens træning, kvalitet, risici, anvendelse og tilsyn. Informationer om modellens anvendelse dækker over overordnede beskrivelser af modellens funktion, type og design, der ikke behøves opfylde det detaljeniveau, som kræves for algoritmisk transparens men bør kunne fortælle noget om, hvordan modellen virker.

Det bør endvidere tydeliggøres, at informationer skal gives både, når modellerne bruges til fuldautomatiseret behandling og når de bruges til beslutningsstøtte. Set fra borgerens perspektiv vil det have lige så stor betydning at få systemisk transparens om modeller brugt til beslutningsstøtte, ikke mindst fordi det må forventes, at myndighederne i højere grad vil bruge automatiseret beslutningsstøtte, når der er tale om særligt bebyrdende eller indgribende afgørelser eller afgørelser, der kræver skønsmæssige eller fortolkningsmæssige vurderinger (se kapitel 6).

Vi anbefaler, at Justitsministeriet tager initiativ til at indføre regler i offentlighedsloven om aktindsigt i informationer om modellens træning, kvalitet, anvendelse og tilsyn, når myndigheder anvender profileringsmodeller til automatiseret beslutningsstøtte eller ved fuldautomatisering.

OFFENTLIGT REGISTER OVER BRUG AF PROFILERINGSMODELLER

Det tredje fokusområde drejer sig om et offentligt tilgængeligt register over myndighedernes brug af profileringsmodeller. Når offentlige myndigheder i dag bruger profileringsmodeller i sagsbehandlingen, har hverken borgeren eller den bredere offentlighed indblik i omfang og anvendelse. Nogle myndigheder vælger at offentliggøre visse oplysninger, mens andre ikke stiller oplysninger til rådighed om deres brug.

Dette forhold udgør efter vores vurdering en udfordring for retssikkerheden, eftersom åbenhed er en forudsætning for, at myndighederne kan stilles til ansvar i eventuelle sager, ligesom den bredere offentlighed og tilsynsmyndigheder har krav på indsigt i myndighedernes ageren. For at sikre denne åbenhed er der behov for et offentligt tilgængeligt register over samtlige myndighedernes brug af profileringsmodeller.

Flere steder i Europa har man indført offentligt tilgængelige registre om offentlige myndigheders brug af kunstig intelligens, og også i EU-Kommissionens udkast til en forordning om kunstig intelligens foreslås det, at man etablerer en offentlig database i EU-regi for visse højrisiko AI-systemer (se boks).

EUROPÆISKE TILTAG OM AI-REGISTRE OG DATABASER

Amsterdam i Holland og Helsinki i Finland oprettede i efteråret 2020 offentligt tilgængelige registre over myndigheders brug af kunstig intelligens – såkaldte algoritmeregistre. Registrene giver overblik over, hvilke AI-systemer der anvendes af de offentlige myndigheder i byerne. For hvert system er der fremlagt oplysninger om: 1. datasæt, 2. databehandling, 3. sikring mod diskrimination, 4. menneskelig kontrol og 5. generelle risici. Registrene giver også et grundlæggende indblik i, hvad algoritmer er, og hvordan de påvirker hverdagen for byens borgere.[41]

I EU-Kommissionens udkast til en forordning for kunstig intelligens er der foreslået etablering af en offentlig EU-database for visse højrisiko AI-systemer, herunder offentlige myndigheders brug af profileringsmodeller.[42] Myndighederne vil være forpligtede til at registrere brugen af profileringsmodellerne[43], og i databasen vil det være muligt at se oplysninger om bl.a. formålet med modellen samt en såkaldt "EU-overensstemmelseserklæring"[44] om, at modellen efterlever forordningens krav til bl.a. datakvalitet, teknisk dokumentation, transparens og information, menneskeligt tilsyn og krav til systemets robusthed, nøjagtighed og cybersikkerhed.

Der er god inspiration at hente i disse eksempler, men offentlige myndigheders brug af profileringsmodeller rejser nogle særlige rettighedsmæssige risici, som efter vores opfattelse kalder på mere fyldestgørende oplysninger om modellernes brug.

Et samlet offentligt tilgængeligt register over brugen af profileringsmodeller bør som minimum omfatte de mest centrale oplysninger om modellens træning, kvalitet, risici, anvendelse og tilsyn med udgangspunkt i den nye regel om aktindsigt, som vi anbefaler foroven. En måde at give disse oplysninger på er ved at offentliggøre AI-konsekvensanalyserne eller et uddrag heraf (se kapitel 4).

Det kan ikke udelukkes, at der i visse tilfælde vil være behov for at tilbageholde oplysninger om, at myndigheden undersøger borgerens forhold, herunder også når der bruges profileringsmodeller. Tilsyns- og kontrolhensynet berettiger imidlertid ikke, at myndigheder undlader overhovedet at oplyse om, at profileringsmodeller finder anvendelse eller at de undlader en generel beskrivelse af, hvorledes de fungerer og anvendes. Hensynet til tilbageholdelse har derfor ikke betydning for etableringen af et offentligt register.

Et sådant register kan efter instituttets vurdering oprettes af Digitaliserings-styrelsen, der netop har til formål at skabe en sammenhængende offentlig service, der giver borgeren et overblik over det offentlige Danmarks digitalisering.[45] Et samlet overblik over offentlige myndigheders brug af profileringsmodeller vil efter vores vurdering være med til at styrke retssikkerheden på området.

Vi anbefaler, at Digitaliseringsstyrelsen, som led i det fællesoffentlige arbejde med kommuner og regioner, opretter et offentligt register over samtlige offentlige myndigheders brug af profileringsmodeller rettet mod borgere.

at AI-konsekvensanalyserne eller et uddrag heraf offentliggøres i registret.

ALGORITMISK TRANSPARENS SOM LED I BEGRUNDELSESPLIGTEN

Det fjerde og sidste fokusområde omhandler forvaltningslovens krav om, at en afgørelse skal indeholde en begrundelse, medmindre afgørelsen fuldt ud giver den pågældende part medhold.[46] Begrundelsespligten vil kunne indeholde oplysninger omfattet af systemisk transparens som f.eks. hvordan modellen bruges af myndigheden. Vores fokus i dette afsnit er på, om begrundelsespligten også indeholder krav om algoritmisk transparens.

Det følger af begrundelsespligten, at en begrundelse skal indeholde en henvisning til de regler, som afgørelsen er truffet ud fra. I det omfang afgørelsen beror på et skøn, skal begrundelsen også angive de hovedhensyn, der har været bestemmende for skønsudøvelsen. Begrundelsen skal angive de kriterier, der er blevet inddraget og som har haft afgørende betydning for afgørelsens resultat.[47]

Begrundelsespligten styrker retssikkerheden ved at give borgere adgang til at kontrollere myndighedens afgørelser, ligesom kravet har betydning for sagens behandling ved en klageinstans eller ved domstolene. Kravet kan også være med til at sikre grundigere sagsbehandling, ensartet praksis og samlet effektivitet hos myndigheden.[48]

Når en afgørelse fuldautomatiseres, vil modellens resultat og den endelige afgørelse være sammenfaldende, og begrundelsespligten vil derfor svare til oplysninger om, hvordan modellen genererer sit resultat.

Når profileringsmodellen anvendes til beslutningsstøtte kan den udgøre et "hovedhensyn" for den endelige afgørelse og dermed indgå i begrundelsen. Dette er navnlig tilfældet, når modellens resultat er en væsentlig rettesnor for eller element i den endelig afgørelse. Desuden følger det af god forvaltningsskik, at der ved større eller væsentlige sagsbehandlingsskridt gives en begrundelse.[49]

En særegen problematik ved begrundelsespligten i relation til profileringsmodeller er de tidligere omtalte ulogiske sammenhænge, en model kan producere (se kapitel 6). Modellerne kan finde sammenhænge, som er komplekse og umiddelbart virker ulogiske for et menneske. Denne egenskab anses for en af modellens styrker, men det udfordrer naturligt nok begrundelsespligten og kan føre til, at afgørelsen mister sin "kontrollerbarhed". Dette er en af grundene til vores anbefaling om grænser for fuldautomatisering (se kapitel 6).

Efter vores vurdering bør myndigheden som led i begrundelsespligten være forpligtet til at informere om de konkrete omstændigheder, som er særegne for borgerens sag. Udfordringen er, at dette kan indbefatte alle niveauer af algoritmisk transparens, fra det lokale til den holistiske.

Begrundelsen skal grundlæggende give svar på, hvorfor borgerens sag fik et bestemt udfald. Dette kræver ofte adgang til modellens oplysninger på lokalt niveau (dvs. transparens i forhold til modellens vurdering af den konkrete borger).

Men også de øvrige niveauer af algoritmisk transparens kan vise sig nødvendige. Navnlig modulær transparens i modellen (dvs. transparens om bestemte dele af modellen) kan være nødvendig, idet den viser, om modellen har vurderet en person usagligt, eller om der er blevet diskrimineret direkte mod personen ved at give svar på, hvilken rolle en bestemt variabel har spillet i modellens vurdering.

Myndigheden skal sikre, at formålet med begrundelsespligten, som både er rettet mod den enkelte borger og en efterfølgende klageinstans, fortsat opfyldes når myndigheden bruger en model.

Modellens begrænsede evne til at varetage saglighedskravet (se kapitel 6) og diskriminationsforbuddet (kapitel 8) skærper efter vores vurdering kravet til begrundelsespligten, hvis den fortsat skal tjene sit retssikkerhedsmæssige formål.

Vi anbefaler derfor, at der i en vejledning fra Justitsministeriet (se kapitel 4) tages udtrykkeligt stilling til behovet for algoritmisk transparens som led i begrundelsespligten både når modellen anvende til fuldautomatisering og når den anvendes til beslutningsstøtte.

Transparens i modellen vil efter vores vurdering være et led i begrundelsespligten efter gældende ret, når modellen anvendes til at oplyse sagen og indgår som et "hovedhensyn" i afgørelsen. Det vil også kunne følge af principperne om god forvaltningsskik, i tilfælde, hvor modellen er blevet anvendt til at prioritere imellem sager. Det er imidlertid mere tvivlsomt, om en myndighed vil anse det for nødvendigt at give oplysninger om modellen og dens resultat i tilfælde, hvor en model er blevet brugt til at "screene" borgere og vurdere, hvem der skal oprettes sager om. I disse tilfælde er modellen blevet brugt, før der overhovedet er oprettet en sag om borgeren. Efter vores vurdering vil det øge retssikkerheden, hvis algoritmisk transparens i modellen indgår i begrundelsespligten også i disse tilfælde. Spørgsmålet om, hvorfor en sag er blevet oprettet om én borger men ikke en anden, kan have stor relevans for blandt andet spørgsmålet om diskrimination. Det vil efter vores vurdering være problematisk, hvis myndighederne ikke i disse tilfælde er forpligtede til at forklare, både hvordan modellen fungerer og hvordan den har haft indflydelse på udvælgelsen af borgere, som der skal oprettes sager om.

Vi bemærker, at der kan være en glidende overgang fra systemisk transparens, som indeholder letforståelige og overordnede beskrivelser af modellens funktion

og design og til algoritmisk transparens. Algoritmisk transparens i form af forklaringer (xAI) om komplekse modeller kan således også tage form af simple og lettilgængelige – men måske teknisk upræcise eller ufuldstændige – forklaringer (se nedenfor). Der kan desuden være forskel på, hvad der er en retvisende og meningsfuld begrundelse (og dermed udgør transparens) for en borger og hvad der er det for et specialiseret tilsyn, der har til formål at føre kontrol med brugen af modellen. Desuden kan der være forskel på, hvad der kan anses for en meningsfuld begrundelse for ikke-teknikkyndige personer og for personer, der besidder domænekendskab.

Begrundelsespligten tjener som nævnt både til, at borgeren forstår sagen udfald og til, at en efterfølgende klageinstans kan prøve afgørelsen[50] og vejledningen bør efter vores vurdering komme ind på, hvilken form for algoritmisk transparens, der efter omstændighederne er nødvendig for at tjene begrundelsespligtens formål.

Algoritmisk transparens vil desuden være en forudsætning for, at der kan etableres et effektivt forudgående og løbende tilsyn med myndighedernes brug af profileringsmodeller (se kapitel 4).

Vi anbefaler, at det i vejledningen (se kapitel 4) tydeliggøres, at profilerings-modellens konkrete vurdering af borgeren altid indgår i begrundelsespligten, herunder også ved myndigheders brug af profileringsmodeller til beslutningsstøtte

ALGORITMISK DILEMMA: KOMPLEKSITET OG TRANSPARENS

Som afslutning på dette kapitel tager vi fat på det, vi har kaldt et algoritmisk dilemma: Balancen mellem modellens kompleksitet og dens algoritmiske transparens – eller sagt på en anden måde, balancen mellem på den ene side modellens adgang til at udføre sin opgave, og på den anden side menneskers mulighed for at få indsigt i dens funktion og resultater. Helt forenklet sagt har mere komplekse modeller typisk en lavere fejlrate, men med højere kompleksitet følger ofte mindre algoritmisk transparens.

I sin kerne er dilemmaet derfor balancen mellem en adgang til at forstå og efterprøve, om modellens afgørelse er korrekt på den ene side og sikring af, at modellen i flest mulige tilfælde når frem til et korrekt resultat på den anden side. Hvordan opnår man som offentlig myndighed denne balance, når man udvikler en profileringsmodel?

I starten af dette kapitel redegjorde vi for, at algoritmisk transparens opnås enten ved, at modellen kan fortolkes (er "interpretable") eller at den kan forklares (er "explainable"). Modeller med lavere kompleksitet er mere transparente og dermed umiddelbart fortolkbare, men har ofte en højere fejlrate, mens modeller med høj kompleksitet alene er forklarbare, men ofte har en lavere fejlrate.

En måde at håndtere udfordringen er at udvikle modeller, som er så enkle, at de kan fortolkes. En almindelig og enkel metode for at sikre, at modellen kan fortolkes, er at bruge en **regulariseringsfaktor** i træningen af modellen (se kapitel 3). Reguleringsfaktoren anvendes til at begrænse antallet af variable som gives vægt i modellen, og dermed reducerer de modellens kompleksitet, og gør det lettere at fortolke den.[51]

I nogle situationer fører træning med en regulariseringsfaktor, som er tilstrækkeligt høj til at sikre en enkel, og let fortolkelig model, til en højere fejlrate, men i mange tilfælde kan enkle modeller opnå samme fejlrate, som mere komplekse modeller. Dette løser så at sige dilemmaet og hvis dette er muligt, er det klart at fortrække. Er det ikke muligt må dilemmaet løses på anden vis.

I litteraturen om maskinlæring er der forslag om, at modeller, der kun kan forklares, men ikke kan fortolkes, ikke bør anvendes i vigtige og indgribende afgørelser.[52] Også i EU-Kommissionens udkast til en forordning for kunstig intelligens er der opstillet visse krav til højrisiko AI-systemer, herunder visse profileringsmodeller anvendt af offentlige myndigheder. Udkastet lægger overordnet op til, at sagsbehandleren skal kunne "fortolke systemets output og anvende det korrekt".[53]

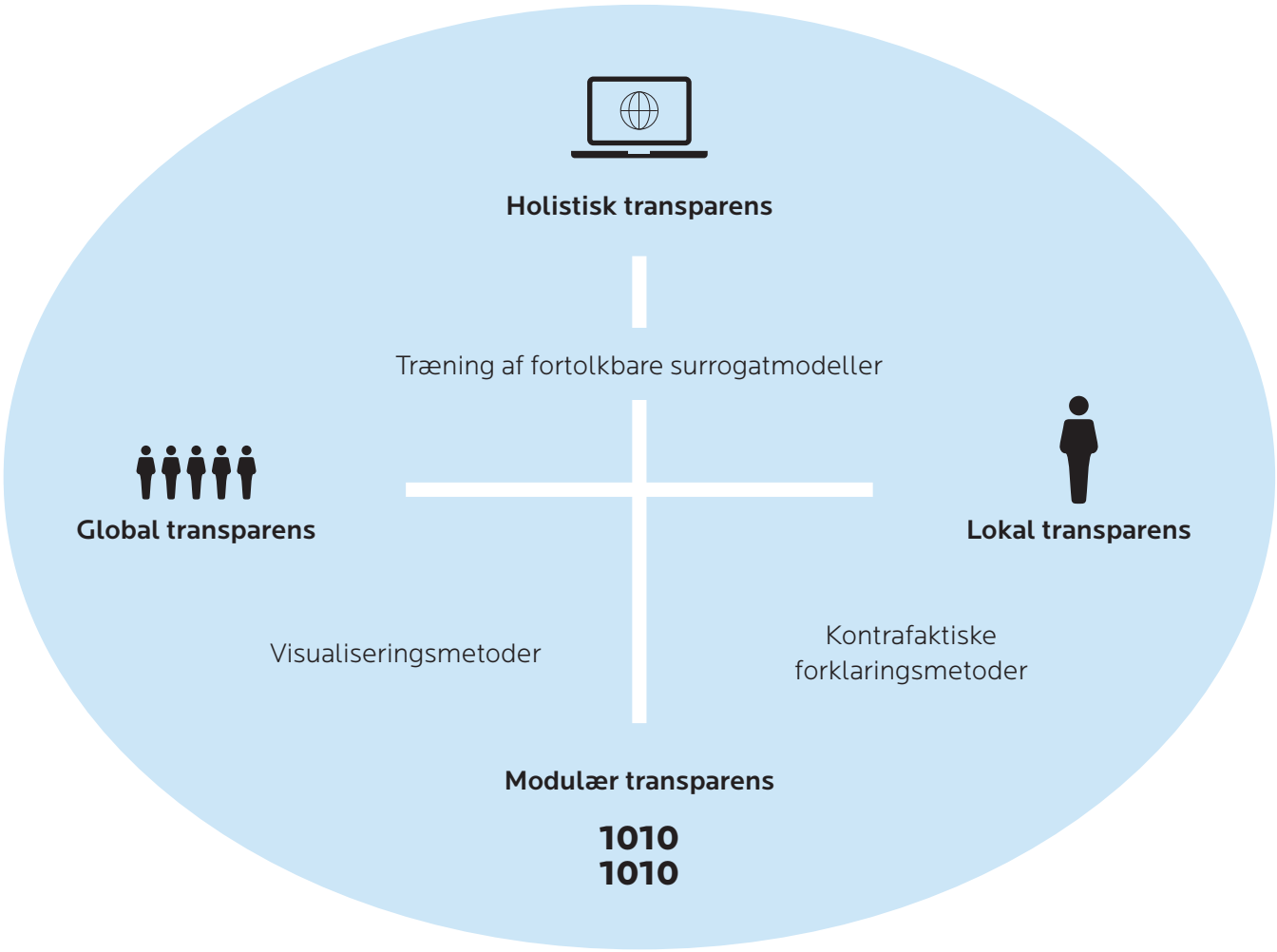
På den ene side bør myndigheder ikke anvende modeller med høje fejlrater, hvis modeller med lave fejlrater findes eller kan udvikles. På den anden side rejser uigennemsigtige modeller helt åbenbare rettighedsmæssige udfordringer også selvom modellen måtte have en lav fejlrate.

Samlet set er det vores vurdering, at myndighederne bør stile efter at anvende den mest muligt algoritmisk transparente model i sagsbehandlingen og under alle omstændigheder redegøre for sit designvalg, herunder brugen af enten en fortolkbar eller forklarbar model.

METODER TIL TRANSPARENS GENNEM "FORKLARBARHED"

Designvalgene om transparens stopper imidlertid ikke her. Som vi har nævnt kan algoritmisk transparens opnås på fire forskellige niveauer og de er ikke alle lige lette at opnå. Når algoritmisk transparens skal opnås gennem forklarbare modeller vil det oftest kræve forskellige tekniske metoder at opnå transparens på de forskellige niveauer (se figur).

TRE METODER TIL FORKLARING AF MODELLERS FUNKTION



Som figuren illustrerer, bruger man forskellige metoder til at sikre de forskellige niveauer af algoritmisk transparens (holistisk, global, modulær eller lokal). Det er i den forbindelse væsentligt at bemærke, at det ikke på forhånd er til at vide, hvilken form for algoritmisk transparens, der er behov for i den enkelte borgers sag.

Dertil kommer, at kvaliteten af forklaringen måles på forskellige parametre, og dette udgør et selvstændigt – retssikkerhedsmæssigt vigtigt – designvalg (se boks).

FORKLARING AF PROFILERINGSMODELLER: FIRE KVALITETSPARAMETRE

En forklarings kvalitet kan måles på fire væsentlige parametre:[54]

- Fidelitet: Kan forklaringen levere en forståelse af modellen, som faktisk svarer til modellen?
- Stabilitet: Gives der forklaringer som minder om hinanden, for personer, som minder om hinanden?
- Repræsentativitet: Hvilke niveauer af transparens – holistisk, modulær, global og lokal – dækker forklaringen over?
- Anvendelighed: Giver forklaringen rent faktisk viden om det, som en borger eller et tilsyn har krav på?

På grund af de designvalg, som både kvalitetsparametre og de tre metoder i figuren indebærer, vurderer vi, at der er behov for, at myndighederne eksplicit i AI-konsekvensanalyserne (se kapitel 4) forholder sig til udfordringerne ved transparens og redegør for designvalg og imødegåelse af retssikkerhedsmæssige risici ved valget. Som nævnt i dette kapitel anbefaler vi yderligere offentliggørelsen af disse konsekvensanalyser eller et uddrag heraf.

De tre metoder i figuren beskrives her kortfattet:

VISUALISERINGSMETODER

For at skabe transparens om en bestemt variabel kan man anvende en række enkle metoder til at analysere og forklare variablens effekt på modellens vurderinger, ved at visualisere den. Dette kan gøres ved at måle den gennemsnitlige effekt, som forskellige værdier af variablen (f.eks. et kendetegn for borgeren) har for modellens vurdering. Analysen kræver både et datasæt af relevante personer (f.eks. modellens træningssæt) og oplysninger om nye personer (de, som der skal træffes afgørelser om). Resultaterne kan plottes på en graf og viser, hvordan ændringer i værdien for den relevante variabel i gennemsnit påvirker modellens vurdering af personer. [55]

Der er tale om let anvendelige forklaringsmetoder, som giver forklaringer, der typisk er nemme at forstå, men de forklarer kun effekten af en bestemt variabel og ikke effekten af de variable samlet set.

KONTRAFAKTISKE FORKLARINGER

For at finde ud af, hvilke af modellens variable, som spiller den største rolle for en bestemt vurdering, kan man benytte såkaldte kontrafaktiske forklaringsmetoder.[56] En kontrafaktisk forklaring forsøger at besvare spørgsmålet "hvorfor vurderede modellen personen, som den gjorde?", ved at pege på, hvordan værdierne for personens variable skulle have været anderledes, for at vurderingen ville have været anderledes.

Kontrafaktiske forklaringer tager udgangspunkt i vurderingen af en specifik person, og beregner den mindste ændring i værdien for en variabel, som ville have ændret modellens vurdering.

Kontrafaktiske forklaringer har den klare fordel, at de er lette at forstå. Til gengæld forklarer de alene effekten af et lille sæt variable for en konkret person. Endvidere kan forklaringens fokus på de variable, som med den mindst mulige ændring ville føre til en anderledes klassificering være misvisende da en klassificering af en person aldrig kun har én årsag.[57] Ofte vil mange af modellens variable bidrage væsentligt til klassificeringen, og der kan derfor findes mange forskellige gyldige, kontrafaktiske forklaringer.

FORTOLKBARE SURROGATMODELLER

Den sidste og måske mest ambitiøse forklaringsmetode er træningen af fortolkbare surrogatmodeller.[58] En surrogatmodel er en umiddelbart transparent model, som er trænet til at vurdere personer på samme måde som den uigennemsigtige model. Konkret erstattes de faktiske værdier for målegenskaben i surrogatmodellens træningssæt med den oprindelige models vurderede værdier for målegenskaben. Derved tilstræber træningen, at surrogatmodellen vurderer personer på samme måde som den oprindelige model, uanset om disse vurderinger er korrekte.

Fordelen ved fortolkbare surrogatmodeller er, at de sikrer global, holistisk transparens. Udfordringen er, at den forklaring, som surrogatmodellen genererer, kan være misvisende. Enten må surrogatmodellen op i en vis kompleksitet for at kunne "efterligne" den oprindelige model og derved risikere selv at blive uigennemsigtig, eller også må den generere forklaringer om den oprindelige model, som ikke nødvendigvis er korrekte og dækkende.[59] Når den oprindelige model er uigennemsigtig, kan man ikke sige hvornår og hvordan forklaringen er misvisende, f.eks. om forklaringen er retvisende for en person men ikke for en anden.

NOTER

1 Akhtar (forthcoming), Transparens i det offentliges brug af maskinlæring, Jarlner og Escherich, Fra velfærdsstat til overvågningsstat Djøf Forlag
2 Se Biran, og Cotton (2017). Explanation and Justification in Machine Learning: A Survey. IJCAI-17 (XAI), Lipton (2017). The Mythos of Model Interpretability. arXiv, Molnar, C. (2019). Interpretable machine learning. A guide for making black box models explainable.
3 Se i almindelighed Biran og Cotton (2017). Explanation and Justification in Machine Learning: A Survey. IJCAI-17 (XAI), Doshi-Velez, og Kim (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv e-prints, arXiv:1702.08608, Guidotti, Monreale, S. Ruggieri, F. Turini, D. Pedreschi, et al. (2019). A Survey Of Methods For Explaining Black Box Models. ACM Computing Surveys 51(5)
4 Burrell (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms.3(1): 2053951715622512.
5 Larsson og Heintz (2020) Transparency in artificial intelligence, Internet Policy Review, 9(2)
6 Molnar (2019) Interpretable machine learning. A guide for making black box models explainable
7 Burrell, (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. 3(1): 2053951715622512, Lipton, (2017). The Mythos of Model Interpretability. arXiv, Rudin, (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1(5): 206-215
8 Molnar (2019). Interpretable machine learning. A guide for making black box models explainable.
9 Se bl.a. Wachter, Mittelstadt og Russell (2020), Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, Harvard Journal of Law and Technology, Selbst og Barocas (2018), The Intuitive Appeal of Explainable Machines 87 Fordham Law Review 1085, Hildebrandt (2018) Algorithmic regulation and the rule of lawPhil. Trans. R. Soc. A.376201703552017035 5 og Motzfeldt og Abkenar (2019) Digital forvaltning, Djøf forlag kapitel 8.
10 Databeskyttelsesforordningen 2016/679 af 27. april 2016 artikel 13 og 14
11 Databeskyttelseslovens § 23
12 Databeskyttelsesforordningen 2016/679 af 27. april 2016 artikel 5, stk. 1, litra b
13 Artikel 29-Gruppens udtalelse (2013) Opinion 03/2013 on Purpose limitation 13/EN WP 203, s. 11

14 Databeskyttelsesforordningen 2016/679 af 27. april 2016 artikel 13, stk. 3 og 14, stk. 4
15 Se Databeskyttelsesforordningen 2016/679 af 27. april 2016 artikel 23 og databeskyttelseslovens § 5, stk. 3. For at gøre brug af undtagelsesadgangen i § 5, stk. 3 skal der indhentes en forudgående udtalelse fra Datatilsynet (databeskyttelseslovens § 28) og Folketingets Retsudvalg skal modtage udkast til bekendtgørelse, som udvalget med et flertal kan beslutte, ikke må udstedes (se besvarelsen af spørgsmål 115 til L 68).
16 jf. navnlig justitsministerens svar på spørgsmål 71 til Retsudvalget: https://www.ft.dk/samling/20171/lovforslag/L68/spm/71/index.htm
17 Se databeskyttelseslovens § 23.
18 Se databeskyttelseslovens § 23 og Retsudvalgets betænkning af 9. maj 2018 til L 68.
19 Se databeskyttelseslovens § 22, stk. 2. Om bestemmelsen anføres i forarbejderne, at bestemmelsen ikke giver adgang til generelt at undtage bestemte former for behandling fra indsigtsretten, jf. Bemærkninger til L 68 fremsat 25. november 2017, s. 191.
20 Databeskyttelsesforordningen 2016/679 af 27. april 2016 artikel 13, 14 og 15 jf. artikel 22
21 Artikel 29-gruppens retningslinjer om automatiske individuelle afgørelser og profilering i henhold til forordning 2016/679 s. 26.
22 Justitsministeriets betænkning 1565/2017 om Databeskyttelsesforordningen (2016/679) – og de retlige rammer for dansk lovgivnings. 293 tilgængelig på https://www.justitsministeriet.dk/sites/default/files/media/Pressemeddelelser/pdf/2017/bet_1_1.pdf
23 Databeskyttelsesforordningen 2016/679 af 27. april 2016 betragtning nr. 71
24 Se bl.a. European Parliament (2019), Understanding algorithmic decision-making: Opportunities and challenges" s.57ff og Kuner, Bygrave, Docksey og Drechsler (2020), The EU General Data Protection Regulation (GDPR): A Commentary, Oxford University Press, samt Goodman, B. and S. Flaxman (2016) European Union regulation on algorithmic decision-making and a "right to explanation" ICML Workshop on human interpretability in ML.
25 Se bl.a. European Parliament Committee on Civil Liberties, Justice and Home Affairs (2013), Report on the Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation) - A7-0402/2013, A7–0402/2013. Se i samme retning Wachter, Mittelstadt og Floridi (2017), Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, International Data Privacy Law, Vol. 7, Issue 2, og Edwards og Veale (2017) Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for, Duke Law & Technology Review, Vol. 16.
26 Offentlighedslovens § 1, stk. 2.
27 Offentlighedslovens § 7
28 Offentlighedslovens § 26
29 Fenger (2018), Forvaltningsret, Djøf forlag s. 413

30 Offentlighedslovens § 26 (forvaltningslovens § 13)
31 Fenger (2018), Forvaltningsret, Djøf forlag s. 425
32 Se ombudsmandens udtalelse i FOB 2010 18-3
33 Offentlighedslovens § 26, nr. 4, se også Fenger (2018), Forvaltningsret, Djøf forlag, s. 425
34 Offentlighedslovens § 30 nr. 2
35 Offentlighedslovens § 33 nr. 2
36 Se også Motzfeldt, H. M. og A. T. Abkenar (2019), Digital forvaltning, Djøf forlag, s. 317
37 Offentlighedslovens § 12
38 Offentlighedslovens § 33
39 Se offentlighedskommissionens betænkning 1510/2009 om offentlighedsloven, Offentlighedsloven, kapitel 11, afsnit 4.4.2, tilgængelig på: https://www.elov.dk/media/betaenkninger/betaenkning_om_offentlighedsloven_del_1.pdf
40 Burrell, (2016), How the machine 'thinks': Understanding opacity in machine learning algorithms 3(1): 2053951715622512.
41 Registrene er tilgængelige her: https://algoritmeregister.amsterdam.nl/en/ai-register/ og her: https://ai.hel.fi/en/ai-register/
42 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final, artikel 60
43 S Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final, e udkastets artikel 51
44 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final, artikel 48
45 For særligt udvalgte projekter om myndighedernes brug af kunstig intelligens (såkaldte signaturprojekter), har Digitaliseringsstyrelsen oprettet et overblik her: https://digst.dk/strategier/kunstig-intelligens/signaturprojekter/
46 Forvaltningslovens § 22
47 Forvaltningslovens § 24
48 Se f.eks. betænkning 657/72 om begrundelse af forvaltningsafgørelser og administrativ rekurs mv. afgivet af Justitsministeriets udvalg vedrørende begrundelse af forvaltningsafgørelser mv., s. 31ff
49 Fenger (2018), Forvaltningsret, Jurist- og Økonomforbundets Forlag s. 633f samt Motzfeldt og Abkenar (2019) Digital forvaltning, Djøf forlag, s. 199. og s. 312
50 Se nærmere Akhtar (forthcoming), Transparens i det offentliges brug af maskinlæring, Jarlner og Escherich, Fra velfærdsstat til overvågningsstat Djøf Forlag
51 Molnar (2019) Interpretable machine learning. A guide for making black box models explainable
52 Rudin, (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1(5): 206-215
53 Kommissionens udkast til forordning om harmoniserede regler for kunstig intelligens COM(2021) 206 final, betragtning 47 og artikel 13 og 14
54 Molnar (2019). Interpretable machine learning. A guide for making black box models explainable.

55 Molnar (2019). Interpretable machine learning. A guide for making black box models explainable

56 Laugel, Lesot, Marsala, Renard og Detyniecki (2017), Inverse Classification for Comparison-based Interpretability in Machine Learning, arXiv e-prints, arXiv:1712.08443, Molnar, (2019), Interpretable machine learning. A guide for making black box models explainable, Wachter, Mittelstadt og Russell (2020) Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, Harvard Journal of Law and Technology

57 Molnar (2019). Interpretable machine learning. A guide for making black box models explainable og Wachter, Mittelstadt og Russell (2020) Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR, Harvard Journal of Law and Technology

58 Molnar (2019) Interpretable machine learning. A guide for making black box models explainable og Rudin, (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead Nature Machine Intelligence 1(5): 206-215

59 Rudin, (2019), Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1(5): 206-215

KAPITEL 8

FORBUDDET MOD DISKRIMINATION

I dette kapitel stiller vi skarpt på de risici, der kan opstå for ulovlig forskelsbehandling eller diskrimination i sagsbehandlingen, når man tager en profileringsmodel i brug. Disse risici kan overordnet opstå på tre måder, som vi gennemgår i det følgende:

For det første kan en diskriminerende praksis blive videreført i modellen ved, at modellen trænes på et datasæt, der indeholder et element af diskrimination. Det kan enten være fordi datasættet ikke er repræsentativt, eller fordi det har varierende kvalitet på tværs af befolkningsgrupper – eller det kan være fordi datasættet afspejler en diskriminerende myndighedspraksis, som bliver indarbejdet i modellens vurderinger.

For det andet kan modellen forstærke risikoen for diskrimination og skabe såkaldte negative feedbacksløjfer, hvor anvendelse af modellen giver anledning til ny diskriminerende praksis i en forvaltning, baseret på modellens vurderinger og resultater.

Begge dele har vi berørt i kapitel 5 og vil gå i dybden med her.

For det tredje kan diskrimination opstå i mere særegne former i modellerne. Brug af profileringsmodeller indebærer en risiko for, at flere hensyn omfattet af diskriminationsforbuddet ikke kan indeholdes i modellen på samme tid. Teknologien er sjældent i stand til at beskytte mod diskrimination inden for hele forbuddets anvendelsesområde, men myndigheden er naturligvis fortsat forpligtet til at overholde lovgivningen om diskrimination. Dette sætter nogle krav og rammer for teknologiens brug og rejser, hvad vi betegner algoritmiske dilemmaer.

De algoritmiske dilemmaer kan f.eks. være, at modellen oftest kun kan måle diskrimination på én ud af flere forskellige måder, at den ikke altid er i stand til at beskytte flere forskellige beskyttede grupper samtidig, samt at den sjældent kan beskytte mod direkte og indirekte diskrimination på samme tid. Hertil kommer, at en model, der varetager hensynet til diskriminationsforbuddet kan risikere at få en for høj **fejlrate** og dermed for alle borgere foretage flere forkerte vurderinger.

Disse forhold rejser alvorlige rettigheds- og retssikkerhedsmæssige udfordringer i brugen af profileringsmodeller. De er baggrunden for anbefalingen om, at modellerne kun undtagelsesvist anvendes til fuldautomatisering (se kapitel 6).

Når modellen bruges til beslutningsstøtte, er det uhyre vigtigt, at risikoen for diskrimination, som opstår i modellen imødegås, bl.a. ved AI-konsekvensanalyser, tilsyn (se kapitel 4) og transparens i og om modellen (se kapitel 7).

I det følgende giver vi indledningsvist en kort beskrivelse af diskriminations-forbuddet, efterfulgt af et overblik over de forskellige måder, hvorpå algoritmisk diskrimination kan opstå. Herefter behandler vi spørgsmålet om bevis-vanskeligheder ved algoritmisk diskrimination, inden vi dykker ned i de tre typer af diskriminationsrisici.

KORT OM DISKRIMINATIONSFORBUDDET

Forbuddet mod ulovlig forskelsbehandling (diskrimination) er reguleret i både international menneskeret, EU-retten og dansk ret (se kapitel 2). I dette kapitel har vi dog overvejende fokus på EU-Domstolens praksis om diskrimination og ligebehandling. Det skyldes, at den EU-retlige beskyttelse kan bruges til at illustrere både væsentlige udfordringer (f.eks. hvordan diskrimination skal bevises) samt mulige løsninger, (f.eks. ved bevisbyrderegler).

EU-retten tjener derfor til at udfolde og forklare nogle af de mere generelle udfordringer, som profileringsmodellerne rejser i forhold til diskriminations-forbuddet selvom EU-retten kun finder anvendelse på en afgrænset del af det offentliges mange sagsområder.

I EU-retten såvel som i andre regelsæt om diskrimination skelnes der mellem direkte og indirekte forskelsbehandling.

Direkte forskelsbehandling foreligger, hvis en person på grund af et beskyttet kendetegn som køn, etnicitet, religion eller tro, handicap, alder eller seksuelle orientering, behandles ringere end en anden i en sammenlignelig situation. Når det fremgår, at grunden til den ringere behandling skyldes et beskyttet kendetegn, vil der være tale om direkte forskelsbehandling, hvilket som udgangspunkt er ulovligt (diskrimination). Inden for EU-retten er der i forskellige direktiver snævret oplistede undtagelsesgrunde for direkte diskrimination. Uden for EU-retten er indgreb i beskyttelsen mod direkte diskrimination mulig, hvis de forfølger et legitimt formål, og hvis indgrebet er egnet, nødvendigt og proportionalt.

Indirekte forskelsbehandling foreligger, hvis en umiddelbart neutral bestemmelse, betingelse eller praksis vil stille en person særligt ufordelagtigt i forhold til andre personer på grund af et beskyttet kendetegn. Indirekte forskelsbehandling er ikke ulovligt, hvis forskelsbehandlingen er begrundet i et legitimt formål og midlerne til at opfylde formålet er egnede og hensigtsmæssige.

I en sag om indirekte forskelsbehandling vil den berørte borger først skulle påvise, at der er tale om, at en gruppe er blevet stillet ringere end andre i en sammenlignelig situation. For at vise, at en gruppe er blevet behandlet ringere,

kræver det grundlæggende, at man kan vise, at en gruppe i den samme situation er blevet behandlet bedre.

Når borgeren har påvist, at der er tale om ringere behandling, vil myndigheden skulle vise, at den indirekte forskelsbehandling forfølger et sagligt formål med midler, der er nødvendige for og egnede til at opnå målet. Hvis myndigheden ikke kan påvise dette, vil der være tale om ulovlig forskelsbehandling (diskrimination).

Hvilke formål, der kan blive anset som saglige, vil afhænge af konteksten, diskriminationsgrunden og samfundsområdet. EU-Domstolens praksis illustrerer, at der som udgangspunkt foretages en meget intensiv prøvelse af om forskelsbehandlingen kan retfærdiggøres. Domstolen har dog givet udtryk for, at der gives en vid skønsmargin på området for socialpolitik og beskæftigelse.[1]

Forbuddet mod ulovlig forskelsbehandling (diskrimination) udgør ikke altid en klar og let anvendelig regel. Især indirekte forskelsbehandling kan være vanskeligt at bevise og baseres på vage kriterier, som kan føre til forskelligartede udfald fra sag til sag. Dette er vigtigt at holde sig for øje, når profileringsmodeller introduceres i sagsbehandlingen.

HVORDAN FORSKELSBEHANDLER EN PROFILERINGSMODEL?

En profileringsmodel virker ved, at modellen vurderer borgere forskelligt baseret på forskelle i personernes variabelværdier. Modellen klassificerer borgerne baseret på deres individuelle oplysninger, og en af de centrale udfordringer ved profileringsmodeller er derfor risikoen for, at modellen udøver ulovlig forskelsbehandling eller diskrimination. I det følgende skitserer vi, hvordan henholdsvis direkte og indirekte forskelsbehandling konkret kan opstå i modellens funktioner.

DIREKTE FORSKELSBEHANDLING

Direkte forskelsbehandling af en beskyttet gruppe forekommer, når modellen anvender en **variabel**, som kendetegner gruppen, til at vurdere **målegenskaben**.[2] Modellen kan f.eks. lægge vægt på personers køn, så det at være mand øger sandsynligheden for at blive klassificeret positivt, mens det at være kvinde øger sandsynligheden for at blive klassificeret negativt (eller omvendt).

Et beskyttet kendetegn kan sagtens optræde som en variabel i modellens **træningsdatasæt** eller funktion, uden at der af den grund er tale om diskrimination. Risiko for direkte diskrimination opstår kun i de tilfælde, hvor den givne variabel har en effekt på modellens vurdering.

Der er flere måder, hvorpå man kan forhindre, at modellen kommer til at diskriminere relevante grupper direkte. En profileringsmodel vil som udgangspunkt behandle alle variable – dvs. oplysninger om borgeren – ens, uanset om de udgør beskyttede kendetegn eller ej, og hvis et af disse kendetegn har en stærk

sammenhæng med en målegenskab vil **læringsalgoritmen** træne en model, der risikerer at udøve direkte diskrimination på baggrund af dette kendetegn.

Den enkleste metode til at undgå dette er, at læringsalgoritmen ikke anvender de relevante egenskaber. Det gør man ved ganske enkelt at sørge for, at beskyttede kendetegn ikke optræder som variable i træningssættet. En læringsalgoritme kan ikke træne en model på variable, der ikke findes i træningsdata og dette garanterer, at modellen ikke direkte diskriminerer de relevante grupper.

En alternativ metode er at træne en blindet model gennem såkaldt særskilte læringsprocesser (eng. "disparate learning processes").[3] En særskilt læringsproces betyder, at læringsalgoritmen underlægges begrænsninger, således at den træner en model, der ikke anvender de variable, der udgør beskyttede kendetegn, selvom disse variable findes i træningssættet.

Begge metoder rejser udfordringer.

For det første kan en blindet model have højere fejlrate end modeller, som anvender de pågældende variable. En læringsalgoritme træner en model til at anvende netop de variable på netop den måde, som gør modellen i stand til bedst muligt at vurdere målegenskaben. Finder den som led i sin analyse en sammenhæng mellem køn og afslag på tilskud eller etnicitet og risiko for langtidsledighed vil den uden videre anvende disse beskyttede kendetegn. Udelukker man anvendelsen af kendetegnene kan det komme til at reducere modellens overordnede evne til at foretage korrekte vurderinger for alle befolkningsgrupper. Den vil med andre ord begå flere fejl samlet set.

For det andet er det en væsentlig udfordring, at ingen af metoderne udelukker indirekte forskelsbehandling, og i visse tilfælde kan gøre denne type forskelsbehandling endnu værre. Vi ser nærmere på dette under afsnittet om særegne algoritmiske diskriminationsrisici.

INDIREKTE FORSKELSBEHANDLING

En profileringsmodel kan også indirekte forskelsbehandle grupper. Indirekte forskelsbehandling betyder i denne sammenhæng, at en model i praksis stiller en gruppe ringere end andre, selvom modellen ikke arbejder med en variabel, der kendetegner den beskyttede gruppe. Indirekte forskelsbehandling af en gruppe kan derfor forekomme selvom det beskyttede kendetegn ikke figurerer i træningssættet eller modellen er gjort blind overfor gruppen.

Et eksempel på dette kunne være sammenhængen mellem forældremyndighed og køn. Forældremyndighed kan være ulige fordelt på tværs af køn, og selv hvis man konstruerer et træningssæt som ekskluderer køn for at undgå direkte diskrimination, kan anvendelsen af forældremyndighed som variabel i praksis komme til at fordele effekterne af modellen ulige på tværs af køn. Et andet eksempel er sammenhængen mellem bopæl og etnicitet. Hvis borgere med en

bestemt etnicitet er koncentreret i bestemte boligområder, kan modellen, når den forskelsbehandler på baggrund af bopæl, have forskellig effekt for personer med forskellig etnicitet.

Indirekte forskelsbehandling kan opstå som et resultat af valget af målegenskab, når målegenskaben er ulige fordelt på tværs af relevante grupper. Hvis en model f.eks. forsøger at vurdere en målegenskab, som oftere findes hos kvinder end hos mænd, vil den klassificere flere kvinder end mænd positivt. I disse tilfælde skyldes forskelsbehandlingen ikke selve modellen, men snarere dét at lægge målegenskaben til grund for, hvordan personer behandles. Den samme type indirekte forskelsbehandling kunne forekomme, hvis myndigheden lagde den pågældende målegenskab til grund i den manuelle sagsbehandling.

Derimod skyldes indirekte forskelsbehandling selve profileringsmodellen, når modellen finder nye sammenhænge mellem variable og relevante grupper. En profileringsmodel kan bestå af mange dele, såsom beslutningspunkter eller variable. Hvert af disse elementer kan påvirke, hvordan den samlede model behandler grupper. I nogle tilfælde kan de måder forskellige dele af modellen virker på udligne hinanden, så modellen samlet set ikke forskelsbehandler. Samspillet mellem modellens enkelte dele kan også have den modsatte effekt – når flere forskellige dele på samme tid forskelsbehandler en gruppe kan modellens samlede forskelsbehandling af gruppen forstærkes. Det er derfor vigtigt at se på, hvordan en model samlet behandler hver af de relevante grupper.

Når modellens sammenhænge fører til forskelsbehandling skyldes det enten, at beskyttede og ikke-beskyttede grupper har forskellige værdier for visse variable eller at sammenhængen mellem variable og målegenskab varierer på tværs af grupperne.

Et eksempel på det første er, at en model, som anvender personers højde til at vurdere en målegenskab vil kunne have forskellig effekt for mænd og for kvinder, af den simple grund at mænd i gennemsnit er højere end kvinder.

Et eksempel på det andet er, hvis der i modellen er en positiv sammenhæng mellem længden af en persons uddannelse og personens forventede livstidsindkomst (målegenskaben), men at sammenhængen er stærkere (eller svagere) for kvinder end den er for mænd.

BEVISVANSKELIGHEDER

Det kan være vanskeligt at bevise indirekte diskrimination og disse vanskeligheder risikerer at blive forstærket ved brugen af en profileringsmodel.

Overordnet er grunden til dette, at der er et vist fortolkningsrum i af forbuddet mod indirekte diskrimination, så at forbuddet kan tilpasses det konkrete scenarie.[4] I forhold til algoritmisk profilering skaber det imidlertid den ulempe, at forbuddet ikke kan sættes på en formel og indlejres i modellens design uden videre.

Forbuddet skaber fortolkningsrum på en række forskellige punkter. Det gælder f.eks. vurderingen af sammenlignelighed (dvs. hvilke grupper, der skal sammenlignes og hvordan)

F.eks. fandt EU-Domstolen i en dom, om hvorvidt orlovsperioder skulle medregnes i fratrædelsesgodtgørelser, at forældreorlov (som hovedsageligt blev taget af kvinder) og orlov i forbindelse med militærtjeneste (overvejende taget af mænd) ikke var sammenlignelige situationer.[5] Dommen er blevet kritiseret ud fra den begrundelse, at sammenligningen ikke burde vedrøre de to former for orlov, men derimod en sammenligning af, hvilke typer af orlov mænd og kvinder typisk tager.[6] Dommen og dens kritik illustrerer, at sammenligning indebærer afvejninger, som kan foretages på forskellige måder. Det er derfor ikke altid sikkert, hvilke grupper, der skal sammenlignes og hvad sammenligningen skal vedrøre.

Anvender vi eksemplet på en model, der skal beregne beløbet på fratrædelsesgodtgørelser, ville man – hvis man fulgte kritikken af dommen – under modellens udvikling skulle sikre sig, at man udvalgte orlovstyper, der var kønsmæssigt repræsentative. Ville man i stedet følge dommens resultat, ville det ikke være nødvendigt at forholde sig til kønsfordeling på tværs af forskellige orlovstyper.

Sammenlignelighedsvurderingen vil altid være en konkret vurdering. Den vil således tage udgangspunkt i, at grupperne skal være i en sammenlignelig situation i forhold til det konkrete spørgsmål, som sagen handler om. Denne usikkerhed i, hvad der i det konkrete tilfælde vil kunne udgøre den rette sammenligning gør det svært at helgardere sig mod risikoen for indirekte forskelsbehandling i udvikling af en model.

Også kravet om ringere behandling er en konkret vurdering og EU-Domstolen har fundet, at det ikke er nødvendigt at bevise, at personer reelt er blevet behandlet ringere, hvis blot tiltaget hypotetisk ville stille nogle ringere på grund af f.eks. deres etnicitet.[7] EU-Domstolens praksis her er dog langt fra entydig og Domstolen har i en anden sag anført, at undersøgelsen af, om der forelå ufordelagtig behandling, skulle ske "specifikt og konkret i lyset af den pågældende gunstige behandling og ikke generelt og abstrakt"[8] Dette illustrerer ligeledes besværlighederne i at tænke diskriminationsforbuddet ind i modellens design.

Særligt kan det være svært at beslutte, hvilke tal der skal sammenlignes for at vurdere om en gruppe er stillet ringere. Det kan give meget forskellige resultater alt efter, hvilke grupper man vælger at sammenligne og hvordan man opgør tallene. Domstolen har udtalt i en sag om kvindelige arbejdstagere, at:

> "den bedste metode til sammenligning af statistiske oplysninger består i at foretage en sammenligning mellem, på den ene side, andelene af de mandlige arbejdstagere, som opfylder - og ikke opfylder - betingelsen efter den anfægtede regel om to års forudgående ansættelse, og, på den anden side, de tilsvarende andele for kvindelige arbejdstageres vedkommende. Det er ikke tilstrækkeligt at betragte antallet af berørte personer, da dette antal afhænger af antallet af aktive arbejdstagere i den pågældende medlemsstat som helhed samt fordelingen af mandlige og kvindelige arbejdstagere i denne medlemsstat."[9]

Den algoritmiske pendant til besværlighederne finder vi i valget af, hvorledes forskelsbehandling overhovedet skal måles i en model (se nedenfor under "Særegne algoritmiske diskriminationsrisici").

Som det fremgår, kan forbuddet mod indirekte ulovlig forskelsbehandling ikke defineres entydigt. Der er variationer i vurderingen på tværs af både samme og forskellige sagstyper og beskyttelsesområder. Som vi har behandlet i kapitel 6, kan regler med fortolkningsrum være svære at sætte på formel. Hertil kommer, at det kan vise sig vanskeligt at bevise en sammenhæng mellem modellens variable og et beskyttet kendetegn fordi modellen kan anvende såkaldte proxyinformationer. Dette forøger de allerede eksisterende bevisvanskeligheder og gennemgås i det følgende.

PROXYINFORMATIONER I PROFILERINGSMODELLER

Profileringsmodeller kan som nævnt finde nye sammenhænge mellem variable og relevante grupper, som – selvom de givne variable ikke dækker over beskyttede kendetegn –kan føre til diskrimination. Dette kan ske i et kompleks samspil mellem modellens forskellige dele, som tilsammen både kan udligne eller forstærke risikoen for, at modellen samlet set udøver indirekte diskrimination.

Fænomenet bygger på, hvad der inden for maskinlæring kaldes "redundant encoding", og indebærer, at beskyttede kendetegn (eventuelt uforvarende) bliver "indkodet" i andre egenskaber.[10] For at kunne vurdere risikoen for indirekte diskrimination er der derfor nødvendigt at identificere disse egenskaber, der er proxyinformationer, dvs. "stedfortrædende" informationer for et beskyttet kendetegn.

I nogle tilfælde kan man relativt let identificere egenskaber, der er typiske proxyinformationer – f.eks. sammenhæng mellem højde og køn. Men netop i algoritmiske modeller kan proxyinformationerne ikke nødvendigvis identificeres,

fordi modellen er i stand til at finde fjernere og mere komplekse statistisk pålidelige, men menneskeligt ulogiske sammenhænge.[11] Dette har vi berørt i kapitel 6.

Bevis for indirekte forskelsbehandling vanskeliggøres desuden af, at modellen ikke altid er algoritmisk transparent. Manglende transparens og ulogiske sammenhænge er to forskellige udfordringer. Det første gør det umuligt eller vanskeligt at få indblik i modellen. Det andet indebærer derimod, at selv hvor det er muligt at få et sådant indblik forekommer modellens sammenhæng ulogisk eller absurd.

For at konstatere indirekte diskrimination skal man vide, hvordan modellen vurderer alle personer med og uden en eller flere (mere eller mindre logiske) proxyinformationer. Dette kræver indsigt i, hvordan modellens mange dele samlet set finder anvendelse på alle personer – det, vi i kapitel 7 har kaldt global, holistisk transparens. Hvis en model ikke er global, holistisk transparent (enten i form af, at den kan **fortolkes** eller **forklares**), vil det være vanskeligt at konstatere indirekte forskelsbehandling. Det kan imidlertid være svært at opnå dette niveau af transparens, hvor man kan få indsigt i hele modellens funktion for alle personer. Som et alternativ kan der laves enkle statistiske analyser af, hvordan modellen vurderer et sæt af personer med og uden en eller flere identificerede proxyinformationer.

EU-retten kan tjene som inspiration til løsning af udfordringerne med proxyinformationer.

Inden for EU-retten er der således i nogle tilfælde fastsat regler, som kræver delt bevisbyrde. Dette er tilfældet for en række af direktiverne, hvor det fastslås, at hvis borgeren fremfører omstændigheder, der "giver anledning til at formode" forskelsbehandling, så skal myndigheden bevise, at ligebehandlingsprincippet ikke er tilsidesat.[12]

Også EU-Domstolen har i sin praksis fastsat regler om bevis. I en dom om ligeløn har EU-Domstolen udtalt, at uigennemskuelige systemer for lønordninger, hvor den berørte borger ikke er i stand til at vide, hvilke kriterier, der anvendes over for vedkommende, og hvordan de anvendes, betyder, at der er en bevismæssig formodning for diskrimination.[13] EU-Domstolen fremhævede i en sag om en kønsdiskriminerende lønordning, at en ordning, der ikke er gennemskuelig som udgangspunkt er i strid med princippet om lige adgang til beskæftigelse, idet den manglende gennemskuelighed forhindrer enhver form for domstolskontrol. Det var derfor op til arbejdsgiveren at bevise, at den førte lønpolitik ikke systematisk stillede kvindelige lønmodtagere ringere ved at oplyse hvordan man anvendte kriterier for personlige tillæg og dermed gøre systemet gennemskueligt.[14] Det var dog fortsat en betingelse, at den berørte person i sagen fremlagde materiale, der godtgjorde, at gennemsnitslønnen for kvinder var lavere end for mænd.

Selvom dommen ikke handler om algoritmiske profileringsmodeller illustrerer den nogle af de pointer, som vi har behandlet i kapitel 7 om transparens og begrundelsespligt.

Vores anbefalinger i kapitel 4 om, at der eksplicit tages stilling til diskriminationsforbuddet i AI-konsekvensanalyserne i kombination med krav om tilsyn, den begrænsede brug af fuldautomatisering (kapitel 6) og krav til transparens (kapitel 7), bidrager efter vores vurdering alle til at adressere bevisvanskelighederne ved proxyinformationer.

TRE TYPER AF DISKRIMINATIONSRISICI

Vi vender os nu mod de tre typer af diskriminationsrisici, vi skitserede i kapitlets indledning: Videreførelse af eksisterende diskrimination; algoritmisk forstærket risiko for diskrimination og endelig særegne algoritmiske diskriminationsrisici.

VIDEREFØRELSE AF EKSISTERENDE DISKRIMINATION

En læringsalgoritme træner modellen på det træningsdatasæt den har til rådighed, uanset hvilken kvalitet datasættet har. Det betyder at forskelsbehandling kan opstå som resultat af **bias** i træningsdatasættet, som skyldes enten den praksis der har genereret data, registreringen af data eller sammensætningen af datasættet. Den første type af diskriminationsrisiko er derfor, at maskinlæring kan videreføre eksisterende diskrimination.

Bias er en tendens i datasættet til systematisk at misrepræsentere sammenhænge mellem variable og målegenskab, og det er således udtryk for, at data i hvert fald i én henseende har lav kvalitet. Hvis de data som anvendes til træning af modellen, eller de data som den behandler når den anvendes, har lav kvalitet, så vil det føre til vurderinger af tilsvarende lav kvalitet.

Lav datakvalitet giver imidlertid ikke i sig selv modellen tendens til at forskelsbehandle. For at føre til forskelsbehandling skal svaghederne i data hænge systematisk sammen med medlemskab af en eller flere beskyttede grupper. Bias kan indebære diskrimination, men kan også mere generelt føre til usaglige afgørelser (se kapitel 6).

Diskrimination kan videreføres på tre forskellige måder:

For det første, kan modellen trænes på data, der afspejler en ulovlig forskelsbehandling af beskyttede grupper i myndighedens praksis. I så fald vil modellen indarbejde denne praksis i sine vurderinger. Hvis myndigheden f.eks. historisk set har haft tendens til oftere at afsige indgribende afgørelser mod borgere med et bestemt køn, alder eller etnicitet, vil træningsdatasættet afspejle denne forskel, som kan blive videreført i modellens vurderinger. Dette kan føre til diskrimination.

For det andet, kan kvaliteten af indsamlet data variere, afhængigt af hvilke grupper det handler om. I dette tilfælde kan datasættet misrepræsentere

sammenhænge mellem variable og målegenskab for nogle grupper, hvilket vil give modellen tendens til forskellig fejlrate for de pågældende grupper og kan føre til diskrimination.

For det tredje, kan der være er systematiske forskelle i, hvem der inkluderes i træningssættet, hvor nogle grupper over- og andre underrepræsenteres. Hvis der er relevante forskelle mellem de grupper som henholdsvis over- og underrepræsenteres, kan læringsalgoritmen udlede sammenhænge, der varierer fra de sammenhænge, som optræder i den befolkning, som modellen skal anvendes på. Det vil give modellen tendens til lav **generaliserbarhed**, og forskellig fejlrate for de to grupper.

For at mindske risikoen for videreførelse af diskrimination er det absolut afgørende at stille krav til datakvalitet (kapitel 5) og tilsyn (kapitel 4).

ALGORITMISK FORSTÆRKET DISKRIMINATIONSRISIKO

En profileringsmodel kan ikke bare videreføre, men også forstærke risikoen for diskrimination. Dette er den anden type af diskriminationsrisiko, som vi behandler.

For at illustrere fænomenet kan vi bruge et meget omtalt eksempel fra litteraturen. Her trænede en gruppe forskere en model til at identificere kønnet på personer i billeder. Til det formål brugte de med vilje et træningsdatasæt, der overrepræsenterede kvinder i køkkener: Af de personer, der var afbildet i køkkener, var to ud af tre kvinder, mens en ud af tre var mænd. Fordi køkkenet udgjorde en letgenkendelig og signifikant variabel, tillagde modellen det imidlertid langt større betydning i sin vurdering af personens køn. Resultatet var en model, der i 84% af tilfældene klassificerede personen i et køkken som en kvinde – selvom tallet i træningsdatasættet kun var 67% Selv relativt små forskelle mellem grupper i et træningssæt kan altså føre til en model, som behandler de to grupper meget forskelligt.

Diskriminationsrisikoen kan også blive forstærket gennem såkaldte **n**egative feedbacksløjfer (se kapitel 5). Hvis modellen, som nævnt foroven trænes på data, der indeholder et element af forskelsbehandling, så vil modellen risikere at diskriminere. Modellens vurdering kan herefter indgå i en myndigheds afgørelser, og disse kan risikere at føre til diskrimination. Når myndighedens afgørelser bruges som input til at træne og opdatere modellen, får modellen endnu stærkere tendens til diskrimination (se eksempel i boks).

Udfordringen ved gamle og misvisende træningsdata kan løses ved at fastsætte krav om periodisk gentræning af data (se kapitel 5). Men det kendetegnende for negative feedbacksløjfer er, at det er vanskeligt at måle eller at opdage forkerte, herunder diskriminerende resultater i modellen og at gentræning med nye data derfor ikke løser problemet. Også af denne grund bør myndighedens tidligste AI-konsekvensanalyser adressere, hvorledes risikoen vil blive imødegået (se kapitel 4 og 5).

NEGATIV FEEDBACKSLØJFE: PRÆDIKTIVT POLITIARBEJDE I USA

Et af de mest omdiskuterede eksempler på negative feedbacksløjfer knytter sig til såkaldt predictive policing i USA. Predictive policing indebærer, at politiet bruger en algoritmisk model til at vurdere, hvordan de skal fokusere deres ressourcer, f.eks. hvilke områder, der skal patruljeres mere eller mindre intensivt.[15] Imidlertid er der blevet udtrykt bekymring for, at de data, som modellerne trænes på, i mange tilfælde afspejler politiets langvarige diskrimination af sorte borgere, f.eks. i form af en tilbøjelighed til i højere grad at undersøge og anholde sorte borgere. Er det tilfældet, vil datasættet indeholde en bias, som afspejler denne forskelsbehandling, og læringsalgoritmen kan træne en model med tendens til at vurdere, at politiets ressourcer bør fokuseres på at patruljere områder med høj koncentration af sorte borgere. Hvis modellens vurderinger fører til, at flere ressourcer koncentreres på at patruljere disse områder, så kan bias i data forstærkes: endnu flere interaktioner mellem politi og borgere vil være koncentreret i disse områder. Det fører til at en opdateret model får endnu stærkere tendens til at forskelsbehandle, ved at vurdere at politiet i endnu højere grad bør fokusere på de pågældende områder.[16]

SÆREGNE ALGORITMISKE DISKRIMINATIONSRISICI

Det forhold, at profileringsmodeller kan føre til diskrimination, og at dette kan ske på forskellige måder

har været erkendt og behandlet i forskningen i over et årti[17], og der findes i dag en rig litteratur som udforsker tekniske løsninger, der forhindrer eller reducerer forskelsbehandling i algoritmisk profilering (se boks).

En central pointe er, at brugen af disse teknikker indebærer flere valg i modellens udformning. Dette er den tredje og mest komplekse type af diskriminationsrisici. Vi har tidligere kaldt disse designvalg for algoritmiske dilemmaer, for de er alle kendetegnet ved, at der ikke altid er nogen klar og entydig løsning, der fuld ud beskytter borgernes rettigheder og retssikkerhed. Tværtimod rummer designvalgene hver en potentiel konflikt mellem forskellige beskyttelseshensyn, der alle er lige vigtige, men ikke alle kan blive tilgodeset i lige høj grad i modellen. Dette fører til en af de mest alvorlige udfordringer ved profileringsmodeller og er derfor også en af de vigtigste grunde til vores anbefalinger om begrænset brug af fuldautomatisering (se kapitel 6), gennemførelsen af AI-konsekvensanalyser og effektivt tilsyn (se kapitel 4).

TEKNISKE LØSNINGER FOR AT MINDSKE DISKRIMINATIONSRISIKO I MODELLEN

Overordnet kan de tekniske løsninger for at mindske risikoen for diskrimination inddeles i metoder til at ændre på træningssættet (præ-processering), metoder til at ændre på træningsprocessen (in-processering), og metoder til at ændre på modellens vurdering (post-processering).[18] Mange af teknikkerne kan naturligvis også kombineres. Studier, der sammenligner de tekniske løsninger, indikerer imidlertid, at der ikke er nogen løsning, som dominerer de andre, forstået som, at den er mindst lige så god som alternative løsninger i hver af disse henseender.[19]

Præ-processering indebærer, at der ændres på træningssættet for at mindske eller forhindre forskelsbehandling, f.eks. ved at fjerne relevante variable i træningssættet. En mere sofistikeret løsning er at redigere i data for at svække eller fjerne de sammenhænge i træningssættet, som ville føre forskelsbehandling af relevante grupper.[20] Overordnet kræver denne løsning en metode til at identificere relevante sammenhænge i datasættet, en metode til at redigere i datasættet for at reducere sammenhængene, og en metode til at identificere de ændringer, som vil føre til den mindst mulige stigning i fejlraten. Det sidste element er afgørende, fordi enhver ændring af værdierne i træningssættet vil have en omkostning for fejlraten, men nogle ændringer kan have langt større effekt på fejlraten end andre.

In-processering indebærer, at der ændres på træningsprocessen, f.eks. ved at give læringsalgoritmen mulighed for at repræsentere interaktioner mellem variable (via **polynomiske variable**, se kapitel 3).[21] Den resulterende model vil, lidt forenklet, arbejde med variable, der repræsenterer sammenhængen med målegenskaben for hver af de relevante grupper (f.eks. højde/kvinder og højde/mænd fremfor højde og køn som forskellige variable). Det betyder, at modellen i nogle tilfælde kan mindske forskelle i fejlraten. En mere sofistikeret form for in-processering er at ændre på læringsalgoritmens **tabsfunktion**, så læringsalgoritmen træner en model, som i mindre grad forskelsbehandler. Udvikleren kan f.eks. definere hårde grænser for, hvilke modeller læringsalgoritmen kan træne, men det er også muligt at tilføje en **regulariseringsfaktor**, som giver modellens forskelsbehandling en negativ værdi i træningsprocessen.[22]

Post-processering indebærer, at der ændres på den måde, modellen vurderer personer. En af de centrale opgaver i udviklingen af en klassificeringsmodel er at definere en værdi for beslutningstærsklen (se kapitel 3), og typisk vil udvikleren vælge en tærskelværdi som minimerer modellens fejlrate. Tærsklen kan imidlertid også defineres sådan, at modellens indirekte forskelsbehandling reduceres, f.eks. ved at vælge en tærskelværdi, hvor modellen har samme fejlrate for to grupper.[23]

I det følgende ser vi nærmere på fem designvalg. Disse designvalg indebærer en risiko for, at et beskyttelseshensyn i modellen beskyttes på bekostning af et andet: valget mellem forskellige måder at måle forskelsbehandling; valget mellem hensyn til forskellige beskyttede grupper; valget mellem beskyttelse mod risikoen for direkte versus indirekte diskrimination; valget mellem beskyttelse mod diskriminationsrisici versus målet om en lav fejlrate; og endelig hensynet til diskriminationsforbuddet versus beskyttelsen af personoplysninger.

Alle disse designvalg skal træffes på en måde som sikrer, at de rettighedsrisici, som modellen ikke er i stand til at beskytte (eller beskytter i ringere grad) er kende og bliver fjernet i myndighedens faktiske brug af modellen.

MÅDER AT MÅLE FORSKELSBEHANDLING I EN MODEL

Forskelsbehandling indebærer lidt forenklet, at en person stilles ringer end andre ikke-beskyttede grupper enten direkte eller indirekte på grundlag af et bestemt kendetegn.

Anvendelsen af de tekniske løsninger beskrevet ovenfor forudsætter i næsten alle tilfælde, at der anvendes en matematisk præcis definition af, hvordan forskelsbehandlingen måles. Der ligger derfor et centralt designvalg heri.[24]

I modellen kan man enten måle forskelsbehandling som en forskel i vurderingen af borgeren på baggrund af beskyttede kendetegn (direkte eller indirekte). Dette er den mest umiddelbare form for forskelsbehandling og indebærer, at en borger får afslag eller en bebyrdende afgørelse direkte eller indirekte på grund af et beskyttet kendetegn. Måske kunne et afslag eller en bebyrdende afgørelse være berettiget og lovlig af andre grunde, men pointen er, at modellen pr. automatik vurderer borgeren ringere på baggrund af det beskyttede kendetegn (eller en proxyinformation).

Som en variation af dette kan forskel også måles på forskelle i fordelingen af fejlraten. Her måles, om man som beskyttet gruppe oftere fejlklassificeres end den ikke-beskyttede gruppe og dermed får et afslag eller en bebyrdende afgørelse oftere fordi modellen er dårligere til at vurdere sager for den beskyttede gruppe.

Endnu en måde at måle forskel på er ud fra modellens tendens til at begå forskellige typer fejl, når den vurderer personer knyttet til forskellige grupper.[25] En model kan begå fejl på flere måder og det er ofte væsentligt, om det er den ene eller den anden type fejl, som modellen begår (se kapitel 3 og 6). Når en model begår fejl, er det så f.eks. fordi den overvurderer eller undervurderer værdien for målegenskaben? Og har den tendens til at overvurdere værdien for én gruppe, men til at undervurdere den for andre grupper?

Udfordringen består i, at det kan være vanskeligt på samme tid at fjerne risikoen for diskrimination målt ud fra de forskellige måder. Det skyldes dels, at udviklingen af en model teknisk er vanskeligere, hvis den skal tage flere forskellige hensyn på

samme tid, og dels at det i mange tilfælde er matematisk umuligt for modellen samtidigt at undgå forskelsbehandling i alle henseender.[26] Denne udfordring kan illustreres af den såkaldte COMPAS-sag (se boks).

FORSKELSBEHANDLING GENNEM FORSKEL I FEJLTYPER: COMPAS-SAGEN

Måske den mest prominente offentlige debat om brugen af profileringsmodeller har knyttet sig til brugen af COMPAS-software i det amerikanske retsvæsen. COMPAS er en profileringsmodel udviklet af Northpointe (i dag Equivante), som er designet til at vurdere hvor sandsynligt det er, at en sigtet eller straffet person begår kriminalitet igen. Den anvendes i mange stater i forbindelse med bl.a. beslutninger om varetægtsfængsling og prøveløsladelse.

COMPAS vurderer en persons risiko for at begå kriminalitet på ny baseret dels på registeroplysninger og dels på personens svar på et spørgeskema.[27] Modellen udtrykker vurderingen som en score mellem nul og ti. Scoren repræsenterer modellens vurdering af, hvilken risikokategori den pågældende person tilhører– en score på 10 betyder at personen hører til de 10% af personer, som har den højeste risiko, mens en score på 5 betyder at personen har en gennemsnitlig risiko.

Den amerikanske NGO ProPublica offentliggjorde i 2016 en kritisk analyse af COMPAS, baseret på data fra strafudmålinger i Broward County, Florida.[28] ProPublicas analyse anførte, at COMPAS systematisk vurderede sorte borgere anderledes end hvide borgere:

- Sorte borgere, som ikke blev anholdt igen inden for to år, blev af COMPAS klassificeret som havende medium til høj risiko for at begå ny kriminalitet næsten dobbelt så ofte, som hvide borgere, der ikke blev anholdt igen (45% vs. 23%). Det vil sige, at COMPAS lavede relativt flere falsk positive fejl i vurderingen af den første gruppe, end i vurderingen af den anden.
- Hvide borgere, som blev arresteret igen inden for to år, blev klassificeret som havende lav risiko næsten dobbelt så ofte, som sorte borgere, der begik nye forbrydelser (48% vs. 28%). Det vil sige, at COMPAS lavede relativt flere falsk negative fejl i vurderingen af den første gruppe, end i vurderingen af den anden.

Målt på fejltype blev sorte borgere derfor forskelsbehandlet, da der var forskel i fejltyperne for henholdsvis sorte og hvide borgere, og fejltypen falsk positiv var den mest indgribende og dermed stillede befolkningsgruppen ringere.

Northpointe og uafhængige forskere har forsvaret COMPAS ved bl.a. at pege på, at etnicitet ikke optræder som variabel i modellen (modellen diskriminerer ikke direkte) samt at modellens fejlrate var ca. 35% for begge grupper (målt på fejlrate var der ingen forskelsbehandling).[29]

Selvom der skal foretages et valg i modellens design af, hvordan man vil lade modellen måle forskelsbehandling, kan myndigheden selvsagt ikke vælge, hvilken type diskrimination de vil beskytte mod og hvilken, den ikke vil. Der er derfor en grundlæggende forskel i, hvad modellen er i stand til, og hvad myndigheden er forpligtet til. Det afgørende i diskriminationsforbuddet er ikke, hvordan den enkelte er stillet ringere men at den enkelte er stillet (urimeligt og uproportionelt) ringere: dette kan både være på grund af modellens vurdering af vedkommende; fordi modellen for den enkelte har en dårligere kvalitet og derfor oftere begår fejl; og fordi modellen oftere begår flere alvorlige fejl for den beskyttede gruppe. Der er ikke tale om beskyttelse som enten-eller men som både-og.

Rettighedsmæssigt må modellens begrænsninger således ikke føre til, at myndigheden stiller borgeren ringere.

FORSKELLIGE BESKYTTEDE GRUPPER

En model kan på samme tid indirekte forskelsbehandle flere grupper, f.eks. en gruppe på baggrund af køn og en anden gruppe på baggrund af etnicitet. Myndigheden skal naturligvis beskytte alle grupper mod diskrimination men for profileringsmodellen bliver dette vanskeligt og det betyder, at rettigheder ikke til fulde kan tænkes ind i modellens design. Mange af de tekniske løsninger vi har nævnt foroven kan kun reducere forskelsbehandling for én gruppe ad gangen, ellers oplever man, at fejlraten vokser dramatisk, hvis løsningen anvendes til samtidigt at reducere ulighed for flere grupper. I sådanne tilfælde er man tvunget til at vælge, hvordan beskyttelsen af forskellige grupper skal prioriteres i modellen.

Den menneskeretlige – herunder i den EU-retlige – beskyttelse mod forskelsbehandling indebærer en vis prioritering. Køn, handicap og etnicitet nyder således mere omfattende beskyttelse end f.eks. alder eller sociale tilhørsforhold. Særreguleringen for de enkelte beskyttede grupper bør inddrages i udformningen af modellen og kan komme til at indebære, at modellen behandler de beskyttede grupper forskelligt afhængig af beskyttelsesområdet for den enkelte gruppe.

Men det er noget andet end det algoritmiske problem, som består i, at modellen slet ikke er i stand til at beskytte flere grupper endsige sikre nuancerne i beskyttelsesområdet for de forskellige beskyttede grupper.

DIREKTE VERSUS INDIREKTE FORSKELSBEHANDLING

Visse tekniske løsninger kræver også en afvejning mellem direkte forskels-behandling og indirekte forskelsbehandling.

Som vi skitserede i starten af dette kapitel, er det muligt at forhindre direkte diskrimination i en træningsmodel ved at "blinde" den for de variable om beskyttede kendetegn. Denne løsning udelukker dog samtidig anvendelsen af nogle tekniske løsninger, som kan reducere indirekte forskelsbehandling. Omvendt involverer anvendelsen af mange af de tekniske løsninger, som reducerer indirekte

forskelsbehandling, at relevante grupper direkte forskelsbehandles, enten af modellen, eller af læringsalgoritmen under træningen af modellen.

Også her er modellen således teknisk begrænset på en måde, som myndigheden er nødt til at kende til og løse for at sikre overholdelsen af diskriminationsforbuddets fulde anvendelsesområde.

DISKRIMINATION VERSUS FEJLRATE

En teknisk løsning, som reducerer en models diskriminationsrisiko målt og forstået på den ene eller anden måde, har i mange tilfælde en omkostning i form af en stigning i modellens samlede fejlrate.[30] Det skyldes, at de fleste tekniske løsninger involverer ændringer i data, som gør data mindre repræsentative, eller ændringer i modellens behandling af data, som reducerer modellens evne til præcist at repræsentere sammenhængene mellem variable og målegenskab. I disse tilfælde fører anvendelsen af en løsning alt andet lige til, at modellen laver flere fejl i vurderingerne, end den ellers ville. Det betyder, at en model, der beskytter mod forskelsbehandling, kan foretage forkerte vurderinger for alle borgere på tværs af befolkningsgrupper og dermed blive retssikkerhedsmæssigt problematisk på et andet område end diskrimination.

Her er designvalget en afvejning af diskrimination versus fejlrate, dvs. mellem, om modellen skal beskytte mod diskrimination, eller om den skal sikre, at der for samtlige befolkningsgrupper begås færrest mulige fejl.

Myndighederne er nødt til også at kende til disse afvejninger, adressere dem i AI-konsekvensanalyserne og sikre, at modellens faktiske brug i sagsbehandlingen ikke fører til en forringelse af borgerens retsstilling.

FORHOLDET MELLEM DISKRIMINATIONSFORBUDDET OG DATABESKYTTELSE

Databeskyttelsesforordningen er på mange måder med til at sikre en effektiv gennemførelse af diskriminationsforbuddet (og øvrige rettigheder efter i EU-retten).[31]

Der kan dog opstå tilfælde, hvor databeskyttelsesretten og håndhævelsen af diskriminationsforbuddet kan trække i forskellige retninger. Det er tilfældet, hvis beskyttelsen af personoplysninger hindrer adgang til oplysninger, som kan være nødvendig for at vurdere, om modellen diskriminerer en borger.

Adgangen til oplysninger om beskyttede kendetegn kan navnlig være nødvendige i to tilfælde: For det første for at sikre, at historisk data, som bliver brugt til træning af modellen, ikke indeholder ulovlig forskelsbehandling, som dermed videreføres til modellen, For det andet for at sikre mod indirekte forskelsbehandling i modellen, hvilket nødvendiggør brugen af beskyttede kendetegn i træningsprocessen.
For at sikre mod risikoen for indirekte diskrimination må man nemlig under træningsprocessen lade læringsalgoritmen måle og justere for de beskyttede kendetegn.

Som vi har redegjort for i kapitel 5 skal behandlingen af personoplysninger ske i overensstemmelse med principperne om såvel dataminimering og formålsbestemthed. Samtidig er kendskab til visse personoplysninger om beskyttede grupper nødvendigt for at identificere, om der sker diskrimineres

For så vidt er problemstillingen ikke anderledes end i sager, hvor der ikke anvendes profileringsmodeller, men fordi modeller anvender flere personoplysninger, kan anvende mere avancerede proxyinformationer og rejser særlige diskriminationsrisici, kan det komme til at udgøre et større rettighedsmæssigt problem end ved manuel sagsbehandling.

I EU-Kommissionens udkast til en forordning for kunstig intelligens (se kapitel 2) er der udtrykkeligt taget stilling til samspillet mellem beskyttelsen af personoplysninger i form af beskyttede kendetegn og forbuddet mod diskrimination. Udkastet lægger op til, at sådanne oplysninger kan behandles, hvis der sikres passende sikkerhedsforanstaltninger, herunder tekniske begrænsninger for videreanvendelse og tekniske metoder for beskyttelse af privatliv og personoplysninger.

Vi anbefaler, at myndigheden i AI-konsekvensanalyserne redegør for, hvordan myndigheden har forsøgt at løse udfordringen med de to beskyttelsesområder og at dette underlægges tilsyn (se kapitel 4). Endvidere finder vi, at tekniske løsninger til at beskytte personoplysninger i træningssættet bør udvælges med problemstillingen for øje (se kapitel 5) Hertil kommer, at vi i vores anbefaling om systemisk transparens om træningsdata stiller krav om, at myndigheden fremlægger aggregerede statistikker (hvor oplysninger ikke kan kobles til en bestemt borger og dermed ikke er personhenførbare) om hvordan modellen anvender beskyttede kendetegn til at vurdere en borger (se kapitel 7).

NOTER

1 Se bl.a. EU-Domstolens dom i sag C-411/05, pr. 68
2 Barocas og Selbst (2016) Big Data's Disparate Impact California Law Review 104(671), Kleinberg, Ludwig, Mullainathan og Sunstein (2019) Discrimination in the Age of Algorithms arXiv e-prints
3 Lipton, Chouldechova og McAuley (2018) Does mitigating ML's impact disparity require treatment disparity? 32nd Conference on Neural Information Processing Systems
4 Wachter, Mittelstadt og Russell (2020) Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI, SSRN Electronic Journal. 10.2139/ssrn.3547922
5 EU-Domstolens dom i sag C-220/02 pr. 64
6 Se f.eks. Evelyn og Watson (2012) EU Anti-Discrimination Law, Oxford University Press s. 153f
7 Se f.eks. EU-Domstolens dom i sag C-7/12 og C-81/12 (som dog begge vedrørte direkte diskrimination)
8 EU-Domstolens dom i sag C-457/17, pr. 48. Se også EU-Domstolens dom i sag C-668/15 pr. 31f
9 EU-Domstolens dom i sag C--167/97, pr. 59. Se desuden Wachter, Mittelstadt, og Russell, Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI (March 3, 2020) s. 37ff
10 Pedreschi, Ruggieri og Turini (2008) Discrimination-aware data mining. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, Dwork, Hardt, Pitassi, Reingold og Zemel (2011). Fairness Through Awareness. arXiv:1104.3913 [cs], Barocas og Selbst (2016). Big Data's Disparate Impact, California Law Review 104(671), Hardt, Price og Srebro (2016). Equality of Opportunity in Supervised Learning arXiv:1610.02413 [cs].
11 Selbst og Barocas (2018), The Intuitive Appeal of Explainable Machines 87 Fordham Law Review 1085
12 Se således direktiv 2000/43 artikel 8, direktiv 2000/78, artikel 10, stk. 1, direktiv 2004/113, artikel 9, stk. 1 og direktiv 2006/54, artikel 19, stk. 1
13 EU-Domstolens dom i sag C-109/88, pr. 10ff
14 Hacker, Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law, s. 23f
15 Perry (2013). Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations, RAND Corporation, Ferguson (2017). The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement, NYU

Press, Richardson, Schultz og Crawford (2018). "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice

16 Ensign, Friedler, Neville, Scheidegger og Venkatasubramanian (2017). Runaway Feedback Loops in Predictive Policing. 1st Conference on Fairness, Accountability and Transparency, arXi

17 Et ofte citeret, banebrydende studie i denne henseende er Pedreschi, Ruggieri og Turini (2008). Discrimination-aware data mining. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM

18 Hajian og Domingo-Ferrer (2013) A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. IEEE Transactions on Knowledge and Data Engineering 25(7): 1445-1459, Berk, Heidari, Jabbari, Kearns og Roth (2018). "Fairness in Criminal Justice Risk Assessments: The State of the Art." Sociological Methods & Research Online first, Friedler, Scheidegger, Venkatasubramanian, Choudhary, Hamilton, et al. (2019) A comparative study of fairness-enhancing interventions in machine learning. Proceedings of the Conference on Fairness, Accountability, and Transparency. Atlanta, GA, USA, Association for Computing Machinery: 329–338

19 Friedler, Scheidegger, Venkatasubramanian, Choudhary, Hamilton, et al. (2019) A comparative study of fairness-enhancing interventions in machine learning. Proceedings of the Conference on Fairness, Accountability, and Transparency. Atlanta, GA, USA, Association for Computing Machinery: 329–338

20 Kamiran og Calders (2009) Classifying without discriminating. 2009 2nd International Conference on Computer, Control and Communication, IEEE, Kamiran, F. and T. Calders (2010). Classification with No Discrimination by Preferential Sampling. Proceedings of the 19th Machine Learning conference of Belgium and The Netherlands, Hajian og Domingo-Ferrer (2013) A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. IEEE Transactions on Knowledge and Data Engineering 25(7): 1445-1459, Feldman, Friedler, Moeller, Scheidegger og Venkatasubramanian (2015) Certifying and Removing Disparate Impact. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, NSW, Australia, Association for Computing Machinery: 259–268

21 Kusner, Loftus, Russell og Silva (2017) Counterfactual Fairness. arXiv e-prints

22 Kamishima, Akaho, Asoh og Sakuma (2012) Fairness-Aware Classifier with Prejudice Remover Regularizer, Berlin, Heidelberg, Springer Berlin Heidelberg, Woodworth, Gunasekar, Ohannessian og Srebro (2017) Learning Non-Discriminatory Predictors Proceedings of Machine Learning Research 65: 1-34, Zafar, Valera, Rodriguez og Gummadi (2017). Fairness Beyond Disparate Treatment & Disparate Impact. World Wide Web Conference, Perth, Australia, Agarwal, Beygelzimer, Dudík, Langford og Wallach (2018). A Reductions Approach to Fair Classification. 35th International Conference on Machine Learning, Stockholm, Sweden

23 Hardt, Price og Srebro (2016) Equality of Opportunity in Supervised Learning. arXiv:1610.02413 [cs]

24 Dwork, Hardt, Pitassi, Reingold og Zemel (2011) Fairness Through Awareness. arXiv:1104.3913 [cs], Hardt, Price og Srebro (2016) "Equality of Opportunity in Supervised Learning arXiv:1610.02413 [cs].
25 Hardt, Price og Srebro (2016) Equality of Opportunity in Supervised Learning. arXiv:1610.02413 [cs]
26 Kleinberg, Mullainathan og Raghavan (2016) Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv e-prints, Chouldechova, (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments Big Data 5(2), Corbett-Davies, Pierson, Feller, Goel og Huq (2017) Algorithmic decision making and the cost of fairness. KDD '17, Berk, Heidari, Jabbari, Kearns og Roth (2018) Fairness in Criminal Justice Risk Assessments: The State of the Art." Sociological Methods & Research Online first
27 Northpointe (2012). Practitioners Guide to COMPAS
28 Angwin, Larson, Mattu og Kirchner (2016) Machine Bias ProPublica, Larson, Mattu, Kirchner og Angwin ibid. How We Analyzed the COMPAS Recidivism Algorithm
29 Dieterich, Mendoza og Brennan (2016) COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity, NorthPointe, Flores, Bechtel og Lowenkamp (2016) False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks Federal Probation 80(2). ProPublica har i sit svar fastholdt kritikken. Larson og Angwin (2016) Technical Response to Northpointe. ProPublica.
30 Friedler, Scheidegger, Venkatasubramanian, Choudhary, Hamilton, et al. (2019) A comparative study of fairness-enhancing interventions in machine learning. Proceedings of the Conference on Fairness, Accountability, and Transparency. Atlanta, GA, USA, Association for Computing Machinery: 329–338
31 Se især databeskyttelsesforordningens betragtning nr. 71. Se også Tischbirek (2020) Artificial Intelligence and Discrimination: Discriminating Against Discriminatory Systems Regulating Artificial Intelligence, Springer, Gellert, De Vries, De Hert og Gutwirth (2013) A comparative analysis of anti-discrimination and data protection legislations, Discrimination and privacy in the information society (pp. 61-89). Springer, Hacker (2018) Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law 55 Common Market Law Review, Issue 4, pp. 1143–1185 samt Guidelines on improving the collection and use of equality data, 2018 fra High Level Group on Non-discrimination, Equality and Diversity, EU-Kommissionen.

BILAG

BILAG 1

TERMINOLOGILISTE

Algoritme: En defineret serie af matematiske eller logiske operationer, som udføres på et datasæt for at løse en opgave.

Beslutningstræ: En type klassificeringsalgoritme, som sorterer borgere gennem en serie af beslutningspunkter, indtil hver borger når et slutpunkt i træet. Hvert beslutningspunkt tester værdien for en bestemt variabel, og sorterer borgerne afhængigt af, om værdien er over eller under beslutningspunktets tærskel.

Bias: Et udtryk for en algoritmisk models iboende tendens til at vurdere borgere på en bestemt måde. Når denne iboende tendens afviger fra, hvordan målegenskaben faktisk er fordelt blandt borgere, giver bias modellen tendens til at fejlvurdere borgerne. En høj bias giver derfor en for "grovkornet" sortering af borgerne med en høj fejlrate til følge.

Datasæt/Træningsdatasæt: Et datasæt, som anvendes med en læringsalgoritme til at udvikle en profileringsmodel, idet det udover de variable også indeholder en værdi for målegenskaben, som for borger fortæller læringsalgoritmen om den pågældende borger besidder eller ikke besidder egenskaben. Der vil oftest være tale om et sæt data bestående af borgersager og en lang række variable af relevant for modellens træning. Datasæt opdeles typisk i træningssæt, testsæt og valideringssæt.

Domænekendskab: Kendskab til f.eks. maskinlæringsmetoder, modeltyper, modelkvalitet og fejltyper, som kan være nødvendige for at omsætte data om modellen til forståelse af modellen.

Fejlrate: Andelen af forkerte klassifikationer i den samlede mængde af klassifikationer. Fejlraten er målestok for modellens evne til korrekt at vurdere, om borgere besidder eller ikke besidder målegenskaben, og bruges derfor ofte som den første og mest centrale målestok for en profileringsmodels kvalitet.

Fejltyper: Typer af fejlklassifikationer, som en model kan foretage. Mest simpelt skelner man mellem fejltyperne falsk positive og falsk negative. Ved et falsk positivt resultat vurderer modellen, at en borger besidder målegenskaben, selvom dette ikke er tilfældet. Ved et falsk negativt resultat vurderer modellen, at borgeren ikke besidder målegenskaben, selvom det faktisk er tilfældet.

Forklaring: Tekniske metoder (også benævnt xAI), der gør det muligt for mennesker at analysere en profileringsmodel og forstå, hvorfor modellen er nået frem til et givet resultat (eng. "explainability"). Forklaringer er relevante for modeller, der er for komplekse til, at de kan fortolkes

Fortolkning: Adgang for mennesker til at forstå, hvordan modellen virker, ved at studere selve modellen (eng. "interpretability"). Mange simple modeller har høj transparens – en enkel logistisk regression eller et kort beslutningstræ er relativt lette at overskue og forstå. Adgangen til at fortolke en model aftager typisk i takt med at dens kompleksitet vokser. Meget komplekse modeller, som dybe neurale netværk eller store tilfældige skove, kan sjældent fortolkes. De er uigennemsigtige sorte kasser (eng. "black boxes"), som der efter omstændighederne kan opnås transparens om ved en forklaring.

Generaliserbarhed: Et udtryk for modellens evne til at vurdere nye data på samme måde som data i træningssættet. En central opgave i udviklingen af en profileringsmodel er at opnå en høj generaliserbarhed og samtidig en lav fejlrate.

Global transparens: transparens om, hvordan modellen virker for alle borgere, som den anvendes på.

Holistisk transparens: transparens om, hvordan hele modellen virker.

Klassificeringsmodel: En model, som har til formål at vurdere, på hvilket niveau hvorvidt en person besidder en målegenskab eller ej.

Klyngeanalyse: En metode til at analysere data for at finde mønstre i data. Klyngeanalyse forudsætter ikke, at man på forhånd kan definere, hvilken egenskab i data man ønsker at modellen vurdere (såkaldt "unsupervised learning").

Kernetrick: Kernetricket (eng. "kernel trick") er en metode, som lader lineære modeller klassificere borgere i datasæt, hvor der er ikke lineære sammenhænge, og det derfor ikke umiddelbart er muligt lineært at afgrænse borgere med og uden målegenskaben. Kernetricket erstatter de almindelige variable i modellens funktion med variable, som måler hvor om og hvor meget en borger, som skal klassificeres, minder om allerede klassificerede borgere. Derved bliver modellen i stand til at trække ikke-lineære afgrænsninger. Kernetricket anvendes ofte i SVM-modeller.

Kunstig intelligens: En samlebetegnelse for teknologier, som lader software løse komplekse opgaver, især når den derved lader teknologi agere delvist selvstændigt, eller løser opgaver, som det ellers kun er mennesker, som har kunnet løse. Mange af de teknologier, som vi behandler i denne analyse, betegnes ofte som kunstig intelligens. Kunstig intelligens er imidlertid et bredt og flertydigt begreb, som også dækker over teknologier, som er meget forskellige fra dem, vi her behandler.

Lineær regressionsmodel: En model, hvor borgere fordeles på en skala efter målegenskabens værdi.

Logistisk regression: En type klassificeringsalgoritme, som vurderer hvordan hver af de variable påvirker sandsynligheden for, at en borger besidder målegenskaben, og sorterer borgerne henholdsvis positivt og negativt afhængigt af om denne sandsynlighed overskrider en tærskel.

Lokal transparens: transparens om, hvordan modellen virker for én bestemt borger eller en bestemt gruppe af borgere.

Læringsalgoritme: En algoritme, som ved at analysere et datasæt og evaluere mulige modellers kvalitet med en tabsfunktion, udvikler en anden algoritme, f.eks. en profileringsmodel.

Maskinlæring: en udviklingsproces, hvor en læringsalgoritme analyserer et datasæt, og tilpasser en anden algoritme, f.eks. en profileringsmodel, så den optimalt løser en opgave på datasættet.

Modulær transparens: transparens om, hvordan en bestemt del af modellen virker.

Målegenskab: Den egenskab, som en klassificeringsalgoritme forsøger at vurdere, om borgere besidder eller ikke besidder.

Naiv Bayes-model: En modeltype baseret på Bayesiansk sandsynlighedsteori. Læringsalgoritmen beregner hvordan hver variabel enkeltvis påvirker sandsynligheden for, at en borger besidder målegenskaben, og modellen klassificerer borgere ved at vurdere, om den samlede betingede sandsynlighed er størst for at have eller ikke have målegenskaben.

Objekter: Objekter i et datasæt er de ting, som data omhandler. I en profileringsmodel er det objekters egenskaber, som modellen vurderer. Objekter i et datasæt kan eksempelvis være personer, familier, virksomheder, eller ejendomme.

Overtilpasning: Når en model trænes til at passe så præcist på træningssættet, at den dårligt kan generaliseres til nye data. En overtilpasset model giver en meget præcis og "finkornet" sortering af borgere i træningssættet, men vil have en høj fejlrate, hvis den tages i anvendelse.

Polynomisk variabel: En kompleks repræsentation af en enkelt variabel eller en repræsentation af interaktion mellem flere variable. Polynomiske variable anvendes typisk til at modellere ikke-lineære sammenhænge og interaktioner i data.

Profileringsmodel: En algoritme, som på baggrund af statistiske data forsøger at vurdere, hvorvidt personer eller persongrupper har eller ikke har en målegenskab.

Regressionsmodel: En model, som har til formål at vurdere, på hvilket niveau en person besidder en målegenskab.

Regulariseringsfaktor: Et element, som kan tilføjes til en læringsalgoritmes tabsfunktion. Ofte anvendes regulariseringsfaktoren til at reducere overtilpasning, ved at definere en omkostning for at give vægt til de variable i modellen. Derved får læringsalgoritmen tendens til at bruge så få variable som muligt, med så lav vægt som muligt, og deraf til at træne en model som alene giver vægt til de variable, som har en stærk sammenhæng med målegenskaben.

Slutbetingelser: Anvendes af læringsalgoritmen for et prædiktivt beslutningstræ til at bestemme, hvornår den skal definere et slutpunkt i træet. Slutbetingelser kan eksempelvis være, at gruppen af personer har nået en vis størrelse, eller at hver person har været igennem et bestemt antal beslutningspunkter.

Splitbetingelser: Anvendes af læringsalgoritmen for et prædiktivt beslutningstræ til at bestemme, hvilken test den skal definere for hvert af beslutningspunkterne i træet. Typisk vil læringsalgoritmen analysere datasættet for at vælge den test, som bedst muligt deler borgerne i én gruppe med og én gruppe uden målegenskaben.

Support Vector Machine (SVM): En type klassificeringsalgoritme, der forsøger at vurdere, om en borger besidder eller ikke besidder målegenskaben. SVM er en lineær model, som tilskriver de variable en vægt, og klassificerer borgeren på baggrund af summen af de vægtede variable. SVM-modeller anvendes ofte med det såkaldte kernetrick.

Tabsfunktion: En central komponent i en læringsalgoritme, som definerer, hvordan læringsalgoritmen matematisk skal evaluere kvaliteten af en model. Læringsalgoritmen forsøger normalt at minimere værdien af denne funktion, det vil sige at specificere den model, hvor tabsfunktionen har den lavest mulige værdi. Almindelige tabsfunktioner er logistisk tabsfunktion (eng. "logistic loss"), kvadreret fejl (eng. "squared error"), og hængsel tabsfunktion (eng. "hinge loss").

Tilfældig skov: En variation af beslutningstræet, hvor en serie af træer konstrueres ved at hvert led tilfældigt begrænses til at vælge mellem et udsnit af de variable, og klassificeringen af en borger er en funktion af sættet af individuelle træers klassificeringer.

Tærskelværdi: Grænsen for, hvornår en person vurderes positivt eller negativt i forhold til målegenskaben.

Undertilpasning: Når en model ikke i tilstrækkelig detaljeret grad repræsenterer sammenhænge i træningsdatasættet. En undertilpasset model giver en uprækis

og "grovkornet" sortering af borgerne i træningsdatasættet, fordi der er væsentlige sammenhænge i data, som den ikke repræsenterer.

Variabel: En egenskab eller et kriterium, som indgår i datasættet med en værdi, f.eks. køn, alder, civilstatus eller indkomst. Hver af værdierne for de variable indgår i modellens vurdering af målegenskaben.

Varians: Varians er modellens følsomhed overfor variationer i træningsdata, når denne følsomhed fører til fejlklassificeringer. Høj varians kan få modellen til at repræsentere sammenhænge, som kun optræder tilfældigt i de historiske data, og som derfor ikke fører til korrekte klassificeringer af nye borgere. Varians er således en målestok for modellens evne til at foretage korrekte vurderinger af borgere, som ikke optræder i træningssættet. En profileringsmodel med høj varians vil typisk være overtilpasset data i træningssættet.

Vægt: En værdi for hver af de variable, som profileringsmodellen justerer den variable med for at vurdere målegenskaben. Træning af vægtene er ofte den centrale opgave for en læringsalgoritme, idet de angiver den rolle, som den enkelte variabel spiller i den samlede model.

BILAG 2

FIRE EKSEMPLER PÅ MYNDIGHEDERS BRUG AF PROFILERING-SMODELLER TIL AUTOMATISERET BESLUTNINGSSTØTTE

Alle cases bygger på aktindsigt og/eller interviews med nøglepersoner foretaget i 2019 samt presseomtale og offentligt tilgængelige oplysninger.

GLADSAXE KOMMUNE: TIDLIG OPSPORING AF SOCIALT UDSATTE BØRN

Gladsaxe kommune gik i midten af 2017 i gang med at udvikle en "dataunderstøttet opsporingsmodel" som led i kommunens projekt "Tidlig opsporing". Det erklærede formål med projektet var at forbedre kommunens evne til at identificere børn med høj risiko for udsathed og mistrivsel, og kommunen håbede, at modellen særligt ville gøre det muligt at identificere børnene på et tidligere tidspunkt i deres liv. Behovet for dette begrundedes dels med solid forskningsbaseret evidens på effekten af tidlige indsatser, og dels med kommunens egne analyser. Sidstnævnte viste, at kun godt en fjerdedel af de børn, som kommunen identificerer som udsatte eller i mistrivsel, opdages i førskolealderen, mens godt en tredjedel opdages efter barnet er fyldt 12 år.

Kommunen gik derfor i gang med at udvikle en model, som kunne forudsige det enkelte barns risiko for at være eller blive socialt udsat på baggrund af en række datapunkter om barnet selv og dets familie. Det første skridt var en analyse af en udvalgt gruppe af kommunens historiske sager med udsatte børn, som kombinerede statistisk analyse af disse sager med fagpersoners vurderinger af fællestræk ved sådanne sager. På baggrund af denne analyse udvalgte kommunen i alt 44 variable, som man vurderede kunne være relevante for modellen. Det drejede sig bl.a. om forældres beskæftigelsesstatus og -historik, statsborgerskab, faderskabsforhold, forældres bopæl, barnets brug af tandpleje, underretninger, pasningsforhold og sprog. Datagrundlaget for modellen ville bestå af træk fra ni registre, inklusiv KMD Momentum (beskæftigelse), CPR, tandplejejournaler, KMD institutionssystem (pladsanvisningen) og SBSYS (underretninger).

Gladsaxe kommune valgte at udvikle et prædiktivt beslutningstræ (se kapitel 3). Intentionen var at modellen regelmæssigt kunne analysere data for samtlige børnefamilier i kommunen, og vurdere den statistiske sandsynlighed for mistrivsel for hver enkelt familie. I de tilfælde, hvor den vurderede sandsynlighed oversteg en bestemt tærskel, ville modellen gøre kommunale sagsbehandlere opmærksom på dette. Disse sagsbehandlere kunne derpå beslutte at foretage en individuel, manuel vurdering af de relevante familier.

Første skridt ville være en indledende vurdering af, om der var grundlag for at forfølge sagen. Andet skridt ville være kontakt til familien for at indhente samtykke til at gennemgå og vurdere sagen i detaljer. Tredje skridt ville være sagsbehandlerens vurdering af, om det i det givne tilfælde ville være ønskeligt at iværksætte yderligere tiltag, f.eks. afklaring af behov og tilbud om støtte. Gladsaxe Kommune understreger i den forbindelse, at det var centralt i planen for modellen, at en familie, hvor barnet ikke var i mistrivsel, men alene havde forhøjet risiko for udsathed, skulle kunne takke nej til sådanne tilbud, og at et sådant afslag ikke ville have fået konsekvenser i form af f.eks. en social børnesag.

Kommunen ansøgte i slutningen af 2017 om tilladelse til at samkøre personfølsomme registerdata til brug for modellen under frikommuneforsøg-II-ordningen. Denne ansøgning blev afvist, med den begrundelse at regeringen ønskede at udvide muligheden for registersamkøring for alle kommuner som del af sin kommende strategi mod parallelsamfund, den såkaldte "ghettopakke". I begyndelsen af marts 2018 dækkede flere store medier historien om modellen med en udpræget kritisk vinkel. Arbejdet med udvikling af modellen fortsatte i resten af 2018, med henblik på at have modellen køreklar til det tidspunkt, hvor Gladsaxe Kommune kunne få tilladelse til at bruge den.

I slutningen af 2018 valgte kommunen imidlertid endegyldigt at sætte udviklingen af modellen i bero. Det skete, før kommunens udviklere lagde sig fast på betingelser for splitbetingelser og stopbetingelser, fastsatte tærsklen for sandsynligheder, eller testede og validerede modellen (se kapitel 3). Beslutningen skyldtes test, som viste, at modellens fejlrate ville blive for høj, først for fremmest fordi antallet af historiske sager var begrænset. Kommunen havde kun 117 sager med udsatte børn i alderen 0-6 år, som kunne fungere som positive eksempler i datasættet.

Inden den endelige beslutning blev truffet, overvejede kommunen to mulige løsninger på denne udfordring. Den første var at træne modellen på et større datasæt, som dels kunne inkludere data om større børn, og dels data om sager fra andre kommuner. Den anden løsning var at inkludere visse variable, som kommunen vurderede kunne have stærk statistisk signifikans, herunder især sundhedsdata fra den kommunale sundhedspleje og det kommunale rusmiddelcenter.

Selvom den konkrete model således er skrinlagt, illustrerer Gladsaxe Kommunes model til tidlig opsporing en type automatiseret beslutningsstøtte, som kunne anvendes både i indsatsen for udsatte børn og på andre dele af det sociale område.

UDBETALING DANMARK: KONTROL MED UDBETALING AF YDELSER

Udbetaling Danmark er en offentlig myndighed under ATP-koncernen, som varetager udbetalingen af bl.a. boligstøtte, folkepension, barselsydelser, børne- og ungeydelser, og førtidspension. I 2014 oprettede Udbetaling Danmark "Den Fælles Dataenhed", som bl.a. fik til opgave at udvikle datadrevne modeller, der kunne styrke kontrollen med, at ydelser udbetales korrekt. Det vil i praksis sige, at Udbetaling Danmarks Dataenhed samkører registre, og anvender profileringsmodeller på disse data til at identificere personer, som har en forhøjet risiko for uberettiget at have fået udbetalt ydelser. Udbetaling Danmark fik lovhjemmel til registersamkøringen i maj 2015, og begyndte derpå udviklingen af de første modeller.

Udbetaling Danmark udvikler og bruger modeller til kontrol på områder, hvor myndigheden har erfaret at der forekommer uberettigede udbetalinger. Det kan f.eks. være ydelser til enlige forsørgere, som reelt er samlevende, barselsydelser til borgere, som har haft fiktiv ansættelse i en virksomhed, eller borgere, som begynder en uddannelse og søger SU uden at få stoppet udbetaling af kontanthjælp. Dataenheden arbejder løbende med at udvælge og udvikle nye modeller baseret på erfaringer hos Udbetaling Danmark og kommunerne. I 2019 brugte Udbetaling Danmark ca. 60 forskellige modeller til løbende analyse af registersamkørte data.

Udbetaling Danmarks modeller varierer betragteligt i både kompleksitet og karakter. Langt de fleste og især de første modeller, som man udviklede, er relativt simple, manuelt konfigurerede "udsøgningsmønstre". Et sådant udsøgningsmønster er en form for meget enkelt beslutningstræ, der udfører en serie af tests på hver person, som fører til at personen klassificeres enten positivt eller negativt. For et manuelt konfigureret udsøgningsmønster er hver test defineret af udvikleren baseret på kvalitativ og kvantitativ analyse af fællestræk ved historiske sager. I de seneste år har man i stigende grad taget mere komplekse modeller baseret på maskinlæring i anvendelse, herunder naiv Bayes-klassifikation og klyngeanalyse (se kapitel 3).

Når modellen er udviklet og implementeret, foregår kontrollen ved, at modellen regelmæssigt – typisk en gang om ugen – analyserer registersamkørte data fra de relevante personer (f.eks. modtagere af den konkrete ydelse). Modellen genererer efter hver analyse en "undringsliste" med de personer, som statistisk set har høj risiko for uberettiget at modtage ydelsen. Sagerne optræder på undringslisten i anonymiseret form, da listen kun indeholder sagsnumre samt en overordnet beskrivelse af hvilke variable, modellen har behandlet.

Næste skridt tages af sagsbehandlere i kontrolafdelinger ved kommunerne og Udbetaling Danmark. Disse kan ved at logge på Udbetaling Danmarks system få adgang til de dele af listen, som vedrører de borgere de er ansvarlige for, f.eks. beboere i den pågældende kommune. Sagsbehandlerne kan vælge at trække sager fra undringslisten ud til kontrol, hvorved de får adgang til den fulde sag i

ikke-anonymiseret form. Sagsbehandleren er i den forbindelse forpligtet til at oplyse borgeren om, at deres sag er blevet udtrukket til kontrol. Sagsbehandleren foretager derpå en manuel og individuel vurdering af, om der er grund til at gå videre med sagen. Hvis det ikke er tilfældet, lukkes sagen igen, og borgeren gøres opmærksom på at sagen er lukket. I modsat fald forfølges sagen, typisk ved at sagsbehandleren som første skridt beder borgeren om at indsende supplerende oplysninger.

Udbetaling Danmark har lovhjemmel til at samkøre data fra de offentlige myndigheders registre, men adgang til de enkelte registre sker efter forhandling med den ansvarlige myndighed, som vurderer muligheden for at give adgang til de pågældende data til brug for den specifikke model. I den første periode arbejdede Udbetaling Danmark med registersamkøring fra relativt få datakilder. Det drejede sig først og fremmest om data fra de pågældende ydelsers egne systemer, som blev samkørt med data fra CPR-, BBR- og CVR-registrene, hvor Udbetaling Danmark trækker data om blandt andet bopæl og flytninger, civilstand og familieforhold, boligforhold, og virksomhedsejerskab og -oprettelse. I dag trækkes data fortsat først og fremmest fra de pågældende ydelsers egne systemer, men også fra CPR, BBR, CVR, SE (virksomheder), og DAR (adresser). Derudover trækkes mere specifikke data fra E-indkomst (indkomstforhold), R75 (skatteoplysninger), Regionsdata (sundhed), Moms, og STAR-data (kontanthjælp og sygedagpenge). Samtidig arbejder modellerne i stigende grad med data om relevante personer i den pågældende borgers netværk, fordi der i nogle tilfælde kan etableres signifikante sammenhænge mellem data om disse personer og en forøget risiko for uberettiget udbetaling. Det kan eksempelvis være en borger som er ansat i en virksomhed, hvis ejer er bosat i udlandet.

Som led i samarbejdet mellem dataenheden og kontrolenheder modtager Udbetaling Danmark løbende tilbagemeldinger på sagsbehandlingen af de udvalgte sager, og benytter resultaterne til at opdatere deres modeller, især med henblik på at reducere antallet af falsk positive, dvs. sager som af modellen vurderes til at have høj risiko, men hvor der ikke kan konstateres uberettiget udbetaling.

Udbetaling Danmark arbejder med ydelser, som udbetales til mange borgere over hele landet. Det betyder at deres datapulje i udgangspunktet er ganske omfattende. Ikke desto mindre kan man opleve udfordringer med at datasættet er så lille, at det er vanskeligt at udvikle en model som har lav fejlrate. Det skyldes at der ofte kun findes et relativt lille antal sager af en given type, hvor der er konstateret uberettiget udbetaling af ydelser. Udviklingen af en model med lav fejlrate er afhængig af en vis mængde af sådanne sager, for at kunne identificere statistisk signifikante forskelle mellem medlemmerne af dette sæt og andre borgere.

Udbetaling Danmark fremhæver selv, at brugen af modeller på registersamkørte data som automatiseret beslutningsstøtte har flere fordele. For det første fører

det til mere ensartet sagsbehandling og effektiviserer sagsbehandlingen ved at lade sagsbehandlere fokusere på de sager som har høj risiko. For det andet gør modellerne det muligt at opdage sager om uberettiget udbetaling, som ellers ville være meget vanskelige at spore, og at opdage sagerne tidligere, især når der er tale om fejl. Endelig reducerer det antallet af falsk positive, dvs. sager hvor borgeren forstyrres af en ubegrundet henvendelse.

STYRELSEN FOR ARBEJDSMARKED OG REKRUTTERING: PROFILAFKLARING FOR NYLEDIGE

Styrelsen for Arbejdsmarked og Rekruttering (STAR) er en offentlig myndighed med ansvar for bl.a. at implementere reformer på ledigheds- og beskæftigelsesområdet. Siden 2013 har STAR arbejdet med at udvikle modeller til at vurdere nylediges risiko for at blive langtidsledige. Det såkaldte profilafklaringsværktøj findes i to varianter rettet mod henholdsvis gruppen af ledige unge under 30 på uddannelseshjælp, og gruppen af dagpengemodtagere generelt. Værktøjerne er designet til at identificere de unge, som har høj sandsynlighed for af egen drift at gennemføre en uddannelse, og de personer i begge grupper, som har høj sandsynlighed for at blive langtidsledige. Udviklingen af værktøjerne blev igangsat i forbindelse med reformer af kontanthjælps- og beskæftigelsessystemerne i 2013 og 2014. De har været testet siden 2014 og 2015, og er opdateret på baggrund af erfaringer og nye data i henholdsvis 2016 og 2017.

STAR iværksatte som første trin i udviklingen af værktøjerne en kortlægning af dels forskning og faglig indsigt i hvilke forhold, som påvirkede de to målegenskaber (langtidsledighed og gennemførelse af uddannelse), og dels internationale erfaringer med udviklingen af profileringsmodeller. Denne kortlægning informerede valget af relevante variable til datasættet, valget af modeltype, udviklingsprocessen og brugerinddragelse, samt udformningen af et spørgeskema, som bruges til yderligere at afdække den lediges baggrund og motivation.

De variable i træningssættet for unge nyledige på uddannelseshjælp inkluderede køn, alder, bopæl, herkomst, forsørgerpligt, uddannelse, helbred, modtagelse af overførselsindkomster, civilstand, tidligere indkomst og tidligere branche. Tilsvarende inkluderede de variable i træningssættet for nyledige dagpengemodtagere køn, alder, bopæl, herkomst, forsørgerpligt, uddannelse, og beskæftigelses- og ledighedshistorik.

Spørgsmålene i spørgeskemaet til unge nyledige behandler bl.a. den lediges lyst til at påbegynde uddannelse, erfaringer med uddannelse, forventninger til gennemførsel, og oplevede barrierer for at gennemføre en uddannelse. Spørgsmålene i spørgeskemaet til dagpengemodtagere behandler bl.a. den lediges forventninger til ledighedsperiode, jobsøgningshistorik, villighed til at rejse langt, skifte branche, eller gå ned i løn, vurdering af jobmulighederne på relevante arbejdsområder, og oplevelse af barrierer for adgang til job.

På baggrund af datasæt med de udvalgte variable udviklede STAR et beslutningstræ for hver af de to grupper (se kapitel 3). I udviklingen blev en række forskellige modeller trænet, med lidt forskellige stop- og slutbetingelser. Disse modeller blev testet på et valideringssæt, og evalueret på baggrund af dels fejlrate og fejltyper, og dels modellens brug af variable og tests, som giver mening i relation til det danske arbejdsmarked.

Det var i udviklingen af den første generation af profileringsmodeller en erklæret intention, at de skulle opdateres løbende, efterhånden som STAR fik adgang til nye data, der kunne sammenholdes med modellernes forudsigelser. Den opdaterede version af værktøjet for unge nyledige blev taget i brug i april 2016. Den er udviklet på baggrund af analyse af data fra ca. 85.000 forløb i perioden 2013-2015 for unge uden en uddannelse på uddannelseshjælp. Den opdaterede version af værktøjet for nyledige dagpengemodtagere blev taget i anvendelse i september 2017. Den reviderede model for nyledige dagpengemodtagere er udviklet på baggrund af analyse af data fra ca. 112.000 dagpengeforløb i perioden 2014-2016, hvor den ledige har udfyldt det daværende spørgeskema. Et centralt element i den opdaterede model for ledige dagpengemodtagere er, at modellen i modsætning til de andre modeller behandler svarene på spørgeskemaet som variable. Det vil sige at den lediges svar på spørgsmål i spørgeskemaet har betydning for hvordan vedkommende vurderes.

De to modeller arbejder med tre forskellige målegenskaber, som modellen vurderer at personen enten besidder eller ikke besidder.

Den første målegenskab er langtidsledighed for gruppen af nyledige på dagpenge, defineret som ledighed i mindst 26 uger, som ikke på noget tidspunkt afbrydes i mindst 7 uger. Modellen identificerer 9 grupper med forskellige kombinationer af karakteristika, der i træningsdatasættet tilsammen udgjorde godt 20% af de nyledige dagpengemodtagere, som har særligt høj sandsynlighed for at blive langtidsledige. I træningsdatasættet blev 67% af borgerne i disse grupper langtidsledige mod 39% af de andre nyledige dagpengemodtagere. Grupperne adskiller sig især fra andre på variablene alder, forventninger til ledighedsperiode, oplevede barrierer, indkomst- og forsørgelseshistorik, og ikke-vestlig herkomst. Data trækkes fra spørgeskemaet, samt fra CPR-, RAM-, og e-indkomst-registre.

Den anden måleegenskab er langtidsledighed for unge ledige på uddannelseshjælp, defineret som ledighed i mindst 52 uger, som ikke på noget tidspunkt afbrydes i mindst 7 uger. Modellen identificerer 15 grupper med forskellige kombinationer af karakteristika, der i træningsdatasættet tilsammen udgjorde godt 10% af de nyledige unge, som har særligt høj sandsynlighed for at blive langtidsledige. I træningsdatasættet blev 62% af de nyledige i disse grupper langtidsledige mod 25% af de andre nyledige unge. Grupperne adskiller sig især fra andre på variablene alder, køn, beskæftigelseshistorik, uddannelse, indkomst-

og forsørgelseshistorik, og ikke-vestlig herkomst. Data trækkes fra spørgeskemaet, samt fra CPR-, KMD-, RAM-, og e-indkomst-registre.

Den tredje måleegenskab er unge ledige på uddannelseshjælp, der hurtigt i job eller uddannelse, defineret som ledighed i højst 6 måneder før job eller uddannelse påbegyndes, og derpå fastholdes indtil mindst et år efter ledighedens begyndelse. Modellen identificerer 7 grupper med forskellige kombinationer af karakteristika, der i træningsdatasættet tilsammen udgjorde godt 10% af de nyledige som søger om uddannelseshjælp. I de historiske data kommer 57% af de nyledige i denne gruppe hurtigt i job eller uddannelse mod 26% af de andre nyledige søgere af uddannelseshjælp. Grupperne adskiller sig især fra andre på variablene beskæftigelseshistorik, uddannelse og alder.

Beslutningstræer er kendetegnet ved høj algoritmisk transparens (se kapitel 7). Det er typisk lettere at formidle, at en person tilhører en gruppe defineret ved et sæt af karakteristika, som i gennemsnit har høj sandsynlighed for langtidsledighed, end at formidle for eksempel funktionen og vægtene i en logistisk regressionsanalyse. STAR anfører i den forbindelse, at valget af beslutningstræ som algoritme blandt andet var motiveret af ønsket om at anvende en model, hvis struktur og vurdering relativt enkelt kunne formidles til sagsbehandlere og berørte borgere.

For sagsbehandlere på f.eks. jobcentre giver værktøjet automatiseret beslutningsstøtte i form af en vurdering, som kan indgå i sagsbehandlerens beslutning om, hvilke tilbud den ledige skal gives, og hvilke initiativer der eventuelt skal iværksættes. For den ledige giver værktøjet en begrundet tilbagemelding på hvordan vedkommende er vurderet. Brugen af værktøjerne er for både ledige og sagsbehandlere frivillig: hvis den ledige vælger ikke at udfylde spørgeskemaet, får sagsbehandlere ikke adgang til en vurdering fra værktøjet.

HORSENS KOMMUNE: PRIORITERING AF KONTROLSAGER OM SOCIAL SVINDEL

Horsens kommune begyndte i starten af 2019 at udvikle en model, som skulle hjælpe kommunens sagsbehandlere med at prioritere deres behandling af kommunens kontrolsager. Udviklingen skete i samarbejde med to speciale-studerende fra Århus Universitet, og modellen var færdigudviklet i sommeren 2019. Efterfølgende har kommunen besluttet ikke at sætte modellen i drift, da den vurderer at datasættet er for lille og modellens fejlrate derfor for høj. Projektet var ifølge kommunen først og fremmest et pilotprojekt, som skulle illustrere mulighederne ved at arbejde med data.

Modellen var udviklet til at analysere de sager om mulig social svindel, som kommunens kontrolafdeling rejser, f.eks. på baggrund af en henvendelse fra en myndighed eller en borger – så at sige en kommunal parallel til de modelanalyser som Udbetaling Danmark udfører (se case herom). Modellen ville for hver af disse sager vurdere, hvad sandsynligheden var, for at den pågældende sag faktisk drejede sig om social svindel. For hver sag ville modellen angive om sagen havde høj eller

lav sandsynlighed for at angå social svindel. På baggrund af denne vurdering ville sagsbehandlere i kommunens kontrolafdeling f.eks. kunne vælge at afsætte mere eller mindre tid til at vurdere sagen.

Processen ville konkret bestå i, at sagen åbnedes, for eksempel gennem en borgerhenvendelse, hvorpå en sagsbehandler ville vurdere om den skulle behandles. Kun hvis det var tilfældet, ville modellen vurdere sagen. Derpå ville sagen, sammen med modellens vurdering, gå videre til de sagsbehandlere, som skulle behandle sagen, og som på baggrund af modellens vurdering kunne prioritere sagen højere eller lavere.

Modellen blev udviklet på et sæt af ca. 1200 historiske sager, som er behandlet i kommunens kontrolafdeling fra efteråret 2016 og frem. Disse sager angik forskellige typer ydelser. Blandt alle disse sager er kun 75 afgjort som social svindel. Modellen analyserede på ca. 35 variable fra de afgjorte sager, herunder bopæl, privat indtægt, offentlige ydelser, civilstatus og -historik, adressehistorik, antal samlevere, køn, og alder. Udvalget af disse variable var baseret på den kommunale kontrolgruppes erfaringer, men også et resultat af, at udviklingen var underlagt visse begrænsninger i form dels af ressourcer til at forberede data til modellen, og dels adgang til registre udenfor kommunal regi, for eksempel SKAT.

Fordi der var så få positive sager – 75 ud af ca. 1200 – ville en udvikling på det reelle sæt sager have ført til en model, som tenderede mod at klassificere alle sager negativt. Selv hvis modellen kategoriserede alle sager i datasættet negativt og altså ignorerede de 75 positive sager, ville den stadig have en succesrate på 93,75%. De relativt få falsk negative resultater, som en sådan model ville resultere i, kunne hurtigt blive overgået af falsk positive resultater, hvis modellen forsøgte at klassificere nogle sager positivt. Kommunen var opmærksom på dette, men vurderede at det var vigtigere at undgå falsk positive vurderinger, hvor modellen kategoriserede de sager, som reelt angik social svindel, som lavrisikosager. Det skyldes at intentionen med modellen var, at alle sager som klassificeredes som høj risiko, skulle vurderes grundigt af en sagsbehandler, som kunne have sorteret de falsk positive vurderinger fra. Omvendt ville det store flertal af sager, som ville være blevet klassificeret som lav risiko, have modtaget en mindre omfattende sagsbehandling, som derfor ikke nødvendigvis ville have fanget en lille gruppe af falsk negative vurderinger. Udviklerne besluttede derfor at konstruere et sæt af træningsdata, som overrepræsenterede de historiske sager, hvor der var forekommet svindel, således at datasættet bestod af lige store mængder sager med og uden social svindel.

Horsens Kommune trænede og testede i alt ni forskellige modeller. De blev først og fremmest evalueret på fejlraten, men for at tilgodese hensynet til især at undgå falsk negative vurderinger blev modellerne også evalueret på deres præcision og genkald (se terminologiliste). Blandt de ni modeller valgte kommunen at færdigudvikle to forskellige modeller. Den ene model var en Support Vector Machine (SVM) (se kapitel 3). Denne model var bedst til at klassificere sagerne

i testsættet målt på fejlrate. Imidlertid har SVM-modeller lav algoritmisk transparens. Af hensyn til behovet for transparens udvikledes også en logistisk regressionsmodel (se kapitel 3). Denne model havde højere fejlrate, men bedre transparens, fordi modellens vægtning af de enkelte variable kan aflæses i koden.

Kommunen pseudonymiserede de behandlede data i udviklingsprocessen, så f.eks. variablen bopæl alene rummede postnummer, ikke den præcise adresse, mens variablen indtægt placerede hver person i en bred kategori snarere end med en præcis indtægt. De pseudonymiserede data kunne derfor ikke umiddelbart henføres til konkrete personer.

Anvendelse af modellen ville ifølge Horsens Kommune ikke have krævet særskilt hjemmel, idet der alene var tale om behandling af data som allerede indgår i den relevante sagsbehandling, på det tidspunkt hvor modellen ville have tilgået disse data.

Den væsentligste udfordring for udvikling af modellen var det lave antal af historiske sager, som er afgjort som omhandlende social svindel. Hvis modellen var blevet sat i drift, ville resultaterne af sagsbehandlingen kunne blive ført tilbage til udvikling af modellen, sådan at datasættet og dermed modellens forudsigelseskraft kunne være blevet forbedret over tid. Det begrænsede datasæt medførte, at modellens fejlrate umiddelbart var så høj, at Horsens Kommune besluttede ikke at sætte den i drift. En anden mulighed, som kommunen overvejede, ville være at udvikle en version af modellen i samarbejde med andre kommuner, hvorved datasættet kunne udvides med sager fra disse kommuner. En sådan udvikling er dog ikke aktuelt igangsat.

INSTITUT FOR
MENNESKE
RETTIGHEDER