

CompLex



Senter for rettsinformatikk / Avdeling for forvaltningsinformatikk

Åse Haram

Forklarbarhet i algoritmisk velferdsforvaltning

Demokrati og rettssikkerhet ved bruk av maskinlæringsmodeller i saksbehandlingen

6/2021



UiO **Det juridiske fakultet**

Henvendelser om denne bok kan gjøres til:

Senter for rettsinformatikk

Postboks 6706 St. Olavs plass

0130 Oslo

<http://www.jus.uio.no/ifp/om/organisasjon/seri/>

ISSN 2703-8777 Complex (online)

Grafisk produksjon: 07 Media AS – 07.no

Forord

Denne oppgaven er skrevet for graden Master i rettsvitenskap våren 2021 ved Institutt for offentlig rett. Det er flere jeg ønsker å takke når jeg nå har kommet i havn.

Jeg vil gjerne takke Torill Christine Edvardsen og Daniel Engehagen i NAV for forslag til oppgavetema og støtte på veien.

Takk til SERI ved Dag Wiese Schartum og Lee Andrew Bygrave for studentstipend og bistand til publisering, og til stipendiat Mona Naomi Lintvedt for gjennomlesning og konstruktive forslag.

Takk til veileder Jon Christian Fløysvik Nordrum som hevet listen og ga gode innspill til forbedringer på struktur og språk.

En stor takk går også til Robindra Prabhu i NAV for smittende faglig engasjement, tålmodige forklaringer og interessante diskusjoner.

Åse Haram, Oslo, november 2021

Sammendrag

Maskinlæringsmodeller kan benyttes til å påvise sammenhenger i store data-sett, til å profilere, klassifisere og predikere. I modelleringen benyttes avanserte algoritmer og store mengder data, og modellene kan derfor være lite forklarbare, det vil si lite forståelige, forutberegnelige og etterrettelige.

Dersom maskinlæringsmodeller skal anvendes i forvaltningens saksbehandlingssystemer, til å automatisere vurderinger eller som beslutningsstøtte til saksbehandlere, må det være mulig å overholde offentlighetsprinsippet og borgernes innsynsrett. I enkeltsaker må dessuten partens rett til kontradiksjon, personvern, privatliv og ikke-diskriminering ivaretas. Modellenes og systemenes forklarbarhet er derfor en betingelse for et velfungerende demokrati og retts-sikre forvaltningsavgjørelser.

I denne oppgaven presenterer jeg forklaringsbegreper og forklaringsmetoder fra den tekniske litteraturen om maskinlæring, og drøfter hvordan disse kan tilpasses forvaltningskonteksten. Gjennom analyse av flere eksempler på algoritmiske beslutningssystemer ser jeg på hvordan disse begrepene og metodene kan bidra til overholdelse av de rettslige krav til forklaring som følger av Grunnloven, forvaltningsretten, EMK og GDPR.

Jeg kommer blant annet frem til at det er et behov for regler for dokumentasjon av algoritmiske beslutningssystemer, og løsninger for å publisere eller gi innsyn som er meningsfulle for borgerne. Dette er nødvendig for å realisere offentlighetsprinsippet og den aktive informasjonsplikten. I noen tilfeller kan det være behov for å legge til rette for medvirkning fra interessenter eller berørte, for å sikre at modellen bygger på et saklig og representativt utvalg av data og at egen-skapen den predikerer for (utgangsvARIABLEN) er relevant.

Når det kommer til enkeltavgjørelser rettet mot en part vil disse måtte forklares både lokalt (på enkeltprediksjonsnivå) og globalt (på overordnet modellnivå), og det bør fremgå av en form for systembeskrivelse hvordan systemet er innrettet og utøver myndighet. En gjennomgang av noen utvalgte lokale forklarings-metoder for komplekse modeller, herunder approksimasjoner og kontrafaktiske forklaringer, tyder på at slike metoder ikke tilfredsstiller forvaltningslovens krav til begrunnelse. Det tilsier at enklere og mer forklarbare modeller bør velges i automatiserte avgjørelser. Det kan være mindre problematisk å ta i bruk

komplekse modeller som beslutningsstøtte, under forutsetning av at saksbehandlerne har tilstrekkelig kompetanse til å forstå modellen og anvende prediksjonene på en forsvarlig måte.

Kravet til forklarbarhet skjerpes for systemer som har mye personopplysninger, personopplysninger av sensitiv karakter (eller proxier for sensitive egenskaper) i datagrunnlaget, fordi dette fører med seg en risiko for diskriminering og usaklig forskjellsbehandling. Da må forvaltningsorganet sørge for dokumentasjon på at risikoen er håndtert gjennom tiltak som rettferdighetsteknikker, sosio-teknisk analyse eller algoritmisk konsekvensanalyse.

Jeg konkluderer også med at algoritmisk forvaltning, og kontrollen med denne, i større grad vil kreve systemperspektiv enn det har vært tradisjon for i forvaltningsretten. Selv om både GDPR og den kommende AI-lovgivningen fra EU inneholder systemkrav, må også forvaltningsretten tilpasse seg teknologien. Man bør vurdere egne saksbehandlingsregler for algoritmisk forvaltning, noe som blant annet kan omfatte et endret begrunnelseskrav.

Innhold

Forord	3
Sammendrag	4
1 Innledning	9
1.1 Tema og problemstilling	9
1.2 Aktualitet	9
1.2.1 Rettsstatlige bekymringer	10
1.2.2 Tilnærminger til forklarbarhet	11
1.2.3 Lovgivning og offentlige dokumenter	12
1.3 Metode	13
1.4 Oversikt over fremstillingen	15
2 Bruk av maskinlæring i velferdsforvaltning	17
2.1 Innledning	17
2.2 Maskinlæring som kunnskaps- og automatiseringsverktøy	17
2.3 Læringsmetoder	20
2.3.1 Forholdet mellom maskinlæring og profilering	21
2.4 Utvikling av en maskinlæringsmodell ved hjelp av veiledet læring	21
2.4.1 Systemdokumentasjon	24
2.5 Potensielle bruksområder med eksempler	24
2.6 Algoritmiske beslutningssystemer	27
2.7 Nærmere om aktuelle bruksområder for velferdsforvaltningen	29
2.7.1 Automatisert skjønnsutøvelse?	30
2.7.2 Maskinlæringsmodellen bidrar til sakens utredning	31
2.7.3 Maskinlæringsmodellen benyttes til kontrollformål	31
3 Forklarbarhet fra et teknisk perspektiv	33
3.1 Innledning	33
3.2 Hva er det som er vanskelig å forklare?	33
3.2.1 Svarte bokser	33
3.2.2 Dimensjonalitet og rasjonalitet	34
3.2.3 Forklarbarhet som funksjon av dimensjon og algoritmetype	36
3.3 Hvorfor er forklarbarhet nødvendig?	41
3.3.1 Brukeren må forstå hvordan modellen virker	41

3.3.2	Modellene må være etterprøvbare for å sikre kontradiksjon og kontroll	41
3.3.3	Det må være mulig å finne ut hva som gikk galt	42
3.3.4	Maskinlæringsmodeller har et problem med slagside	43
3.4	Hvordan kan modellene forklares? Begreper og tilnærminger.....	45
3.4.1	Globale forklaringer	46
3.4.2	Lokale forklaringer.....	48
3.4.3	Approximasjoner for å forklare komplekse modeller.....	52
3.4.4	Forklaring som formidling og interaksjonsdesign	54
3.4.5	Eksempel fra NAV: Forklaring av prediksjoner på sykefraværets lengde	56
3.4.6	Forholdet mellom transparens og forklarbarhet	59
3.5	Oppsummering: Hvordan skal forvaltningen forklare algoritmiske avgjørelser?	60
4	Forklarbarhet som forutsetning for et velfungerende demokrati.	63
4.1	Innledning	63
4.2	Behovet for en åpen og forståelig forvaltning.....	63
4.2.1	Ytrings- og informasjonsfrihet.....	63
4.2.2	Offentlighetsprinsippet og den passive informasjonsplikten	64
4.2.3	Aktiv informasjonsplikt	65
4.2.4	Deltakerdemokrati og brukermedvirkning	66
4.2.5	Krav til klart språk	67
4.3	Dokumentasjon og innsyn i digital forvaltning.....	68
4.3.1	Dokumentbegrepet og arkivplikten	68
4.3.2	Dokumentasjon av automatiserte saksbehandlingssystemer	70
4.3.3	Formidling og forståelighet.....	72
4.3.4	Oppsummering.....	72
4.4	Hvordan kan innsyn og medvirkning sikres i algoritmiske beslutningssystemer?	73
4.4.1	Modellene og beslutningssystemene må dokumenteres	74
4.4.2	Åpenhet krever aktiv formidling.....	76
4.4.3	Medvirkning for mer representative modeller.....	80
4.4.4	Oppsummering.....	82
5	Forklarbarhet som forutsetning for en rettssikker velferdsforvaltning. .	84
5.1	Innledning	84
5.2	Grunnleggende rettslige krav om ivaretagelse av privatliv og personvern ved algoritrisk forvaltning.....	85

5.2.1	GDPR og kravene til åpenhet, innsyn og informasjon til den registrerte	85
5.2.2	Gir GDPR en rett til forklaring på avgjørelser basert på maskinlæring?	88
5.2.3	SyRI-saken: Et algoritmisk beslutningssystem i strid med retten til privatliv og personvern.....	91
5.2.4	Oppsummering.....	98
5.3	Retten til kontradiksjon ved algoritmiske avgjørelser	99
5.3.1	Algoritmisk karakterfastsettelse ved IB og Datatilsynets intervensjon	99
5.4	Betydningen av lokale forklaringsmetoder	103
5.4.1	Lokale approksimasjoner og vedtaksbegrunnelser.....	103
5.4.2	Kontrafaktiske forklaringer og vedtaksbegrunnelser.....	106
5.4.3	Lokale forklaringer i forbindelse med beslutningsstøtte ...	106
5.4.4	Menneskelige versus algoritmiske begrunnelser.....	109
5.4.5	Oppsummering.....	111
5.5	Forklarbarhet og algoritmisk forskjellsbehandling.....	112
5.5.1	Avdekking av slagside med rettferdighetsverktøy.....	114
5.5.2	Avdekking av slagside ved hjelp av sosio-teknisk analyse: AMS-algoritmen for oppfølging av arbeidssøkende.....	115
5.5.3	Oppsummering.....	118
6	Avslutning	119
6.1	Oppsummerende punkter.....	119
6.2	Systemkrav og kontroll ved algoritmisk forvaltning.....	120
6.3	Regler for algoritmisk saksbehandling	122
7	Henvisninger	124
7.1	Litteratur.....	124
7.2	Lover	131
7.3	Internasjonale rettskilder.....	132
7.4	Forarbeider og offentlige dokumenter.....	132
7.5	Rettsavgjørelser.....	134
7.6	Oversikt over samtaler	135
7.7	Oversikt over konferanser, kurs og seminarer.....	135

1 Innledning

1.1 Tema og problemstilling

Tema for denne oppgaven er forklarbarhet av maskinlæringsmodeller brukt i offentlig forvaltning. «Forklarbarhet» sikter til måter å belyse hvordan maskinlæringsmodellene gjør sine beregninger og resultatene de produserer. Det kan også sikte til et krav om at de skal være mulige å forstå.

Min innfallsvinkel til temaet er forvaltningsrettslig. Jeg ser på hva forklarbarhet kan innebære dersom maskinlæringsmodeller integreres i saksbehandlingen. Den overordnede problemstillingen er: Hva vil det som ofte omtales som et krav til forklarbarhet innebære i en forvaltningsrettslig kontekst, og hvordan kan kravet realiseres hvis man skal bruke maskinlæringsmodeller til å forberede grunnlaget for eller fatte avgjørelser i velferdsforvaltningen?

1.2 Aktualitet

Maskinlæring er betegnelsen på ulike teknikker basert på data science og statistikk, som er rettet mot å identifisere mønstre og sammenhenger i store datasett. Maskinlæringsmodeller kan bidra til forvaltningens systemer og avgjørelser på ulike måter. Offentlig forvaltning, ikke minst velferdsforvaltningen, sitter på store mengder data om innbyggerne som kan brukes til å modellere sammenhenger og predikere utvikling. Teknologien har et potensiale for å fremme likebehandling, tjenester som er i forkant og individuelt tilpassede løsninger. Den kan styrke det faglige grunnlaget for avgjørelser og bidra til større gjennomsliktighet i prosessene.

Utviklingen vi ser i dag blir beskrevet som datadrevet. Datatilfanget øker og er mer detaljert på personnivå, datamaskinene får større regne- og lagringskapasitet og det skjer mye utvikling innenfor analyseteknikker. Maskinlæring gir en helt ny evne til å analysere disse store og komplekse datamengdene.¹

Maskinlæringsteknikker ble først utviklet i en 'low-cost, low-risk setting'. Basert på data om forbruksmønstre kunne man lage kunde profiler og spesialtilpasse nettreklame og anbefalinger for dermed å øke kundetilfredshet og salg. Men

1 Lovett (4.3.2020).

teknologien har bredt om seg, og samme type metoder brukes nå til å ta avgjørelser med konsekvenser for menneskers helse, økonomi og muligheter innenfor velferd, utdanning og arbeidsmarkedet. Det aktualiserer et behov for åpenhet og ansvarlighet.²

Maskinlæringsmodeller er komplekse og kan endre seg autonomt basert på dataene de tilføres. De kan være uforutberegnelige, lite etterprøvbare og vanskelige å forstå. Uten tilstrekkelig åpenhet og kontroll kan algoritmisk forvaltning også lede til svekket rettssikkerhet, innsyn og tillit.

1.2.1 Rettsstatlige bekymringer

Bruken av maskinlæring i forvaltning reiser flere nye problemstillinger for demokratiske rettsstater.

FNs spesialutsending for ekstrem fattigdom har advart mot utviklingen som finner sted innen velferdsforvaltninger rundt i verden. Han ser bruken av kunstig intelligens-systemer (som også omfatter maskinlæring) som et element i den generelle utviklingen mot digitaliserte og automatiserte velferdsstater rundt om i verden. Han mener teknologien ofte ikke brukes for det gode, til å utvikle bedre tjenester, styrke velferden til befolkningen eller fremme åpenhet og tillit i relasjonen mellom stat og borger. I stedet fremtrer den digitale velferdsstaten som et Foucaults panoptikon³. Den overvåker, kobler sammen data for å profilere innbyggerne, den mistenkeliggjør og straffer. De som er mest utsatt er de som har lavest inntekt, dårligst helse og lavest utdanning – altså den velferdsstaten først og fremst skal være til for. Ressursene som settes inn på systemer for å avdekke trygdemisbruk og skattesvindler står ikke i forhold til omfanget av problemet. I mange land er det private aktører som utvikler og drifter systemene, og de opererer mer eller mindre frikoblet fra menneskerettighetene og rettsstatlige verdier.⁴

Algorithmwatch, en NGO som overvåker og rapporterer om utviklingen av automatiserte beslutningssystemer i europeiske land, er bekymret for at systemene møter for lite kritikk og motstand. De går under radaren hos store deler av befolkningen. Det kan være lite kjent at systemene er der, og vanskelig å

2 Burrell (2016) s. 2.

3 *Foucaults panoptikon* henviser til Michel Foucaults historiske analyse av straff og makttøvelse. Fængselsarkitekturen som Jeremy Bentham foreslo på 1700-tallet, med et vaktårn i midten og cellene rundt, var tuftet på antagelsen om at fangene aldri kunne vite når de ble observert og ville dermed tilpasse sin adferd deretter. Foucault brukte dette til å vise hvordan makt internaliseres. Makten virker ved at den enkelte disiplinerer seg selv, og det kan gjelde enten overvåkning skjer fysisk eller digitalt.

4 FNs spesialrapportør på ekstrem fattigdom (2019) s. 4–5, 10, 13, 23.

overskue deres mulige konsekvenser. Organisasjonen advarer mot at manglende åpenhet er en trussel mot demokratiet.⁵

Det er mange eksempler internasjonalt på at implementeringen av kunstig intelligens-systemer i velferdsforvaltningen ikke har vært underlagt tilstrekkelig demokratisk og rettslig kontroll. Flere systemer har blitt trukket tilbake på grunn av offentlig kritikk eller fordi de har falt igjennom i domstolsprøving. Jeg vil omtale noen av dem i denne oppgaven. I alle tilfellene kan problemet føres tilbake til manglende transparens eller forklarbarhet, og til at man trolig ikke har innsett at algoritmene utøver myndighet. Personene det gjelder har for eksempel ikke blitt gitt mulighet til å forstå grunnlaget for avgjørelsen for dermed å kunne bestride den, de har ikke blitt tilstrekkelig informert i forkant om en forestående avgjørelse eller systemene er iverksatt uten at relevante brukergrupper har blitt hørt. Eksempelene demonstrerer derfor behovet for å innrette algoritmiske beslutningssystemer så de er i tråd med rettsstatlige og demokratiske verdier.

1.2.2 Tilnærminger til forklarbarhet

«Forklarbar AI» og «Ansvarlig AI» er områder i rask utvikling, hvor begreper, definisjoner og avgrensninger er omdiskutert. Selv om grenseoppgangen kan bli litt kunstig, kan det være til hjelp å se forklarbarhet både som en teknisk orientert diskusjon og en rettslig orientert diskusjon. Beaudouin et al. kaller det «the computer science angle» og «the legal and policy angle».⁶

Fra et teknisk, datavitenskapelig ståsted handler forklarbarhet om metoder for å generere forklaringer på prediksjoner fra komplekse modeller, ofte kalt svarte bokser. Dette forskningsområdet omtales ofte som *XAI – explainable AI*. Her er en viktig motivasjonsfaktor brukerens operasjonelle behov. For at en bruker skal kunne interagere meningsfylt med systemet, er det nødvendig med en viss grad av forståelse. Ellers vil systemet kanskje ikke tas i bruk eller ikke brukes til fulle. Forklaringer trengs derfor for å etablere menneskets tillit til maskinen.⁷

Fra et rettslig og samfunnsmessig ståsted kan man si at forklarbarhet handler om å ansvarliggjøre aktører som bruker kunstig intelligens med tanke på å sikre borgernes rettigheter og friheter. Dette feltet omtales av noen som *FAT – Fairness, Accountability and Transparency in AI*.⁸

⁵ Algorithmwatch (2020).

⁶ Beaudouin et al. (2020) s. 3, EUs høynivå ekspertgruppe på kunstig intelligens (2019).

⁷ Beaudouin et al. (2020) s. 6.

⁸ Se for eksempel: <https://dl.acm.org/doi/proceedings/10.1145/3351095>, «FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency».

Maskinlæringsteknologien har som nevnt noen problemområder. Den bærer med seg en risiko for bias, vilkårlighet, diskriminering og usaklig forskjellsbehandling. Uten åpenhet og kontrollmuligheter kan man ikke overskue om systemet kan få uønskede virkninger for grupper i samfunnet eller fange opp systemsvikt på en effektiv måte. Uten tilstrekkelig innsyn og begrunnelse kan heller ikke borgerne ivareta sine interesser i enkeltsaker.⁹ Forklarbarhet innebærer da å fremme kjennskap til systemenes eksistens, måten de er innrettet og bidrar til myndighetsutøvelse, og hvordan de kan forklares for ulike grupper av brukere og berørte.

1.2.3 Lovgivning og offentlige dokumenter

Regjeringen lanserte en strategi for kunstig intelligens i januar 2020. I strategien blir det uttalt at offentlig sektor er et satsningsområde, og at det blant annet innbefatter saksbehandlingsløsninger med elementer av kunstig intelligens. Strategien er ikke spesielt konkret, og enkelte steder får man inntrykk av at vanlig automatisering (med faste algoritmer) blandes sammen med kunstig intelligens-teknikker. Det blir også henvist til overordnede verdier og prinsipper med begreper som overlapper hverandre, men man kan i hvert fall trekke ut at regjeringen er opptatt av å ivareta rettssikkerhet og personlig autonomi, og at den ser forklarbarhet og transparens i systemene som en forutsetning for dette.¹⁰

Forvaltningslovutvalget leverte sin innstilling i 2019. Her blir det sagt at bruk av maskinlæring i saksbehandlingen byr på flere utfordringer, blant annet i relasjon til åpenhet og kontroll. Det blir også uttalt at «[m]askinlæringsmodeller mv. gir nye muligheter for automatisering av vurderingspregede kriterier». Utover disse uttalelsene nøyer utvalget seg med å si at det er viktig å være oppmerksom på problemstillingene fremover, og at de støtter det svenske forvaltningslovutvalgets standpunkt om at det er for tidlig å foreslå regulering. Samtidig er utvalget opptatt av at forvaltningsloven er og skal være teknologinøytral.¹¹

Arkivlovutvalget, som også sluttførte sitt arbeid i 2019, går mer inn på kunstig intelligens i sine drøftinger enn Forvaltningslovutvalget. Utkastets § 10 har en bestemmelse om dokumentasjon av systemer for hel- eller delautomatisert retts-

9 Flere internasjonale soft law-initiativer kan knyttes til dette ansvarliggjøringsperspektivet, som EU høynivå ekspertgruppes retningslinjer for etisk AI, OECDs prinsipper for ansvarlig AI og utvikling av ISO og IEEE -standarder for etisk og forståelig kunstig intelligens.

10 KMD (2020).

11 NOU 2019: 5 s. 261 og 268. Schartum (2019) s. 54 har kritisert Forvaltningslovutvalget for å være for tilbakeholdne og lite nytenkende. Han mener utkastet ikke gir «et godt nok grunnlag for en fremtidsrettet digital forvaltning», og at det ikke er noen grunn til å vente med regulering av saksbehandling med lærende algoritmer.

anvendelse, som omfatter eventuell bruk av maskinlæring. De er også inne på spørsmålet om arkivering av lærende algoritmer.¹²

Personvernforordningen (GDPR) er sentral for maskinlæringsmodeller som anvendes i velferdsforvaltning, fordi modelleringen gjerne krever store datasett med personopplysninger. Både individuelle rettigheter og systemkrav i GDPR får betydning for hvilke krav til forklarbarhet som stilles til algoritmiske beslutningssystemer.

Det finnes foreløpig ikke lovgivning som spesifikt angår utvikling og anvendelse av beslutningssystemer som integrerer kunstig intelligens, men det er på gang i EU. EU-kommisjonen la frem en White Paper i februar 2020, og i april 2021 la kommisjonen frem sitt forslag til en forordning for å harmonisere regler for bruk av kunstig intelligens-systemer. Forslaget har en risikobasert tilnærming, slik at systemer som regnes som høyrisiko må oppfylle strenge krav til åpenhet, kvalitet, dokumentasjon og kontroll. Systemer som benyttes av velferdsforvaltningen for å tildele, utmåle eller kalle tilbake ytelser og tjenester vil bli ansett som høyrisiko.¹³

1.3 Metode

Da tema for oppgaven var valgt, innså jeg at det var behov for å sette seg inn i og beskrive noen grunnleggende trekk ved teknologien og hvordan maskinlæringsmodeller utvikles, som en bakgrunn for å nærme seg forklarbarhet. Jeg deltok på to heldagsseminarer om maskinlæring beregnet på jurister og ikke-teknologer, og fulgte det åpne nettkurset «Elements of AI» som er laget av Universitetet i Helsinki og senere oversatt til norsk av NTNU.¹⁴

Systemer basert på maskinlæring er så langt ikke implementert i stort omfang i norsk offentlig sektor. Mange virksomheter er i en fase hvor de ser på hvordan de best kan analysere og nyttiggjøre seg stordata, og hvordan maskinlæring eventuelt kan inngå i dette. Det er få eksempler på beslutningsstøtte- eller kontrollsystemer med kunstig intelligens som faktisk er produksjonssatt.¹⁵ Jeg fant det derfor også nødvendig å se nærmere på hvordan disse teknikkene *kan* komme til å bli anvendt i saksbehandlingen. Til det formålet baserte jeg meg på internasjonale eksempler, informasjon fra norske offentlige etater og presenta-

12 NOU 2019: 9 s. 274.

13 EU-kommisjonens forslag til regulering av KI-systemer, Explanatory Memorandum pkt. 37 og 47 (EU-kommisjonen 2021).

14 NTNU og Universitetet i Helsinki (2020).

15 Reutter og Broomfield (2019) s. 1.

sjoner fra konferanser om maskinlæring og offentlig forvaltning som jeg har deltatt på det siste året.

Temaet bød også på noen metodiske utfordringer. Maskinlæring og algoritmisk forvaltning er et svært dynamisk område. Teknologien utvikler seg raskt og det er diskusjoner på tvers av fagdisipliner og sektorer. Begrepene er ikke definert eller samordnet. Det er derfor et kontinuerlig behov for å sammenholde, koordinere og videreutvikle begreper. Det krever en slags oversettelsesvirksomhet (som data scientister gjerne kaller *mapping*) hvor man forsøker å føre fagene nærmere sammen.

Cobbe har uttalt seg om dette. Hun mener det er en forståelseskluft mellom forskning på forklarbar og ansvarlig maskinlæring og forvaltningsrett. De begreper og metoder for forklaring som utvikles, lar seg ikke uten videre oversette og overføre:

While much research has sought to improve the standards of ADM systems, this has often not considered legal conceptions or standards for decision making. As a result, the processes and metrics for fair, accountable, and transparent machine learning developed through this research do not always translate easily to legal frameworks. There therefore exist gaps in understanding between technical research and administrative law as well as between administrative law and the processes and technical characteristics of ADM.¹⁶

Jurister må med andre ord bidra til å utvikle begrepene og evaluere metodene for forklaring, slik at de lar seg overføre til forvaltningsretten. For meg har det derfor vært et siktemål å bruke begreper fra den tekniske litteraturen i den rettslige analysen, men med noen tilpasninger. Jeg tror også en verdi ved å nærme seg jussen fra den tekniske siden er at man da kan oppdage andre problemstillinger og vinklinger enn man ellers ville gjort.

Min tilnærming til temaet, sammen med rettskildebildet, har medført at oppgaven bygger på en god del teknisk litteratur og litteratur i krysningsfeltet mellom datavitenskap og juss eller samfunnsvitenskap. Litteraturen er supplert av samtaler med teknologer, som har vært til god hjelp for min forståelse.¹⁷

Som jeg skrev ovenfor i avsnitt 1.2.3 er det foreløpig få rettskilder som direkte angår beslutningssystemer som integrerer maskinlæring. Lovforslaget fra

16 Cobbe (2018) s. 4. ADM står for *algorithmic decision-making system*, dvs. algoritmisk beslutningssystem.

17 Se avsnitt 7.6 for oversikt over samtaler.

EU-kommisjonen (Artificial Intelligence Act)¹⁸ ble offentliggjort 21.4.21, under to måneder før innleveringsfristen på oppgaven. Jeg henviser til bestemmelser i en del fotnoter og enkelte steder i teksten, men har ikke rukket å innlemme det fullt ut i analysen.

Jeg kjenner ikke til noe autoritativ rettspraksis om krav til algoritmiske beslutningssystemers forklarbarhet. I kapittel 5 vier jeg en del plass til en underrettsavgjørelse fra en nederlandsk domstol og et forhåndsvarsel fra Datatilsynet. Dette er altså ikke rettskilder av noen spesiell vekt, men de er relevante fordi det nettopp er sparsomt med konkrete avgjørelser å vise til. De er interessante fordi de kan gi en pekepinn om hvordan slike spørsmål kan komme til å bli vurdert fremover.

I norsk forvaltningsrettslig litteratur er forklarbarhet omtrent ikke behandlet foreløpig.¹⁹ Det er dermed et stort behov for forskning på området. Fordi det foreløpig mangler beskrivelser og analyser av algoritmiske beslutningssystemer fra en norsk kontekst viser jeg til flere internasjonale eksempler i teksten.

1.4 Oversikt over fremstillingen

I kapittel 2 gir jeg en nærmere beskrivelse av teknologien og ser på hvordan maskinlæring kan tenkes brukt i velferdsforvaltningen. Jeg omtaler ulike måter algoritmene kan lære eller trenes på, beskriver forskjellen mellom faste og lærende algoritmer i automatisering av avgjørelser og går igjennom utviklings-trinnene for en maskinlæringsmodell. Jeg introduserer også noen begreper som er knyttet til det man kan kalle algoritmisk forvaltning.

Kapittel 3 tar for seg forklarbarhet fra et teknisk ståsted – «the computer science angle»²⁰. Hensikten er å utforske hva det er ved maskinlæringsteknologien som gjør den obskur og nødvendig å forklare. Jeg presenterer ulike forklaringsbegreper og forklaringsmetoder fra den teknisk orienterte litteraturen, og utleder hvilke begreper som vil være nødvendige fra et forvaltningsrettslig synspunkt.

18 EU-kommisjonen (2021).

19 Dag Wiese Schartum har ved enkelte anledninger understreket at forklarbarhet kan være et problem ved bruk av maskinlæringsmodeller i offentlig saksbehandling. (For eksempel: Schartum (2019) og innlegg på eForvaltningskonferansen 15.3.21). Utover det kjenner jeg ikke til annet enn det som er skrevet i NOU 2019: 5 og NOU 2019: 9, og i studentoppgavene JOU 2020: 5 *Kunstig intelligens – Muligheter og risikoer i velferdsforvaltningen* og JOU 2020: 4 *Borgernes rettigheter i møte med det digitale samfunn og den offentlige forvaltning*.

20 Jf. Beaudouin et al. (2020) s. 3.

Jeg beskriver også et eksempel fra NAV på hvordan forklaring kan foregå i praksis.

I kapittel 4 ser jeg på hvordan algoritmiske beslutningssystemer kan utfordre borgernes rett til en åpen og forståelig forvaltning. Forklarbarhet i denne sammenhengen blir et spørsmål om hvordan man kan gjøre systemenes eksistens kjent for offentligheten, vise hvordan de virker og sikre muligheten for innsyn og medvirkning fra borgerne.

Temaet for kapittel 5 er rettssikkerhet. Jeg omtaler tre eksempler på algoritmiske beslutningssystemer (det nederlandske kontrollsystemet SyRI, IB sin karakteralgoritme fra 2020 og den østerrikske AMS-algoritmen for oppfølging av arbeidssøkende) for å finne ut hva slags forklaringer som er nødvendige for å ivareta borgernes rett til privatliv, personvern, kontradiksjon og ikke-diskriminering ved algoritmisk forvaltning. Jeg vurderer også om forklaringsmetodene som omtales i den tekniske litteraturen er egnet til å ivareta forvaltningsrettslige krav til begrunnelse og forsvarlig skjønnsutøvelse.

I kapittel 6 oppsummerer jeg mine funn. Jeg gjør også noen betraktninger om systemkrav, kontroll og saksbehandlingsregler som er tilpasset algoritmisk forvaltning.

2 Bruk av maskinlæring i velferdsforvaltning

2.1 Innledning

Maskinlæringsteknologien kan underbygge og forsterke utviklingen mot mer digitaliserte, automatiserte velferdsstater. På et overordnet nivå kan maskinlæring være et godt hjelpemiddel for nytenkning og utvikling av programmer, ytelser og tjenester i velferdsstaten. Innenfor saksbehandling kan maskinlæring bidra til automatisering og beslutningsstøtte. Teknologien kan potensielt bidra til økt effektivitet og høyere kvalitet i saksbehandlingen.²¹ I dette kapitlet vil jeg se nærmere på hvordan.

Avsnitt 2.2–2.3 gir en overordnet fremstilling av hvordan maskinlæring virker, og hvordan maskinlæring kan brukes til å generere profiler for å blant annet å forutse adferd, behov eller utvikling. Avsnitt 2.4 er en kort og abstrahert beskrivelse av fasene i utviklingen av en modell ved hjelp av veiledet læring.

Avsnitt 2.5–2.7 presenterer noen reelle og tenkte eksempler på hvordan maskinlæring kan bli brukt i velferdsforvaltningens saksbehandling, og introduserer noen begreper som kan være til hjelp for å analysere algoritmiske beslutningssystemer og algoritmisk forvaltning.

2.2 Maskinlæring som kunnskaps- og automatiseringsverktøy

Maskinlæring brukes for å avdekke og modellere sammenhenger og avhengigheter mellom variabler i store datasett. Maskinlæringsalgoritmene brukes for å klassifisere nye observasjoner og predikere fremtidig utvikling. Teknologien er en kobling av datavitenskap, statistikk, matematikk og informatikk. Det er utviklet mange ulike teknikker. Hvilken teknikk man velger avhenger av formålet og hva slags data man har tilgang til. Modellene er dynamiske ved at de kan lære av nye observasjoner, tilpasse seg og slik forbedre treffsikkerheten på fremtidige prediksjoner. Som Amariles skriver: Det er mye mer effektivt å lære data-maskinen hvordan den skal lære, enn å programmere hvert enkelt skritt den

21 FNs spesialrapportør på ekstrem fattigdom (2019), Veale og Brass (2019), Olsen et al. (2019), Kaun (2021).

skal ta. Maskinlæring kan derfor defineres som: «Systemer som blir bedre til å løse en bestemt oppgave når mengden av informasjon og erfaring øker.»²²

Maskinlæring er en underkategori av kunstig intelligens, ved siden av blant annet robotikk og språkgjenkjenningsteknologi. Kunstig intelligens brukes som betegnelse på mange ulike teknikker som simulerer menneskelig intelligens i måten de løser problemer og fatter avgjørelser på.²³

Det er algoritmer som foretar beregningene som inngår i en maskinlæringsmodell. En algoritme kan beskrives som et program som inneholder en oppskrift eller instruks i programmeringsspråk (kode) til datamaskinen. En modell er settet av de regler som algoritmen har sluttet seg frem til.²⁴

Algoritmer kan brukes til å automatisere avgjørelser. Det er viktig å skille mellom faste og lærende algoritmer. De har ulik funksjon i et rettslig beslutningssystem. Faste algoritmer er regelbundet, og basert på en bestemt instruks om trinnvis fremgangsmåte. Det er den samme regelen som appliseres på hvert tilfelle. De utfører for eksempel logiske operasjoner med formen «Hvis X ... så Y».²⁵

Faste algoritmer er grunnsteinene i dagens automatiserte saksbehandling, ofte omtalt som ekspertsystemer. I systemer for automatisert rettsanvendelse utfører faste algoritmer vilkårsbaserte vurderinger som ikke har innslag av skjønn. I slike systemer er alle rettslige tolkningsspørsmål håndtert på forhånd, *ex ante*, og disse blir transformert til programkode. Et system basert på faste algoritmer er statisk og forutsigbart. Det endrer seg ikke uten at det omprogrammeres av utviklere og implementeres i ny versjon. Man kan si på forhånd hva avgjørelsene vil bli.²⁶

Utlendingsdirektoratet (UDI) er en offentlig etat som har helautomatisert enkelte kategorier av saker, som for eksempel statsborgerskap. Her skjer det ingen læring, sakene er helt regelbundet og vurderes etter forhåndsdefinerte kriterier. UDI har hjemmel i utlendingsforskriften § 17-7c til å treffe helautomatiske avgjørelser uten søkerens samtykke. I bestemmelsen står det eksplisitt at de helautomatiserte avgjørelsene ikke kan bygge på skjønnsmessige vilkår, «med

22 OECD (2019), Teknologirådet (2018), NTNU og Universitetet i Helsinki (2020), Amariles (2020) s. 276.

23 Larsson og Heintz (2020) s. 7.

24 NTNU og Universitetet i Helsinki (2020). Barocas og Selbst (2016) s. 677 viser til at modellen er «[t]he accumulated set of discovered relationships».

25 Bing (1977) s. 199, Schartum (2018) s. 200 flg.

26 Schartum (4.3.2020), NOU 2019: 5 s. Forvaltningslovutvalget betegner dette som «rettslig kravspesifikasjon».

mindre avgjørelsen er utvilsom».²⁷ Det finnes flere andre eksempler også. Ved Lov-2020-12-04-132 fikk for eksempel NAV hjemmel i til å helautomatisere avgjørelser om pensjon for ulike yrkesgrupper.

Ved maskinlæring er det ikke en ekspert som programmerer instruksjonene på forhånd. Maskinlæringsalgoritmen slutter seg frem til reglene ved inferens, det vil si statistisk analyse av data. Man snakker om å trene algoritmen, og at den lærer. Der faste algoritmer er statiske, kjennetegnes maskinlæringsalgoritmer ved autonomi og adaptivitet. *Autonomi* betegner at algoritmen kan lære fra data uten menneskelig programmering, og *adaptivitet* betegner at algoritmen endrer seg etter hvert som datagrunnlaget endrer seg.²⁸

Maskinlæring er altså en form for regelproduksjon, men forholdet mellom algoritme og rettsregel ser foreløpig ikke ut til å være mye behandlet i juridisk litteratur.²⁹ Australske Administrative Law Council skriver at faste algoritmer uttrykker rettsreglene i et beslutningssystem som *konstruert kunnskap*, mens dynamiske algoritmer uttrykker reglene gjennom *tillært kunnskap*, som de induktivt har sluttet seg frem til.³⁰ Amariles mener at måten reglene oppstår på er diffus og nærmest unndrar seg forvaltningens og borgernes oppmerksomhet: «[T]hey are drawn continuously from data flows and, thus, are hardly identifiable and recognizable to both the administration and the citizen.» Det er altså behov for å utforske disse problemstillingene nærmere.³¹

Maskinlæringsmodeller hjelper med å håndtere usikkerhet. Man bruker en modell fordi det er usikkert hvordan et fenomen forholder seg eller kommer til å utvikle seg videre. Med en kvantifisert sannsynlighet for et bestemt scenario har man et grunnlag for å ta en avgjørelse, selv om man ikke har sikker viten.³²

Maskinlæringsmodellens output kalles ofte *prediksjoner*. Ordet i seg selv understreker at det ikke er sikker kunnskap. At to variabler korrelerer på en statistisk signifikant måte betyr ikke at det er noe årsak - virkning forhold mellom dem. Derfor omtales også maskinlæringsmodeller som hypotesegenererende. De er nyttige verktøy for å oppdage sammenhenger man ellers ikke ville fått med seg, fordi de skjuler seg i komplekse og omfangsrike datasett. En påvist korrelasjon

27 UDI hørings svar (5.9.19).

28 NTNU og Universitetet i Helsinki (2020).

29 Coglianese og Lehr (2017) skriver om «*rule-making by robot*» og «*adjudication by algorithm*», men dette er i tilknytning til amerikansk forvaltningsrett.

30 Administrative Law Council (2004) s. 17-19. (Administrative Law Council førte tilsyn og ga råd i forvaltningsrettslige spørsmål i Australia. Det er nå underlagt riksadvokatembetet.)

31 Amariles (2020) s. 275.

32 NTNU og Universitetet i Helsinki (2020).

kan så være utgangspunktet for å gjøre videre undersøkelser, som på sikt kan gi kunnskap.³³

2.3 Læringsmetoder

Maskinlæringsmodeller kan systematiseres etter hvordan de lærer.³⁴ Ved *veiledet læring* (supervised learning) sitter man på historiske data med både inngangs- og utgangsvariabler. Utgangsvariablene representerer «fasiten». Formålet er å lage en modell som kan kategorisere nye tilfeller korrekt etter hvert som de inntreffer. Basert på settet med forhåndskategoriserte treningsdata lager algoritmen sine regler for hvilke tilfeller som tilhører i hvilken kategori. Prediksjonen er en beregnet sannsynlighet for at en observasjon eller tilfelle tilhører en klasse. Over tid, etter hvert som algoritmen lærer gjennom nye data, kan reglene forbedres og prediksjonene bli mer treffsikre.³⁵ *Prediktiv analyse* er basert på veiledet læring. NAVs sykefraværsmode (se avsnitt 3.4.5) er et eksempel på klassifisering ved veiledet læring, hvor utgangsvariablen er sykmelding eller friskmelding etter et visst antall uker. Prediksjonen blir oppgitt som en sannsynlighet.

Ikke-styrt læring (unsupervised learning) benyttes når man ikke har oversikt over datasettet, eller ønsker å undersøke om andre relevante korrelasjoner eller mønstre kan påvises. Data er altså ikke kategorisert, det er ikke noen ferdig «fasit» eller utgangsvariabler. Algoritmen lager selv reglene basert på likheter i egenskapene ved dataene.³⁶

Forsterket læring (reinforcement learning) anvendes i situasjoner hvor man har et mål, men ikke noen oppskrift på hvordan det skal nås. Man bruker lite eller ingen treningsdata, og modellen lærer seg gjennom prøving og feiling hva som fungerer. Forsterket læring bruker man først og fremst nevrale nett til.³⁷ En utfordring med forsterket læring er at algoritmene optimerer for enhver pris. De vil ikke ta andre hensyn enn å nå det angitte målet, og det er derfor viktig med kontrollmekanismer som sikrer at algoritmene holder seg innenfor etiske rammer.³⁸

³³ Teknologirådet (2018).

³⁴ Kurs i NAV (26.2.2020), Lovett (4.3.2020), Datatilsynet (2018) s. 6.

³⁵ Amariles (2020) s. 277.

³⁶ Amariles (2020) s. 278, Lovett (4.3.20).

³⁷ Amariles (2020) s. 278, Lovett (4.3.20).

³⁸ Lovett (4.3.20).

2.3.1 Forholdet mellom maskinlæring og profilering

Med *profilering* sikter man til kategorisering av personer basert på likhetstrekk og korrelasjoner i et datagrunnlag, og eventuelt det å ta avgjørelser basert på tilhørighet til en gruppe eller kategori. GDPR art. 4 nr. 4 definerer profilering som følger:

enhver form for automatisert behandling av personopplysninger som innebærer å bruke personopplysninger for å vurdere visse personlige aspekter knyttet til en fysisk person, særlig for å analysere eller forutsi aspekter som gjelder nevnte fysiske persons arbeidsprestasjoner, økonomiske situasjon, helse, personlige preferanser, interesser, pålitelighet, atferd, plassering eller bevegelser;

Maskinlæring er godt egnet til å lage profiler, både basert på veiledet og ikke-veiledet læring. Veiledet læring innebærer at man har en hypotese om profilens sammensetning, som testes på dataene. Når man bruker ikke-veiledet læring er det algoritmen som identifiserer profilene uten at de er kjent på forhånd.

Ved prediktiv analyse kan maskinlæringsmodellens tillærte regler om profiler brukes til å beregne og slutte seg frem til en persons fremtidige handlinger eller egenskaper.³⁹ Når slutninger om en enkeltperson foretas basert på gruppens egenskaper fremfor en reell individuell vurdering, oppstår det såkalte *group-to-individual problem*.⁴⁰ Det gir en risiko for feil, vilkårlighet, urettferdighet og diskriminering i avgjørelsene.⁴¹ Dette er bakgrunnen for at GDPR har et forbud mot helautomatiserte avgjørelser basert på profilering i artikkel 22, ved siden av regler som krever åpenhet, informasjon og innsyn i hvordan personopplysninger behandles. Jeg kommer tilbake til det i kapittel 5.

2.4 Utvikling av en maskinlæringsmodell ved hjelp av veiledet læring

Utviklingsløpet til en maskinlæringsmodell kan deles inn i faser.⁴² Fasene utgjør de ulike stegene man går igjennom når man bygger en maskinlæringsmodell og setter den i produksjon. De glir over i hverandre og er gjerne *iterative* (gjentakende). Gjennom prøving og feiling undersøker man hvilke variabler som skal med og hvordan de virker inn på prediksjonene.

39 Prediktiv analyse bør ifølge Greenstein sees som et mektig verktøy for kontroll og intervensjon, fordi modellene har autonomi og en form for normativitet – de kan bidra til å skape den virkeligheten de forutser. (Greenstein 2017 s. 99).

40 Oswald (2018) s. 12.

41 Gil-González og de Hert (2019) s. 610.

42 Fremstillingen her er basert på presentasjonen på NAVs kurs om maskinlæring 26.2.20.

Forslaget til regulering av KI-systemer fra EU-kommisjonen tyder på at flere aspekter ved modellutviklingen vil bli regulert på sikt, men inntil videre er det GDPR som har størst rettslig betydning for modellutviklingen.⁴³

1) Problemformulering

Man starter med en problemstilling som enten kommer fra, eller forankres hos, det som ofte kalles *produkteier*. I NAV-systemet vil det for eksempel være dem som har ansvar for en bestemt ytelse eller forretningsområde.

Det må dreie seg om et problem som ikke lar seg løse ved et regelbasert system. Det bør ikke være *for* komplekst, det vil si beroende på for mange variabler. Da vil det være nødvendig med veldig mye data.

2) Utvalg og innhenting av data

Utvalg og innhenting av data krever tett samarbeid med produkteier, og er basert på hypoteser om ulike variablers sammenheng og relevans for problemet. Riktig utvalg av relevante egenskaper og dataenes kvalitet er vel så viktig som mengden av data.⁴⁴

Datainnhøsting krever en rekke avklaringer og tillatelser. Det må skje i henhold til personvernlovgivningen, herunder prinsippene om formålsbegrensning og dataminimering i GDPR artikkel 5, som innebærer at man ikke bruker mer data enn det som er nødvendig. Valg av variabler og datakilder er som nevnt en iterativ prosess. Det gjøres tilpasninger etter hvert som modellen er under utprøving.⁴⁵

3) Databearbeidelse

Det er viktig å gjøre seg godt kjent med dataene, og til dette trenger man også å samarbeide med fagsiden. Det kan være behov for å rense eller rydde opp, i dataene, for eksempel ved å håndtere datapunkter som mangler, anomalier, eller feil i merkingen.⁴⁶ Både ved utvalg og bearbeidelse av data er det viktig at produkt-eier er involvert med sin fagkunnskap, ofte kalt *domenekunnskap*. Personer som har jobbet innenfor et fagområde kjenner vil som regel ha god kunnskap om historikken og konteksten for dataene. Slik kunnskap er blant annet viktig med tanke på å unngå uønsket slagside eller *bias*.

⁴³ EU-kommisjonen (2021).

⁴⁴ Datatilsynet (2018) s. 7.

⁴⁵ Jaedick (4.3.20).

⁴⁶ I EU-kommisjonens lovforslag inneholder artikkel 10 *Data and Data Governance* en rekke kvalitetskrav til datasettene som brukes til trening, validering og testing av modellen og hvordan de skal håndteres (EU-kommisjonen 2021).

4) Valg av algoritme, modellering og testing

Det finnes mange ulike typer algoritmer for maskinl ring, og disse er i kontinuerlig utvikling. Hva slags algoritme man velger beror p  kompleksiteten i dataene, og behovet for forklarbarhet. Ofte kan det v re tilfellet at en mer avansert algoritme gir st rre treffsikkerhet, men at det samtidig gj r modellen mindre forklarbar. Se figur 8 i avsnitt 3.2.3 for en illustrasjon av denne sammenheng.

F r man trener algoritmen p  dataene deles datasettet inn i treningssett og testsett. Det kan foreg  ved en tilfeldig oppsplitting. Alternativt kan man bruke en statistisk metode til dette. Testdataene brukes for   finne ut hvor godt modellen presterer. Hvis modellen er *overfitted* (den er trent slik at den er for godt tilpasset treningsdataene), vil det for eksempel vise seg ved at treffsikkerheten faller n r man pr ver modellen mot testdataene.

Det er viktig med m l p  algoritmens treffsikkerhet som er mer beskrivende enn en generell suksessrate (andelen korrekte prediksjoner). En *feilmatrix* (confusion matrix) gir et overblikk over hvordan modellen tar feil. Man beregner andelen ekte positive og negative, og falske positive og negative for hver klasse av utfall. Figur 1 viser en ikke-utfyllt feilmatrix for en klassifiseringsalgoritme med to klasser av mulige utfall:

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figur 1: Feilmatrix for klassifiseringsmodell ⁴⁷

En feilmatrix er ogs  et styringsverkt y. Dersom det kan ha alvorlige konsekvenser   klassifisere feil, vil man for eksempel prioritere   ha en lav andel falske

⁴⁷ Kilde: <https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/>, oppslag 3.1.21.

negative. Dersom avgjørelsene kan oppleves belastende eller stigmatiserende er det viktig med en lav andel falske positive. I et system som predikerer risiko for uregelmessigheter med tanke på uttrekk for kontroll (se for eksempel om SyRI i avsnitt 5.2.3), vil man antakelig prioritere å la noen unnsnippe en kontroll selv om det egentlig var grunnlag for det (falske negative) fremfor å utsette personer for ugrunnet kontroll (falske positive).

Ulike mål på modellens feilrate og presisjonsnivå er også viktige med tanke på å fange opp og håndtere mulige diskriminerende eller urettferdige avgjørelser. Man ønsker for eksempel at modellen skal være like treffsikker uavhengig av befolkningssegment eller gruppetilhørighet. Det kalles *demografisk paritet*. Jeg kommer tilbake til algoritmisk slagside og diskriminering i avsnitt 5.5.

2.4.1 Systemdokumentasjon

Beskrivelsen her viser at modellutviklingsprosessen inneholder flere trinn, med problemformulering, diverse avklaringer, bearbeidelse av data, testing og verifisering. Dersom disse trinnene dokumenteres, kan det gi sporbarhet for vurderingene som er foretatt og en mulighet for den som utvikler systemet (for eksempel NAV eller et annet forvaltningsorgan) til å demonstrere etterlevelse. Det kan også være et utgangspunkt for å kreve innsyn for offentligheten. Slik dokumentasjon kan dermed i seg selv sees på som et bidrag til å forklare modellen eller beslutningssystemet. Jeg kommer tilbake til dette i avsnitt 3.5 om forklaringsbegreper og kapittel 4 om dokumentasjon som grunnlag for en åpen forvaltning.⁴⁸

2.5 Potensielle bruksområder med eksempler

Internasjonalt er det særlig to bruksområder for maskinlæringsmodeller som trer frem innenfor velferdsforvaltning, og det er barnevern (for å identifisere utsatte barn) og kontroll (for å avdekke juks og svindel med ytelser). I UK og USA brukes prediktive modeller for å sile bekymringsmeldinger til barnevernet for videre oppfølging eller for å fange opp utsatte barn tidlig. Det har også vært testet ut i Danmark, med den såkalte Gladsaxe-modellen. Tanken er at den prediktive modellen kan styrke saksbehandlerens beslutningsgrunnlag. I tillegg til sin egen erfaring og kompetanse kan saksbehandleren lene seg på statistisk kunnskap om risikofaktorer som er utledet av dataene.⁴⁹ Andre eksempler på

⁴⁸ EU-kommisjonens forslag til regulering av KI-systemer inneholder bestemmelser om dette, som artikkel 11 om *Technical documentation* og artikkel 12 om *Record-keeping*. (EU-kommisjonen 2021).

⁴⁹ Oswald (2018), Brown (2019), Algorithmwatch (2020). Et eksempel på en slik prediktiv modell i barnevernet er *Allegheny Family Screening Tool*, som blir omtalt i kapittel 4.

bruksområder er utmåling av arbeidsledighetstrygd, tildeling av bostøtte og identifikasjon av sosialt utsatt ungdom.⁵⁰

I Norge benytter bruker Skatteetaten en prediktiv maskinlæringsmodell for uttrekk av saker for kontroll av ligningen⁵¹, og Lånekassen for bokkontroll.⁵² Politiet bruker ulike verktøy som integrerer maskinlærning, blant annet i etterforskning.⁵³ Dette er bare noen eksempler.

Også NAV ønsker å utvikle beslutningssystemer som integrerer maskinlærning. Flere av prosjektene NAV IT, utvikling og data jobber med dreier seg om å gi beslutningsstøtte til veiledere (se figur 2). Prosjektet med beslutningsstøtte til veiledere ved innkalling til dialogmøte 2 blir omtalt nærmere i kapittel 3 og 5.

	Beslutningsstøtte til veiledere Trenger den arbeidsledige bistand fra NAV?		Fange opp avvik og forhindre potensielle feilutbetalinger
	Beslutningsstøtte til veiledere Bør den sykmeldte kalles inn til dialogmøte med NAV?		Matching: CV vs bemanningsbehov (stillinger)
	Beslutningsstøtte til veiledere Hvilke arbeidsrettede tiltak bør anbefales?		Maskinlærning for produksjon av syntetiske data

Figur 2: Bruksområder for kunstig intelligens i NAV⁵⁴

Jeg har ikke sett noen systematisert oversikt over hvordan maskinlæringsmodeller brukes eller kan tenkes brukt i velferdsforvaltning. Det følgende er en blanding av reelle og tenkte eksempler. Flere av bruksområdene reiser spørsmål om personvern, privatliv og retten til kontradiksjon for dem som får algoritmiske avgjørelser rettet mot seg.

50 Algorithmwatch (2020), Amariles (2020), Brown (2019), Munro (2019).

51 Datatilsynet (2018) s. 11.

52 Lånekassen (30.1.19): «Lånekassen bruker kunstig intelligens i bokkontroll». I meldingen opplyser de at de avdekker dobbelt så mange tilfeller av juks med bostøtte når de bruker kunstig intelligens som ved tilfeldig utvalg. <https://www.lanekassen.no/nn-NO/om-lanekassen/media-/nyhende/nyhetsarkiv/lanekassen-bruker-kunstig-intelligens-i-bokkontroll/>, oppslag 27.11.20.

53 Røger (Seksjon for digitalt politiarbeid ved Oslo politidistrikt, innlegg på LDOs årskonferanse 29.10.20).

54 Kilde: Lyngstad (14.5.20).

1) **Behandling av enkeltsaker**

Innenfor saksbehandling kan man se for seg flere ulike anvendelsesområder for maskinlæring. Det kan være tale om beslutningsstøtte til skjønnspregede vurderinger, forslag til formuleringer eller vedtaksutkast basert på saker som likner, kategorisering og siling av saker (for eksempel skille mellom saker til manuell og automatisert behandling), avviksvarsling (systemet gir beskjed dersom et vedtaksutkast avviker fra det som er vanlig for saker på området). Maskinlæring kan også være til hjelp ved systematisering og søk i praksis.⁵⁵ Det kan også være at noen vurderinger kan helautomatiseres, noe jeg kommer litt inn på nedenfor.

2) **Kontrollformål**

Maskinlæring kan brukes til å avdekke feilutbetalinger eller tilfeller av misbruk. Det kan være basert på analyse av historiske data om misbruk eller algoritmer som finner anomalier.

3) **Individuell tilpasning av ytelser og tjenester**

Basert på kunnskap om hva som har fungert for andre i tilsvarende situasjon, kan man gjøre veiledningen mer relevant og tjenestene mer skreddersydd for brukerne. (Man kan sammenligne det med måten strømmetjenester som Netflix anbefaler innhold til kundene på.) NAV ser for seg å dele arbeidssøkende inn i segmenter (profiler) etter behov. Dersom dataene indikerer at et personer som tilhører et bestemt segment klarer seg like godt på egen hånd, kan man heller fokusere innsatsen på segmenter som erfaringsmessig har nytte av veiledning og oppfølging.⁵⁶

4) **Påvirke adferd**

Maskinlæring kan brukes i forbindelse med dulting (*nudging*)⁵⁷ av innbyggerne mot ønsket atferd eller for å være i forkant med veiledning. Informasjonen kan tilpasses brukernes individuelle behov.

5) **Identifisere sårbare grupper og individer**

Prediktiv analyse er velegnet til å oppdage sårbarhet eller risiko som skyldes en kompleks interaksjon mellom variabler som ikke ville blitt fanget opp ellers.

55 Utenriksdepartementet bruker for eksempel språkgenkjenningsteknologi for å søke opp relevant informasjon i landrapportene fra sine utestasjoner. Før de begynte å bruke maskinlæring i 2018 var det uoversiktlig og vanskelig å finne frem i de store tekstmengdene. (KMD, 2020).

56 Lyngstad (presentasjon på LDOs årskonferanse 29.10.2019), KMD (2020). Et slikt system for å differensiere arbeidssøkere finnes i Østerrike, den såkalte AMS-algoritmen. Systemet beskrives i mer detalj i avsnitt 5.5.2.

57 Dulting (*nudging*) defineres av Store norske leksikon som «små, uformelle virkemidler for å påvirke atferd i gunstige retninger». Begrepet stammer opphavelig fra adferdsøkonomi.

6) Generere syntetiske data

Ved å generere syntetiske data (data som etteraper ekte data, men likevel ikke er knyttet til reelle personer) kan man simulere effekten av lovgivning, tiltak og virkemidler på forhånd.⁵⁸

2.6 Algoritmiske beslutningssystemer

Maskinlæringsmodeller kan altså tenkes integrert i saksbehandlingen på forskjellige måter, som eksemplene ovenfor viser. Man kan ha maskinlæringsbaserte programvareutvidelser, som for eksempel utfører oversetting eller skriftgjenkjenning, som kobles på et beslutningssystem (såkalte «plug-ins».⁵⁹ Det kan være systemer med ulike grader av automatisering, ulike kombinasjoner av faste og lærende algoritmer, og ulike grader av menneskelig involvering.

Betegnelsene ADS og ADM opptrer ofte i litteratur om maskinlæring. Begge er forkortelser for *algoritmisk beslutningssystem* (algorithmic decision-making system). Det defineres som systemer som bruker kunnskap fra dataanalyse, ofte av persondata, til å fatte eller understøtte avgjørelser.⁶⁰

Algorithmwatch mener at man må se på algoritmiske beslutningssystemer som «sosio-tekniske rammeverk». Med det ønsker de å få frem at systemets betydning og konsekvenser beror på konteksten – organisatorisk, sosialt, politisk og økonomisk.⁶¹ Dette er et viktig poeng, med konsekvenser for hva åpenhet og forklarbarhet faktisk vil innebære ved algoritmisk forvaltning. Fra et forvaltningsrettslig ståsted blir det for snevert å fokusere på å forklare maskinlæringsmodellen uten å forholde seg til beslutningsprosessen som helhet, og konteksten rundt. Jeg vil komme nærmere tilbake til dette i de neste kapitlene.

I litteratur om maskinlæring og kunstig intelligens finner man også begreper for å beskrive menneskets rolle i et algoritmisk beslutningssystem (figur 3). *Human-in-the-loop* viser til situasjoner der et menneske kan gripe inn i eller overstyre beslutningsprosessen (for eksempel en saksbehandler), *human-on-the-loop* viser til den som designer og drifter systemet (typisk en utvikler eller data scientist),

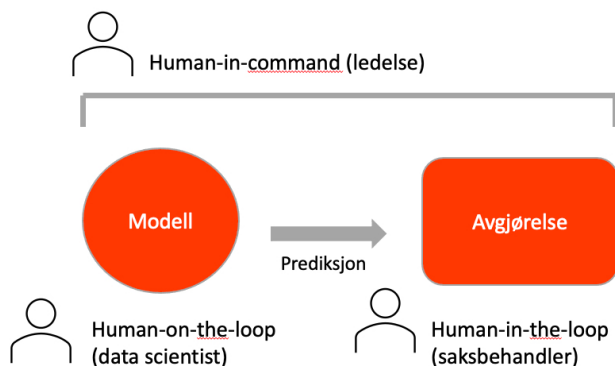
58 NAVs høringsinnspill til nasjonal strategi for kunstig intelligens (4.8.19).

59 Veale og Brass (2019) s. 123.

60 Castelluccia og Le Métayer (2019) s. 5, Olsen et al. (2019) s. 2. Jeg har også sett ADS brukt som forkortelse for «automated decision-making system».

61 Algorithmwatch: <https://algorithmwatch.org/en/automating-society-introduction/>, oppslag 7.12.20. Algorithmwatch overvåker og rapporterer som nevnt om algoritmiske beslutningssystemer i europeiske land.

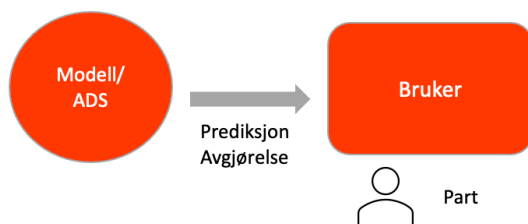
mens *human-in-command* viser til den overordnede kontrollen med systemets rettslige og økonomiske konsekvenser (typisk en ledelsesoppgave).⁶²



Figur 3: Aktørbegreper

Det er også vanlig å snakke om «brukere» i beskrivelser av interaksjonen mellom menneske og maskinlæringsmodell. Brukeren kan både være beslutningstakeren dersom det er et menneske i loopen, eller personen avgjørelsen retter seg mot (figur 5). Jeg skriver litt mer om brukerbegrepet i avsnitt 3.3.1.

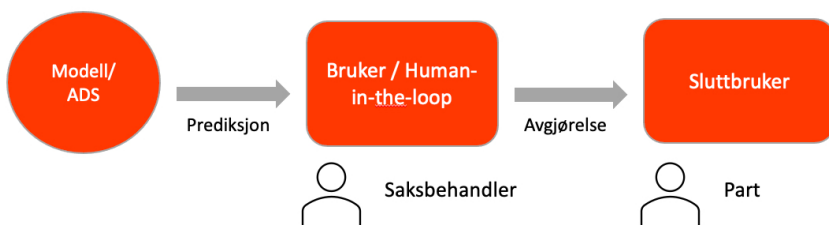
Beslutningssystemet kan altså være innrettet mot at maskinlæringsmodellen fatter automatiske avgjørelser, eller at modellens prediksjoner brukes av beslutningstakeren til å informere avgjørelsen. Brown et al. bruker betegnelse *algorithmic decisions* og *algorithm-assisted decisions* for å få frem denne distinksjonen.⁶³



Figur 4: Prediksjonen inngår i automatisert avgjørelse rettet mot parten

62 EUs høynivå ekspertgruppe på kunstig intelligens (2019) pkt. 65.

63 Brown et al. (2019) s. 4.



Figur 5: Modellen som beslutningsstøtte for en saksbehandler

Overført til velferdsforvaltningens saksbehandling betyr det at maskinlæringsmodellen kan inngå i selve rettsanvendelsen (figur 4) og kunne ut i et automatisert vedtak eller en prosessledende avgjørelse rettet mot en part, eller at dens prediksjoner bidrar til saksbehandlerens beslutningsgrunnlag og slik utgjør del av sakens utredning (figur 5.) Som et samlebegrep for det bruker jeg *algoritmisk forvaltning*.

NAVs pågående prosjekter er som nevnt bare av den andre varianten, hvor maskinlæringen er beslutningsstøtte. Olsen et al. mener også at det er som beslutningsstøtte til menneskelige saksbehandlere at kunstig intelligens i størst grad vil prege offentlig forvaltning de nærmeste årene.⁶⁴ Jeg vil likevel drøfte både bruk av maskinlæring som beslutningsstøtte (figur 5) og maskinlæring som beslutningstaker (figur 4).

2.7 Nærmere om aktuelle bruksområder for velferdsforvaltningen

For at maskinlæring skal ha en funksjon i rettslige beslutningssystemer, må det være et skjønnsrom. Dersom lovgivningen kan transformeres til statiske regler i form av faste algoritmer er det ikke nødvendig eller aktuelt å benytte maskinlæring.⁶⁵

Rettighetsbestemmelser i velferdslovgivningen kan være av både fast og mer skjønnspreget karakter. Barnetrygd og pensjon er eksempler på ytelser som følger faste, skjønnsfrie regler. Behovsprøvd ytelser er beregnet på å ivareta konkrete og individuelle behov, og disse kan ha skjønnspregede vurderinger både på vilkårsiden (rettsanvendelsesskjønn) og rettsfølgesiden (fritt skjønn

⁶⁴ Olsen et al. (2019) s. 10.

⁶⁵ Bygrave (2020) s. 4.

ved utmåling).⁶⁶ Det er altså slike bestemmelser det kan være aktuelle å knytte maskinlæring til. Men i første omgang vil kanskje det mest aktuelle være å benytte maskinlæring til mindre vurderinger eller enkeltvilkår.

Jeg har ikke forutsetninger her og nå for å gå inn i en nærmere analyse av akkurat hvilke ytelser eller typer av bestemmelser som kan egne seg, og viser til det jeg har sagt ovenfor om generelle anvendelsesområder. Teknologien kan som sagt også gjøre at man velger å innrette ytelsessystemet og formulere vilkårene annerledes.⁶⁷

2.7.1 Automatisert skjønnsutøvelse?

En maskin kan ikke brukes til å utøve skjønn i tradisjonell forstand: Å ta stilling til hva som vil være et rimelig og hensiktsmessig resultat ut fra en konkret helhetsvurdering.⁶⁸ Men en mulighet kan være å bruke maskinlæring til å identifisere mønstre fra tidligere forvaltningspraksis. Når modellen appliseres på nye saker, kan det fremme effektivitet og likebehandling.

Det kan imidlertid by på en utfordring med justering av algoritmen. Dersom utfallet i sakene algoritmen har vurdert i neste omgang inngår i læringsgrunnlaget kan den utvikle en selvforsterkende tendens som gjør at «skjønnnet» over tid blir mer sneversynt. Man får det som kalles *feedback loops*, en tilbakekoblingsmekanisme.⁶⁹ Et alternativ hvor algoritmen ikke lærer etter at den er implementert vil heller ikke være tilfredsstillende, fordi praksis ikke vil utvikle seg på den dynamiske måten den ellers ville gjort.

Olsen et al. foreslår en hybrid variant der noen av sakene behandles i en automatisert rutine med maskinlæring, mens den andre delen behandles av mennesker. Slik kan modellen lære av den menneskelige skjønnsutøvelse og utvikle praksis, mens man fortsatt vil få fordelene av effektivitet og likebehandling som algoritmene bidrar til. De kaller det en «administrative Turing test». Men da har de altså gått bort fra en helautomatisert løsning.⁷⁰

66 Syse (2016) s. 40.

67 Broomfield (samtale 7.12.20). Teknologirådet (2018 s. 71) advarer imidlertid om at for mye individuell tilpasning og «hyperindividualisering» gjennom prediktiv analyse og risikoskåring av innbyggerne kan føre til at man undergraver solidariteten som velferdsstaten bygger på.

68 Bing (1977) s. 201–203.

69 Dette er i og for seg ikke bare et problem for algoritmisk forvaltning, men forekommer ellers også. Ikdahl har beskrevet det hun kaller rettskildesløyer i forvaltningens forståelse og anvendelse av praksis. Rettsanvenderens forhåndsforståelse preger utvalg og presentasjon av rettskildene, mens argumenter som ikke passer inn i bildet nedtones eller utelates. Det gir over tid en selvbebreftende og innsnevret rettsanvendelsesprosess (Ikdahl 2016 s. 314–315).

70 Olsen et al. (2019).

Et alternativ vil være å utforme skjønnsbestemmelser på en måte som egner seg for algoritmisk vurdering. Algoritmer kan programmeres til å optimere for ulike mål eller vekte variablene på bestemte måter, og dermed bidra til å oppnå bestemmelsenenes formål. Maskinlæringen kan da fungere som substitutt til det vi ellers tenker på som skjønnsutøvelse.⁷¹

Jeg går ikke videre med denne diskusjonen her, ettersom hensikten var å peke på noen av problemstillingene som reises. For mitt formål vil jeg forutsette at det er mulig å automatisere deler av rettsanvendelsen i en sak ved hjelp av maskinlæring.

2.7.2 Maskinlæringsmodellen bidrar til sakens utredning

Som beskrevet i kapittel 2 er maskinlæring godt egnet til å håndtere usikkerhet. Modellen kan predikere sannsynlig utvikling på områder hvor man ikke har sikre svar.

Forvaltningen fatter hele tiden avgjørelser uten sikker kunnskap, og maskinlæring kan være nyttig for å styrke beslutningsgrunnlaget. Et eksempel er NAVs modell for prediksjon av sykefraværets lengde til bruk i sykefraværsoppfølgingen (se avsnitt 3.4.5). Modellen predikerer hvor lenge en person kommer til å være sykmeldt basert på data om sykefravær fra andre i tilsvarende situasjon, og veilederen bruker prediksjonen til å vurdere om det skal kalles inn til dialogmøte 2.

2.7.3 Maskinlæringsmodellen benyttes til kontrollformål

Maskinlæring benyttes også i systemer som er innrettet for å oppdage mulig svindel med trygdeytelser. En algoritme som er utviklet for dette formålet leter etter uvanlige eller uventede kombinasjoner av egenskaper hos en bruker, eller bruker historiske data på tidligere saker med juks. Dersom systemet predikerer at det er en risiko, blir brukeren merket for nærmere undersøkelser. Dette kalles «*red flagging*». SyRI-systemet som omtales i kapittel 5 er eksempel på det.

Red-flagging innebærer altså at saken utredes nærmere før det fattes en endelig avgjørelse. Men det finnes også eksempler på kontrollsystemer som fatter helautomatiske avgjørelser. I Australia hadde man systemet *Online Compliance Intervention*, som fikk tilnavnet «Robodebt». Systemet opprettet automatiske tilbakebetalingskrav på trygdeytelser dersom mottakeren ikke dokumenterte at det ikke var utbetalt for mye.⁷²

⁷¹ Schartum (4.3.20).

⁷² Commonwealth Ombudsman (2017).

Robodebt-systemet ble sterkt kritisert i offentligheten, og Ombudsmannen trakk blant annet frem i sin rapport at helautomatiseringen gjorde at man i praksis hoppet over et saksbehandlingstrinn. Der saksbehandlerne tidligere hadde tatt aktive skritt for å hente inn korrekt informasjon, ble det i større grad overlatt til brukerne å opplyse saken. Sakene ble derfor ikke lenger like forsvarlig utredet før avgjørelsen. Dokumentasjonskravet var dessuten blitt mer firkan- tet i forbindelse med automatiseringen, og brukerne hadde ikke blitt tilstrekke- lig informert på forhånd hvilken dokumentasjon det var nødvendig å ta vare på.⁷³

Kontrollsystemer kan utgjøre en ekstra rettssikkerhetsrisiko sammenlignet med systemer som tildeler og beregner velferdsytelser. Det blir en balansegang mel- lom statens behov for å unngå at det utbetales ytelser uten at det er grunnlag for det, og hensynet til borgernes rett til kontradiksjon og privatliv. Jeg kommer tilbake til dette i kapittel 5.⁷⁴

73 Commonwealth Ombudsman (2017).

74 Motstandere av SyRI-systemet kaller prosjektet sitt «*Bij voorbaat verdacht*», mistenkt på forhånd. De mener det bør være et forbud mot systemer som kobler innbyggernes data uten at det er grunnlag for mistanke, fordi det strider imot den demokratiske rettsstaten og uskyldspresum- sjonen. (<https://bijvoorbaatverdacht.nl>, oppslag 8.12.20.)

3 Forklarbarhet fra et teknisk perspektiv

3.1 Innledning

I det forrige kapittelet viste jeg ulike måter å bruke maskinlæring i offentlig forvaltning. Noen av bruksmåtene finnes det eksempler på, mens andre bruksmåter er tenkelige uten å være realisert.

Bruken av maskinlæring kan gjøre det vanskelig å forstå hvordan forvaltningen vurderer og avgjør forvaltningssaker. Derfor er det fra ulike hold påpekt at bruken må *forklares* for utenverdenen. I dette kapittelet skal jeg redegjøre for hvordan man fra et teknisk perspektiv har tilnærmet seg spørsmålet om hvordan forklare modellens prediksjoner. Det er et vanskelig terreng å manøvrere i, fordi begrepene ikke har fast innhold. Forklarbarhet kan for eksempel både vise til et iboende trekk ved en maskinlæringsmodell, men også noe man tilfører ved hjelp av ulike forklaringsmetoder og formidlingsteknikker.⁷⁵

Selv om jeg tar utgangspunkt i den teknisk orienterte litteraturen, forsøker jeg samtidig å koble på forvaltningsretten som ledd i oversettervirksomheten jeg beskrev i metodeavsnittet (se avsnitt 1.3). Jeg avslutter derfor kapittelet med en foreløpig vurdering av forklaringsbegrepene og deres anvendbarhet for forvaltningsavgjørelser.

3.2 Hva er det som er vanskelig å forklare?

3.2.1 Svarte bokser

Man hører ofte uttrykket *svart boks* i forbindelse med kunstig intelligenssystemer og maskinlæring.⁷⁶ Det er ikke noe presist, analytisk begrep, men en metafor for situasjoner hvor man ikke får vite eller ikke kan forstå hvordan algoritmen virker. Maskinlæringsmodellen produserte et resultat (en prediksjon), men man aner ikke hvorfor resultatet ble som det ble.

⁷⁵ Heller ikke uttrykket «forklarbarhet» er det enighet om. Det kan både omtales som *explainability*, *interpretability* og *intelligibility*. Noen vil si at disse begrepene alle viser til det samme, mens andre vil si at det er nyanseforskjeller. Beaudouin et al. (2020) s. 7–8.

⁷⁶ *Opasitet* (ugjennomskinnelighet) er et uttrykk for det samme.

I en mye sitert artikkel fra 2016 påpeker Burrell at svart boks kan betegne veldig ulike situasjoner. Hun skiller mellom omstendigheter der informasjon om modellen bevisst holdes tilbake, der brukeren eller parten ikke får noen egnet forklaring eller at modellen er så kompleks at den ikke lar seg begripe.⁷⁷

To kjente eksempler på bevisst tilbakeholdelse av informasjon er COMPAS-systemet fra USA og SyRI-systemet fra Nederland.⁷⁸ COMPAS er en modell for å predikere sannsynligheten for at en straffedømt faller tilbake til kriminell aktivitet. Modellen brukes av dommere i det amerikanske rettsvesenet til å fatte avgjørelser om prøveløslatelse og kausjon. Selv om modellen anvendes i offentlig sektor, er den utviklet av et privat selskap som beskytter koden og tilhørende informasjon av forretningsmessige grunner. SyRI er et kontrollsystem for trygde- og skattemyndighetene som ble brukt for å trekke ut personer til ekstra kontroll. Nederlandske myndigheter fryktet at befolkningen ville innrette seg og manipulere systemet dersom de fikk innsyn i algoritmen. SyRI blir omtalt nærmere i avsnitt 5.2.3.

Svart boks kan også vise til situasjoner der brukeren eller den som berøres av avgjørelsen mangler den nødvendige kunnskap om kode og statistikk til å forstå algoritmene.⁷⁹ Uten meningsfullt innsyn i modellen eller en forståelig og hensiktsmessig begrunnelse for utfallet, kan avgjørelsen da fremstå som en sort boks. Jeg kommer tilbake til forbindelsen mellom forklarbarhet og formidling i avsnitt 3.4.4 og senere i kapittel 4 og 5.

Svart boks beskriver dessuten den kompleksiteten som oppstår i samspillet mellom de store datavolumene og koden i maskinlæringsalgoritmer. En algoritme kan bestå av millioner av variabler som er forbundet i et nettverk av avhengigheter. Hver for seg kan datasettene være oversiktlige og koden enkel, og det kan være fullstendig åpenhet om dem, men de er likevel ubegripelige. Burrell mener man kan se det som et misforhold mellom algoritmenes høydimensjonalitet på den ene siden, og behovet for språklig mening og menneskelig resonnement på den andre. Dette vil jeg gå nærmere inn på i det følgende.⁸⁰

3.2.2 Dimensjonalitet og rasjonalitet

Den matematiske barrieren mot forståelse omtales gjerne som *dimensjonalitetsforbannelse*. Den handler om dynamikken i modeller med mange variabler.

77 Burrell 2016 s. 3–5. Artikkelen heter “How the machine ‘thinks’: Understanding opacity in machine learning algorithms”.

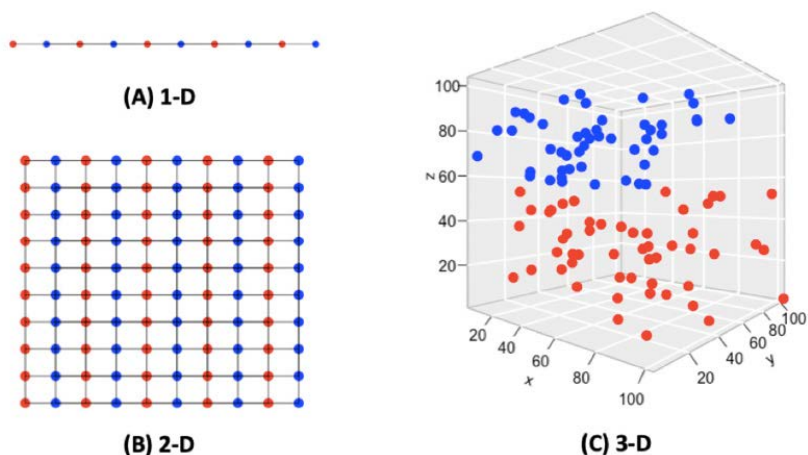
78 COMPAS står for Correctional Offender Management Profiling for Alternative Sanctions, og SyRI for Systeem Risico Indicatie.

79 Burrell (2016) s. 4.

80 Burrell (2016) s. 5.

For hver variabel eller egenskap man tilfører datasettet øker antallet potensielle datapunkter eksponensielt. For n variabler får man n dimensjoner. Mengden data og antallet datapunkter går langt over den menneskelige gripeevne. Til sammenligning klarer menneskehjernen bare å holde styr på 7 ulike enheter eller variabler om gangen.⁸¹

For å ha god representativitet i dataene må man ha mange observasjoner. Men selv med store datasett er det uunngåelig at datapunktene blir tynt og ujevnt fordelt (se figur 6). Maskinlæringsalgoritmer kompenserer for manglende datapunkter gjennom numeriske beregninger, det er det som gjør disse metodene så nyttige. Man ønsker jo nettopp en modell som har funksjonelle regler uten at den må trenes på alle tenkelige og utenkelige kombinasjoner av egenskaper. Men hva algoritmene faktisk gjør – hvordan innsatsvariabelen x blir til utfall y inni den svarte boksen – er ikke enkelt å formidle. Strümke sier det omtrent slik: «Numeriske beregninger i høydimensjonale rom lar seg ikke uten videre forklare analytisk til en bruker eller sluttbruker.»⁸²



Figur 6: Dimensjonalitetens forbannelse⁸³

⁸¹ Wachter et al. (2018).

⁸² Strümke (14.5.20).

⁸³ Kilde: <https://cofactorgenomics.com/curse-of-dimensionality-wk-16/>, oppslag 24.7.20. For hver variabel man tilfører modellen, øker antallet potensielle datapunkter eksponensielt. Når man beveger seg fra 2 til 3 dimensjoner kan man se hvordan datapunktene går fra full representativitet til at de tynnes ut og etterlater tomrom uten data.

Språklig sett handler forståelse (blant annet) om å identifisere seg med den andres logikk eller *ressonement*. Det er også det man intuitivt leter etter når man søker å forstå en prediksjon fra en maskinlæringsmodell. Men algoritmer fortolker ikke diagnosekoder, pikselverdier eller kjøpshistorikk inn i en sosial, kulturell og språklig kontekst slik mennesker gjør. Man kan derfor ikke gjenfinne noen menneskelig logikk i måten den klassifiserer på.⁸⁴

Poenget kan illustreres med dype nevrale nett som er trent opp til billedgjenkjenning. Såkalte *saliency maps* viser hvilke deler av bildet systemet fokuserer på for å klassifisere objektet. Det kan man se på som en slags forklaring, men den sier egentlig ikke noe om *hvordan* nettverket differensierer mellom husky og tverrfløyte (se figur 7).



Figur 7: Husky eller tverrfløyte? Forklaring ved hjelp av saliency maps.⁸⁵

Forklaringsmetoder som ikke viser hvordan algoritmen virker eller «resonner», vil ikke helt svare til våre forventninger til en begrunnelse i jussen. Det kan derfor være relevant å spørre om algoritmiske forvaltningsavgjørelser bør underkastes de samme krav til begrunnelse som menneskelige avgjørelser. Det drøftes i avsnitt 5.3.4.

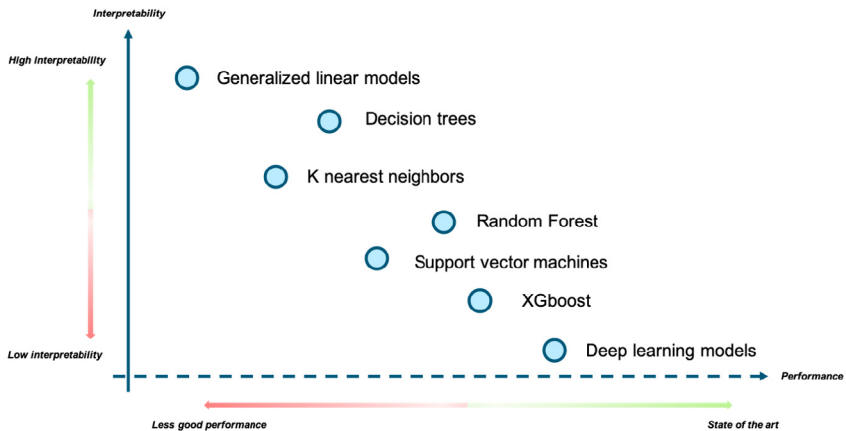
3.2.3 Forklarbarhet som funksjon av dimensjon og algoritmetype

Det er en utbredt oppfatning at noen algoritmetyper er mer obskure og uforståelige enn andre. Dette ser man ofte presentert i en variant av figuren under (figur 8). I den mest forståelige enden av skalaen finner man lineære modeller og beslutningsstrær, mens dype nevrale nett befinner seg i den minst forståelige.

⁸⁴ Doshi-Velez et al. (2017). Sandis (6.7.20) forklarer dette ved å sitere Wittgenstein: «If a lion could speak, you would not understand it».

⁸⁵ Kilde: Rudin (2019) s. 209.

Rudin omtaler de enkleste maskinlæringsmodellene som grunnleggende eller intuitivt forståelige (inherently interpretable). De er små, med få variabler eller få nivåer, og har en enkel sammenheng mellom inngangs- og utgangsverdier.



Figur 8: Sammenhengen mellom algoritmetype, ytelse og forklarbarhet.⁸⁶

Beslutningslister og beslutningstrær er algoritmer som kan gi intuitivt forståelige modeller så lenge de ikke blir for store. Jeg skal beskrive et eksempel på hver av dem.

Rudin og hennes team har utviklet modellen CORELS som en substitutt for den proprietære algoritmen COMPAS omtalt ovenfor. Hensikten var å demonstrere at enkle modeller kan være treffsikre, samtidig som at de ikke har de problematiske sidene som svarte bokser har.

CORELS-modellen er en beslutningsliste med noen enkle 'hvis-så' regler. Reglene anvendes i den rekkefølgen de står. Modellen predikerer om en straffedømt vil bli arrestert på nytt innen to år har gått. Den er altså et mål på tilbakefall til kriminell aktivitet, kalt *recidivism*. Prediksjoner om tilbakefall brukes som nevnt til å ta avgjørelser om blant annet kausjon og prøveløslatelse i det amerikanske rettsvesenet. CORELS har bare noen få linjer med kode, og kan fremstilles slik (figur 9):

⁸⁶ Kilde: <https://towardsdatascience.com/explainable-ai-the-data-scientists-new-challenge-f7cac935a5b4>, oppslag 6.7.20.

Table 1 | Machine learning model from the CORELS algorithm

IF	age between 18–20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21–23 and 2–3 prior offences	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest	

Figur 9: CORELS modell for prediksjon av tilbakefall til kriminell aktivitet ⁸⁷

Modellen er intuitivt forståelig for både brukeren (dommeren) og sluttbrukeren (den straffedømte), og den er forutberegnelig. Den viser at sannsynligheten for tilbakefall i stor grad er relatert til personens alder, kjønn og tidligere straffbare forhold («priors»). Testing viser at den er like treffsikker som COMPAS, som er en proprietær svart boks. COMPAS brukes i stor utstrekning i det amerikanske rettsvesenet.⁸⁸

Det neste eksempelet er en fiktiv modell som er laget for opplæringsformål. Den er nyttig fordi den er oversiktlig og illustrerer hvordan et beslutningsstre utvikles for å utføre klassifisering.

Også beslutningstrær klassifiserer data ved hjelp av beslutningsregler. Beslutningsreglene er utledet fra treningsdataene ved veiledet læring, og appliseres i en bestemt rekkefølge hver gang. Når modellen skal trenes starter man med hele treningsdatasettet ved treets rot. Hver regel leder til en oppdeling av datasettet ved en node (knutepunkt), som er utgangspunktet for treets greiner. Det finnes flere ulike algoritmer utviklet for dette formålet.⁸⁹

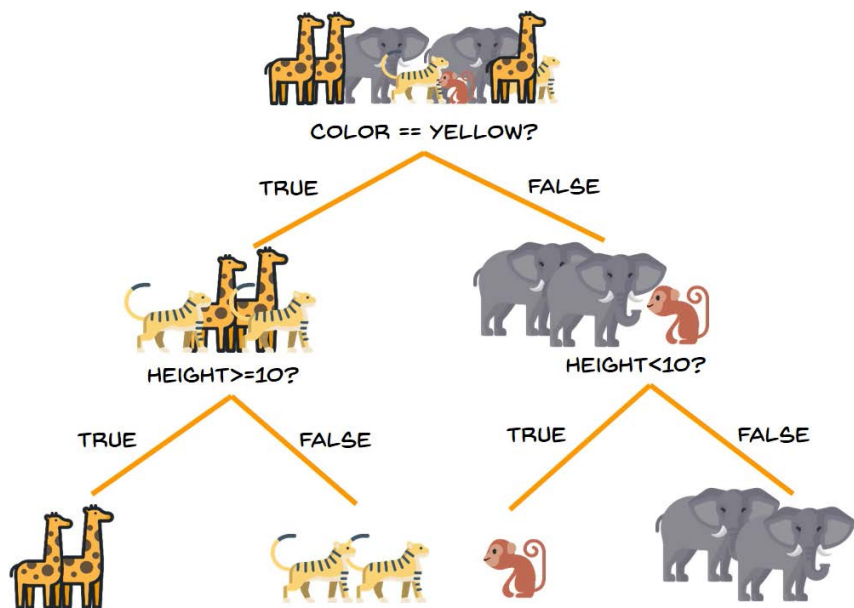
For å sikre at delingen foretas på en optimal måte bruker man *entropi* og *informasjonsgevinst*. Entropi er et mål på graden av vilkårlighet og uforutsigbarhet («uorden») i dataene på en gren. Informasjonsgevinst er et mål på reduksjonen av entropi etter en deling. Målet er derfor å oppnå høyest mulig informasjonsgevinst og mest mulig reduksjon av entropi ved hver deling av datasettet. Ved siste ledd har man blader, som representerer den endelige klassifiseringen.⁹⁰

⁸⁷ Kilde: Rudin (2019) s. 4.

⁸⁸ Omtalt i Rudin (2019) og Rudin (8.8.19).

⁸⁹ Simplilearn (2018).

⁹⁰ Simplilearn (2018).



Figur 10: Beslutningstre for klassifisering.⁹¹

Den første beslutningsregelen er farge = gul, fordi det gir størst reduksjon av entropi og høyest informasjonsgevinst. Men det er fortsatt uorden i klassene, så det er nødvendig å fortsette med delingsprosessen. På neste nivå er høyde den mest optimale beslutningsregelen ved begge knutepunktene. Med disse to nivåene er populasjonen klassifisert fullstendig, og entropien er 0.

Når den ferdige modellen appliseres, vil den ved hvert knutepunkt ta stilling til hvilken gren observasjonen mest sannsynlig tilhører. Modellens prediksjon er en sannsynlighet for at en enkelthendelse eller observasjon tilhører en klasse.⁹²

Beslutningslister og beslutningstrær oppfattes som intuitivt forståelige fordi det er mulig å observere hvilken logikk som styrer modellens kategorisering av datasettet. Man kan se akkurat hvordan x ble til y i dette tilfellet og hvordan utfallet vil bli for andre tenkelige inngangsvariabler. Guidotti et al. formulerer det slik:

⁹¹ Kilde: Simplilearn (2018).

⁹² Simplilearn (2018 a).

A model may be completely interpretable, i.e., we are able to understand the whole logic of a model and follow the entire reasoning leading to all the different possible outcomes. In this case, we are speaking about global interpretability.⁹³

Hvis man ser for seg en enkel modell brukt i en forvaltningsavgjørelse kan forvaltningsorganet gi parten innsyn i modellen for å vise resonnement bak prediksjonen (utfallet i saken). Parten kan da konstatere om avgjørelsen er riktig, og undersøke hva utfallet vil bli for andre. Det gir forutberegnelighet og dessuten en viss trygghet for at modellen er rettferdig og ikke forskjellsbehandler på usaklig grunnlag. Man vil derfor komme langt i å skape tillit og ivareta retten til kontradiksjon gjennom å publisere eller gi innsyn i modellen.

Men dette er bare et tenkt idealtilfelle. Antakelig er det ikke realistisk at modellene som utvikles og brukes til automatiserte avgjørelser eller beslutningsstøtte i forvaltningen vil være så enkle og gjennomsiktede. Et mer sannsynlig scenario er at det utvikles modeller med mange variabler og avanserte algoritmer, og det skyldes at det ofte er en korrelasjon mellom algoritmetype og modellens prestasjon. Figur 8 viser også denne sammenhengen: Jo mer avansert algoritme (jo svartere boksen blir), desto sikrere blir prediksjonene.

Rudin mener riktignok at denne sammenhengen er overdrevet, fordi det er konteksten og dataene som avgjør hvilken metode som gir det beste resultatet. Der som man har meningsfulle, tabulære data med god representativitet kan enkle modeller bli like treffsikre som komplekse, som vist ovenfor i eksempelet der CORELS ble sammenholdt med COMPAS.⁹⁴

Etter Rudins syn bør det derfor være et prinsipp at man velger den enkleste metoden når det ikke går på bekostning av kvaliteten på prediksjonene. Da får man jevnt over mer forståelige modeller. Dette er også foreslått som rettesnor i EUs høynivå ekspertgruppes prinsipper for etisk kunstig intelligens.⁹⁵

Til tross for et slikt prinsipp om å velge den enkleste algoritmen som gir akseptabel ytelse, kan det være at man kommer til (fortrinnsvis etter utprøving) at de enkleste modellene ikke har den ønskede graden av treffsikkerhet eller prestasjon til at man vil bruke dem i en beslutningsprosess. Da står man fortsatt overfor et forklarbarhetsproblem. I neste avsnitt vil jeg omtale noen grunner til at modellene må kunne forklares.

⁹³ Guidotti et al. (2018) s. 6.

⁹⁴ Rudin (8.8.19).

⁹⁵ Rudin (2019). EUs høynivå ekspertgruppe på kunstig intelligens (2019) skriver på s. 34 under overskriften «Forklarlighed»: «Har du undersøgt og forsøgt at anvende den enkleste og mest fortolkningsvenlige model til den pågældende anvendelse?».

3.3 Hvorfor er forklarbarhet nødvendig?

3.3.1 Brukeren må forstå hvordan modellen virker

En viktig motivasjonsfaktor for å forske på forklarbarhet i datavitenskap er brukerens behov for å forstå modellen. Dersom den som skal bruke systemet til å ta beslutninger ikke forstår modellen, vil den ikke bli brukt som forutsatt eller kanskje ikke bli tatt i bruk over hodet. Hvis modellen kan forklares gir det grunnlag for en mer kritisk og meningsfylt interaksjon mellom modell og bruker, som i sin tur kan føre til bedre modeller og økt tillit hos brukeren.⁹⁶

Dette vil naturligvis også være situasjonen dersom en maskinlæringsmodell skal integreres i en saksbehandlingsprosess i offentlig forvaltning. Imidlertid blir bildet av én type relasjon mellom modellen og en bruker for enkelt og unyansert.⁹⁷ En forvaltningsprosess involverer mange aktører med svært ulike behov og forutsetninger. Det betyr at det også kan bli behov for ulike forklaringer alt ettersom forklaringen retter seg mot parten, saksbehandleren, andre ansatte i etaten, klageorganet, offentligheten og så videre. Et eksempel på en slik nyansering finnes i avsnitt 5.4.3, hvor jeg analyserer lokale forklaringer rettet mot henholdsvis en part og en saksbehandler.

3.3.2 Modellene må være etterprøvbare for å sikre kontradiksjon og kontroll

Maskinlæringsmodeller kan lære autonomt mens systemet er live (i drift), eller i mer kontrollerte former som en programvareoppdatering.⁹⁸

Dersom modellen (algoritmen) har justert seg mellom to tidspunkter, vil man ikke få samme resultat begge ganger selv om betingelsene for øvrig er de samme. Etterprøvbarehet krever derfor en form for loggføring, for å kunne reprodusere avgjørelsene og kontrollere systemet.⁹⁹

Etterprøvbarehet kan sies å ha to nivåer, med ulike behov for dokumentasjon. På det enkleste nivået kan det gjelde muligheten for å kontrollere en enkeltprediksjon. I slike tilfeller kan det være tilstrekkelig å ha tilgang på algoritmens kilde-

96 Ribeiro et al. (2016), Royal Society (2019).

97 Man finner heller ikke noen nyansering av ulike aktørbehov i EUs forslag til regulering av KI-systemer. Artikkel 13 i *Transparency and provision of information to users* sier at systemene skal designes på en måte som gjør at brukeren kan forstå dem. «Users» sikter bare til personer som bruker systemet i en profesjonell kapasitet (f.eks. i etaten), og ikke parten i en avgjørelse, jf. artikkel 3 nr. 4. Det er altså tale om et systemkrav og ikke noen individuell rettighet.

98 Uten at jeg kan drøfte det i detalj her, antar jeg at modeller som lærer autonomt kan være problematiske å bruke i forvaltningsavgjørelser på grunn av manglende forutberegnelighet. Dette ble også vurdert av domstolen i SyRI-saken, se avsnitt 5.2.3.

99 EU-kommisjonens lovforslag har en bestemmelse om automatisk logging av hendelser i artikkel 12: *Record-keeping*. (EU-kommisjonen 2021).

kode og vektene (hvordan de ulike egenskapene vektet i det som utgjør prediksjonen), på tidspunktet for avgjørelsen.¹⁰⁰ Denne formen for etterprøvbarhet muliggjør altså grunnleggende kontradiksjon, der man kan kontrollere at det er brukt riktig inngangsverdi og at prediksjonen er riktig. Den kan også muliggjøre innsynskrav rettet mot å forstå hvordan et system fatter avgjørelser.

På det mer omfattende nivået kan etterprøvbarhet dreie seg om et behov for å rekonstruere modellen. Det vil være nødvendig ved systemrevisjon eller dersom selve grunnlaget for modellen (dens forutsetninger eller måten prediksjonene anvendes) bestrides.¹⁰¹ I disse tilfellene holder det ikke med algoritmens kildekode og vektene. Man må også ha de komplette settene med treningsdata for å kunne gjenskape vektene.¹⁰²

Jeg kommer tilbake til dette i avsnitt 4.4.2 om innsyn, avsnitt 5.3.1 i eksempelet om IB sin karakteralgoritme og i avsnitt 6.2 om systemkrav og kontroll.

3.3.3 Det må være mulig å finne ut hva som gikk galt

Hvis en svart boks gjør feil vet man ikke nødvendigvis hva det var som gikk galt. Det kan også være vanskelig å oppdage at feilen har skjedd. Jo mer forståelig eller forklarbar en modell er, jo enklere er det å identifisere feil og forbedre modellen.¹⁰³ Forklarbarhet er derfor et viktig hjelpemiddel for utviklerne i deres forbedringsarbeid.

Et trekk ved komplekse modeller er at de kan være ustabile og uberegnelige. I maskinlæringslitteraturen omtales feilkilder i form av *confounding* (sammenblanding eller forvirring) og *perturbasjoner*. Ved sammenblanding har modellen lært seg til å legge vekt på helt irrelevante faktorer. Et diagnoseverktøy som analyserer røntgenbilder bør for eksempel ikke legge vekt på at bildet er merket «portable», og en modell som skal skjelne mellom husky og ulv bør ikke legge avgjørende vekt på at det er snø i bakgrunnen. Da vil modellene være tilbøyelige

100 Prabhu (samtale 5.1.21). Det vil også være teknisk mulig å lagre en kopi av modellen i den versjonen som ble brukt til å fatte avgjørelsen.

101 Man kan kanskje vurdere å innføre en mulighet for systemklager på algoritmiske beslutningssystemer. Cobbe (2018 s. 8) er inne på dette: «The fact that individual automated decisions are heavily influenced by the processes around the system (i.e. selection of training data; design, construction, and training of models; and testing of systems) means that in order to properly evaluate those decisions it may be necessary to also evaluate those broader processes. As a result, the distinction between judicial review of policy and judicial review of individual decisions which exists for human decision-making may be significantly blurred or eroded for ADM.»

102 Dette kan by på problemer i forhold til GDPR. Dersom mange har valgt å slette sine data slik at de ikke lenger inngår i treningsdataene, vil det også gjøre modellen mindre reproducerbar (Prabhu, samtale 5.1.21).

103 Rudin (2019).

til å gjøre feil dersom de anvendes på data fra en litt annen kontekst enn den tillærte.¹⁰⁴

Perturbasjoner dreier seg om det som for mennesker vil være ørsmå detaljer som vi ikke legger merke til, som forstyrrer modellen. Det kan være en desimal for lite eller en forskjøvet piksel, som kan vippe modellen av pinnen og gjøre at prediksjonen skifter totalt. En husky klassifiseres som en tverrfløyte, og en skolebuss som en struts. Perturbasjoner kan være både vilkårlige og målrettede. Målrettede perturbasjoner kan brukes til å utføre angrep (game systemet), som hjelpemiddel for å finne svakheter med tanke på forbedring eller for å generere kontrafaktiske forklaringer.¹⁰⁵ (Kontrafaktiske forklaringer omtales også i avsnitt 3.4.2 og 5.4.2.)

Eksempelene jeg viste til her er hentet fra nevrale nett for billedgjenkjenning. Slike problemer vil være mindre uttalt for maskinlæringsmodeller det er naturlig å se for seg i velferdsforvaltning, med analyse av tekst og tabulære data. Det kan likevel være en mulig kilde til uforutsigbarhet og vilkårlighet.

3.3.4 Maskinlæringsmodeller har et problem med slagside

Bias (slagside, skjevhet) i maskinlæringsmodeller er en problemstilling det etter hvert har blitt mye oppmerksomhet på, særlig i forbindelse med forskning på ansvarlig og etisk kunstig intelligens. Skjevheter skyldes den sosiale, historiske og kulturelle preging av data, algoritmer og deres bruk, og gjør seg gjeldende i alle faser i en modells livssyklus. De kan være kodet i selve datasettet eller knytte seg til de forutsetninger og valg som utviklerne tar når de velger ut variabler, vasker data, justerer parametere og setter terskler. Slagside kan føre til avgjørelser som strider imot diskrimineringslovgivning, menneskerettigheter og etiske prinsipper dersom man ikke er bevisst og kompenserer for dette.¹⁰⁶

Mehrabi et al. har identifisert hele 23 ulike former for slagside i sin oversiktsartikkel om temaet.¹⁰⁷ Jeg nøyer meg med å nevne noen av dem, som en bakgrunn for behandlingen av temaet i avsnitt 5.5.

Innledningsvis er det viktig å være oppmerksom på at data ikke er objektive eller «sanne». En maskinlæringsmodell utvikles gjerne med utgangspunkt i historiske data som etaten allerede sitter på, som ble samlet inn, systematisert og

104 Eksempelene er hentet fra Ribeiro (2016) og Rudin (2019).

105 Wachter et al. (2018), Rudin (2019), EUs høynivå ekspertgruppe på kunstig intelligens (2019), Lovett (4.3.2020), Prabhu (samtale 5.1.21).

106 Burrell (2016), Kaufman (innlegg på LDOs årskonferanse 29.10.20).

107 Mehrabi et al. (2019).

merket i en annen kontekst og med et annet mål for øyet. Prabhu (data scientist ved NAV IT, utvikling og data) sier:

Noe av faren ved bare å bruke data man henter ut fra fagsystemer, er at opplysningene kan være farget av hvordan NAV alltid har jobbet og tenkt at virkeligheten ser ut. Data er registrert inn etter en systematikk bestemt av NAV.¹⁰⁸

Når historiske data gjenbrukes i en modell vil de antakelser og holdninger som preget innsamlingen bli med videre og fortsette å strukturere virkeligheten.

Relatert til dette er historisk slagside, som viser til at algoritmene er ikke mer objektive enn de historiske dataene de trenes på. De fanger opp mønstre i data som gjenspeiler underliggende samfunnsstrukturer, og bruker disse til å gjøre prediksjoner. På den måten kan algoritmene gjenskape og forsterke ulikheter som samfunnet ønsker å gå bort fra.¹⁰⁹

Et mye omtalt eksempel er det amerikanske selskapet Amazons algoritmiske rekrutteringsverktøy, som systematisk foretrakk mannlige kandidater fremfor kvinnelige. Slagsiden skyldtes at algoritmen var trent på å gjenkjenne mønstre fra CVer og søknader mottatt over periode på 10 år, og disse var altoverveiende fra menn. Når den identifiserte ord som «kvinne» eller aktiviteter og erfaringer knyttet til det å være kvinne, ble CVen nedrangert.¹¹⁰

Dersom algoritmen trenes på faktiske utfall kan nye data som viser (for eksempel) at kvinner ikke jobber i Amazon, bekrefte algoritmens «regel» og styrke tendensen til at kvinner siles automatisk ut av rekrutteringsprosesser. Det kalles for en *vicious feedback loop*, en ondsinnet tilbakekoblingsmekanisme.

Man kan ikke korrigere for slagside og unngå diskriminering ved å fjerne beskyttede egenskaper som etnisitet, kjønn eller religiøs tilhørighet fra datasettet. Fordi de utgjør så grunnleggende strukturerende prinsipper i de fleste samfunn vil algoritmene likevel plukke opp mønstrene gjennom såkalte *proxy-variabler* eller stedfortredende variabler. Med det menes at mange variabler i datasettet er indikatorer også for andre egenskaper. For eksempel kan navn, adresse, matvaner og interesser alle være proxy-variabler for en persons etnisitet. Modellen kan dermed produsere tilnærmet samme resultater med samme

108 MEMU (15.1.18): «Stordata kan hjelpe folk ut i jobb», <https://memu.no/artikler/store-data-mengder-kan-hjelpe-folk-i-jobb/>, oppslag 16.12.20.

109 Mehrabi et al. (2019).

110 Reuters (11.10.18): «Amazon scraps secret AI recruiting tool that showed bias against women», <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>, oppslag 16.12.20.

treffsikkerhet, men uten at «etnisitet» er eksplisitt nevnt som kategori. Modellen kan systematisk forskjellsbehandle, men det kan være mindre åpenbart enn det ellers ville vært.¹¹¹

En annen form for slagside har sammenheng med representativitet i dataene. Det vil ofte være slik at det er en majoritet eller bestemte grupper det finnes mest kunnskap om og data på, mens det kan forekomme minoriteter med liten tilstedeværelse i dataene.¹¹² Det innebærer at det vil være større sikkerhet i prediksjoner rettet mot medlemmer av majoriteten, mens prediksjoner rettet mot minoriteter vil ha høyere feilrate eller være mer upresise. Det kan få urettferdige og urimelige konsekvenser, som i noen tilfeller kan utgjøre diskriminering.¹¹³

Et eksempel er *Allegheny Family Screening Tool (AFST)*. Det er en prediktiv analysemodell som skal hjelpe saksbehandlere med å sile bekymringsmeldinger til barnevernet i Allegheny fylke i Pennsylvania, USA. I sin kritikk av modellen sier Virginia Eubanks at verktøyet sammenkobler informasjon fra offentlige databaser, og har tilgang til veldig mye data om dem som er registrert som brukere av offentlige tjenester og ytelser. Barn i familier som av ulike årsaker har liten tilstedeværelse i disse databasene løper en større risiko for å få en falsk negativ skår. Systemet har ikke data til å predikere risikoen, og utsatte barn fanges kan-skje ikke opp.¹¹⁴ (Systemet omtales også i avsnitt 4.4.3.)

Jeg kommer tilbake til forholdet mellom forklarbarhet og algoritmisk diskriminering i avsnitt 5.5. I det neste avsnittet vil jeg gi et innblikk i hvordan forklarbarhet angripes i den tekniske litteraturen, og jeg presenterer et eksempel på hvordan NAV har nærmet seg problemstillingen.

3.4 Hvordan kan modellene forklares? Begreper og tilnærminger

Et sentralt begrepspar i litteratur om forklarbarhet er *global og lokal forklaring*. Den globale forklaringen beskriver hvordan maskinlæringsmodellen virker på et overordnet nivå, mens den lokale forklaringen angår enkeltprediksjonen. Man kan se på det som forskjellen mellom å forklare selve modellen og begrunne et resultat, eller mellom regelen og dens konkrete anvendelse.¹¹⁵

111 Goodman og Flaxman (2016), Doshi-Velez et al. (2017).

112 Minoritet i betydningen en mindre gruppe individer med noen felles egenskaper, ikke nødvendigvis med felles identitet.

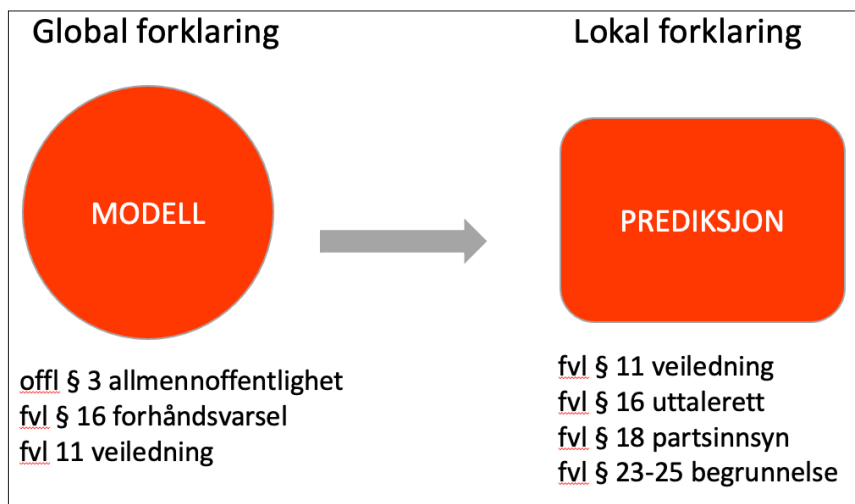
113 Goodman og Flaxman (2016).

114 Wired (15.1.18).

115 Beaudouin et al. (2020) s. 8., Rudin (2019).

Wachter et al. påpeker at en global forklaring kan være *ex ante* (på forhånd) og uavhengig av modellens faktiske anvendelse, mens den lokale forklaringen per definisjon må være *ex post* (i etterkant) fordi den gjelder et konkret utfall.¹¹⁶ Denne distinksjonen har betydning for diskusjonen om retten til forklaring i GDPR artikkel 22 (se avsnitt 5.2.2) og for måten man kobler de ulike forklaringsene til en saksbehandlingsprosess etter forvaltningsloven.

Hvis man tenker seg modellen anvendt i en saksbehandlingsprosess, må forvaltningen gi globale forklaringer for å besvare innsynskrav eller for å veilede og varsle borgerne om algoritmiske avgjørelser. Dersom det foreligger en avgjørelse i form av en prediksjon, må forvaltningen *ex post* kunne forklare den for parten. Da er flere av forvaltningslovens saksbehandlingsregler aktuelle (se figur 11).



Figur 11: Global og lokal forklaring i forbindelse med forvaltningsvedtak ¹¹⁷

3.4.1 Globale forklaringer

Global forklaring kan bety ulike ting.¹¹⁸ Det kan for det første bety en beskrivelse av den overordnede sammenhengen mellom inngangs- og utgangsvariablene.¹¹⁹ Et enkelt eksempel er den lineære modellen under (figur 12). For alle

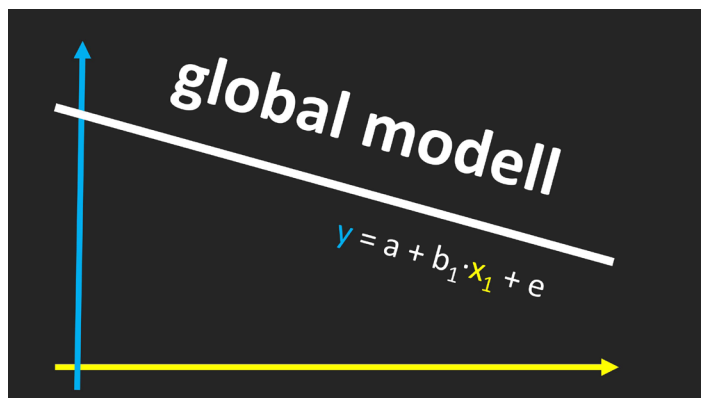
¹¹⁶ Wachter et al. (2017).

¹¹⁷ Figuren er en forenkling av hvordan de ulike reglene kan aktualisere seg. Hensikten er å få frem forskjellen mellom global og lokal forklaring.

¹¹⁸ Det er ikke et fast definert begrep, og ulike forfattere legger litt ulike ting i det. (Løland, samtale 17.8.20).

¹¹⁹ Kurs i NAV (26.2.20).

inngangsverdier x kan man lese ut av modellen hva y vil bli. Fordi funksjonen oppfører seg likt over hele gyldighetsområdet er modellen i seg selv global, og lokal og global forklaring sammenfaller.¹²⁰



Figur 12: Global modell ¹²¹

En global forklaring beskrives også som en angivelse av hvordan modellen opptrer «i snitt», eller hvordan den *pleier* å oppføre seg for bestemte grupper av personer. (For eksempel: 'Modellen vil vanligvis predikere et sykefravær på x uker for personer med diagnose y .) I tilfeller hvor man ikke står overfor en enkel og forutsigbar funksjon som i eksempelet ovenfor, kan man si noe om dette når modellen er trent og testet.¹²²

Beaudouin et al. har et bredere perspektiv på global forklaring. De sammenligner den med en brukerveiledning (se figur 13). Den inneholder et overblikk over modellen (den overordnede sammenhengen mellom inngangs- og utgangsverdier), hvordan algoritmen lærer, hva algoritmen optimerer for, hvilke treningsdata som er brukt, eventuelle kilder til skjevhet i treningsdataene og hvilke begrensninger modellen har. De mener at dette vil være tilfredsstillende for de aller fleste formål. At det er en brukerveiledning tilsier at stoffet må formidles på en måte som er tilpasset brukergruppen.¹²³

¹²⁰ Løland (14.5.20).

¹²¹ Kilde: Løland (14.5.20).

¹²² Prabhu (samtale 1.7.20), Løland (samtale 17.8.20).

¹²³ Beaudouin et al. (2020) s. 13.



- **Overview of the model,**
- **How it learned,**
- **What training data,**
- **Limitations to the model and use restrictions.**

Figur 13: Global forklaring som brukerveiledning, og for ansvarliggjøring ¹²⁴

Global forklaring er altså et fleksibelt og skalerbart begrep. Det kan være den modellnære varianten som angir sammenhengen mellom x og y , eller som gir en liten oppsummering av hva som er modellens anvendelsesområde og gjennomsnittsprediksjon. Men global forklaring kan også brukes i en videre betydning, som er mer passende for «legal and policy»-perspektivet (omtalt i avsnitt 1.2.2). Der inngår informasjon om beslutninger og forutsetninger som er tatt i utviklingsfasen, modellens muligheter og begrensninger og hvordan den er forutsatt implementert i en beslutningsprosess. I sin mer utvidede betydning blir en global forklaring en form for dokumentasjonskrav, der aktøren (forvaltningsorganet) ansvarliggjøres gjennom å måtte vise at modellen er utviklet på en forsvarlig måte.

I avsnitt 3.2.3 viste jeg CORELS-modellen som et eksempel på en intuitivt forståelig modell. Fordi den er så enkel, er det ikke nødvendig å forklare nærmere hvordan modellen har beregnet seg frem til et resultat. Men det kan likevel være nødvendig å stille dens utviklere til ansvar gjennom å spørre om dataene den bygger på er representative, hvordan beslutningsreglene ble utledet av dataene, hvorfor re-arrestasjon innen to år i det hele tatt er en egnet indikator på tilbakefall til kriminell aktivitet, hvor treffsikker modellen er og om den presterer ulikt for ulike befolkningsgrupper. Slike utvidede globale forklaringer kan altså være påkrevd selv om modellen er enkel.

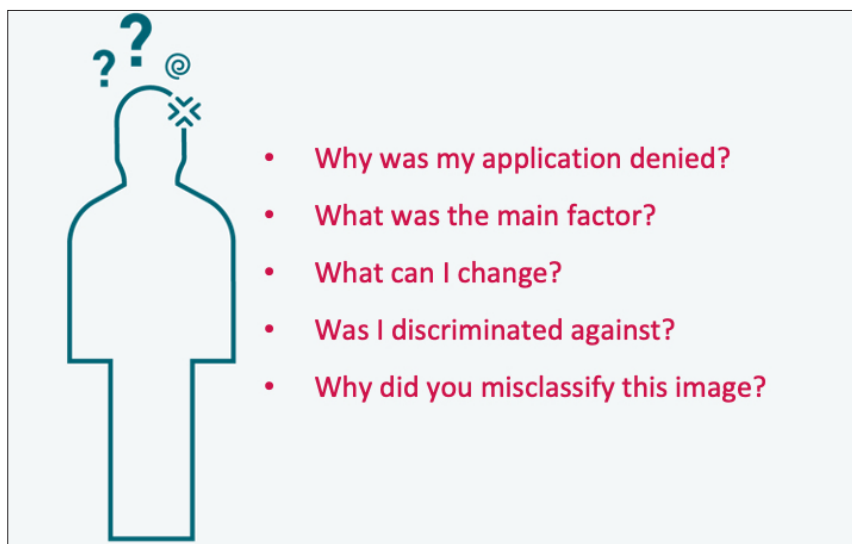
3.4.2 Lokale forklaringer

En lokal forklaring skal som nevnt belyse modellens prediksjoner eller enkelt-utfall. Hvilken informasjon som inngår i en lokal forklaring kan variere alt etter

¹²⁴ Kilde: Beaudouin et al. (2020) s. 13.

hva som er teknisk mulig og hensiktsmessig, forutsetningene til den som spør og hva informasjonen skal brukes til.¹²⁵

Flere forfattere har vært opptatt av å fremheve at det nettopp er brukerens behov som må stå i sentrum for hvordan det forklares. En bruker som spør «Hvorfor fikk jeg dette utfallet?» ønsker ikke først og fremst å få vite hva som foregikk inni den svarte boksen (såkalt *matematisk transparen*s), men å få svar som er til hjelp i sin konkrete situasjon.¹²⁶ Beaudouin et al. illustrerer det på denne måten:



Figur 14: Brukerens spørsmål til modellens enkeltutfall.¹²⁷

125 Lipton (2017), Ribeiro (2016), Wachter et al. (2017), Beaudouin et al. (2020), Løland (samtale 17.8.20).

126 Sandis (6.7.2020).

127 Kilde: Beaudouin et al. (2020) s. 14.

Overført til en forvaltningssak, kan man se for seg spørsmål som disse (figur 15):



Figur 15: Partens spørsmål til avgjørelser fra forvaltningen.¹²⁸

Med andre ord er det området man befinner seg på som avgjør hva som skal forklares og hvordan. Ifølge Olsen et al. vil forklarbarhet på forvaltningsrettens område dermed handle om de rettslige kravene til begrunnelse:

[W]e think the quest for black box explainability (which we call mathematical transparency) should give way to explainability in the legal sense (giving grounds for decisions).¹²⁹

Jeg er enig i at det er konteksten som definerer kravene til forklaring, men samtidig at det kan bli for snevert å redusere forklarbarhet til begrunnelse. Forklarbarhet angår flere sider av relasjonen mellom borgerne og forvaltningen, som jeg vil vise i kapittel 4 og 5.

Fordelen med lokale forklaringer er at de ikke nødvendigvis forutsetter at man ser inn i den svarte boksen eller fullt ut kjenner systemets underliggende logikk. De kan dermed brukes til å belyse prediksjoner fra komplekse modeller som

¹²⁸ Fritt etter Beaudouin et al. (2020) s. 14.

¹²⁹ Olsen et al. (2019) s. 5.

ellers ville vært lite forklarbare.¹³⁰ Det finnes ulike forklaringsmetoder, og jeg vil presentere noen av dem i det følgende.

1) Forklaring basert på parametere

En lokal forklaring kan for eksempel være en angivelse av hvilke faktorer eller variabler som spilte inn i prediksjonen, og hvordan de ble vektet. En måte å identifisere faktorene på er å benytte *approksimasjoner*, som er en forenklet versjon av den egentlige, underliggende modellen. En slik metode blir beskrevet i mer detalj i neste avsnitt.¹³¹

2) Kontrafaktiske forklaringer – «Hva hvis?»

Den kontrafaktiske forklaringen beskriver hva slags endring i input som ville føre til et annet utfall. Den mest optimale kontrafaktiske forklaringen viser hva som er den minste justeringen av egenskaper som skal til for å endre utfallet.¹³² Et avslag på lån ledsages for eksempel av informasjon til kunden om hva som må endres for at søknaden skal innvilges i neste omgang.

Ett enkelt utfall kan belyses ved hjelp av mange kontrafaktiske forklaringer. Utfordringen er å velge relevante hypotetiske «hva hvis?»-situasjoner. Forklaringen bør være uformet med tanke på egenskaper sluttbrukeren har mulighet for og vil være villig til å endre. Man kan for eksempel påvirke sin kredittverdighet, men ikke sin alder.¹³³ Å gi gode kontrafaktiske forklaringer krever derfor domenekunnskap. Overført til låneeksempelet bør man kjenne til egenskaper og preferanser hos bankens kunder for å vite hva som vil være funksjonelle kontrafaktiske forklaringer for dem. Jeg skriver litt mer om kontrafaktiske forklaringer fra et forvaltningsrettslig perspektiv i avsnitt 5.4.

3) Forklaring ved bruk av eksempler – «case-based»

Forklaring ved eksempler er en forklaringsstrategi hvor man gir eksempler på tilfeller med lignende egenskaper, hvor utfallet har blitt det samme eller et annet enn i gjeldende tilfelle. Det kan både gi en forståelse for hvordan modellen opptrer mer generelt og bidra til å underbygge at den enkelte prediksjonen stemmer.¹³⁴

130 Castelluccia og Le Métayer (2019), Beaudouin et al. (2020).

131 Ribeiro et al. (2016), Beaudouin et al. (2020). Det er flere ulike forklaringsmetoder og det skjer mye utvikling innenfor dette feltet, men her har jeg begrenset meg til å omtale metoden som har relevans for eksempelet fra NAV i avsnitt 3.4.5.

132 Løland (14.5.20).

133 Løland (14.5.20). Data scientister snakker derfor om at kontrafaktiske forklaringer bør være «actionable»; orientert mot hva personen selv kan foreta seg for å endre situasjonen.

134 Lipton (2017), Doshi-Velez et al. (2017).

4) Forklaring gjennom visualisering

Visualiseringer forklarer gjennom å vise hva et nevralt nett har fokusert på for å komme frem til en prediksjon eller klassifisering. Det er særlig aktuelt for bildegjenkjenningssystemer. De omtales som saliency maps, feature importance med mer. Et eksempel på det viste jeg i figur 7 med bildene av huskyen. Dette er teknikker som ikke er så aktuelle for bruk i velferdsforvaltning, så jeg går ikke nærmere inne på slike forklaringsmetoder her.

3.4.3 Approksimasjoner for å forklare komplekse modeller

Med komplekse modeller (svarte bokser) har man altså begrenset tilgang til hva som skjer fra inngangsverdi x til utgangsverdi y . Det står i skarp kontrast til det som var tilfellet for de gjennomsiktede, intuitivt forståelige modellene jeg viste i avsnitt 3.2.3.

Fordi en kompleks modell kan variere på en uforutsigbar måte kan det være vanskelig å si noe overordnet og meningsfylt om hvordan den opptrer (det jeg kalte å gi en global forklaring i snever forstand).¹³⁵ I stedet kan man bruke metoder for å belyse mindre områder av modellen. Slike forklaringsmetoder går ut på å trene en enklere modell, også ved hjelp av maskinlæring, som legges «oppå» modellen som skal forklares.

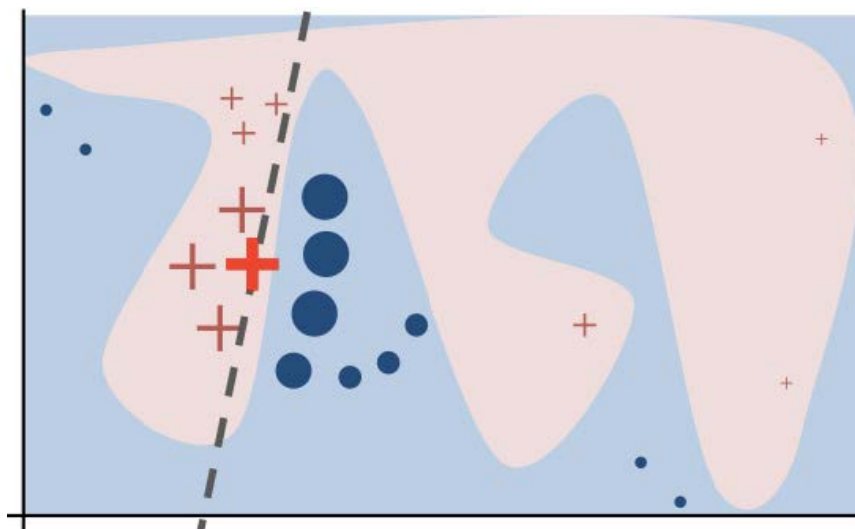
Slike metoder kan være *agnostiske* eller *modelluavhengige*. (SHAP-metoden som NAV har brukt til å forklare sykefraværsmodellen, er et eksempel på det.) Det betyr at de kan anvendes på alle komplekse modeller, uavhengig av algoritmetype. Hensikten er å forklare enkeltprediksjoner basert på analyse av inngangs- og utgangsvariablene til maskinlæringsmodellen. Det vil si at man forklarer *ex post* og lokalt, uten at det er nødvendig med innsyn i algoritmens kode, treningsdataene eller annen dokumentasjon knyttet til modellens utvikling. Man trener altså ikke se inni den svarte boksen, det holder med inngangs- og utgangsverdier.¹³⁶

I figur 16 representerer den lyserøde og lyseblå bakgrunnen den komplekse modellen, som opptrer svært uforutsigbart og ikke lar seg forklare globalt. I stedet forklarer man enkeltprediksjonen + (det røde uthevede plusstegnet) ved å trekke ut et begrenset utvalg av observasjoner i + sitt nærområde og lage en forståelig modell, i dette tilfellet en lineær funksjon, representert ved den svarte stiplede linjen. Det betyr at forklaringen er enkel, men den vil ikke gjelde for alle

¹³⁵ Løland (samtale 17.8.20), Kurs i NAV (26.2.20).

¹³⁶ Løland (14.5.20). Her må det understrekes at dersom etaten selv utvikler modellen vil den jo ikke være en fullstendig svart boks. Poenget med agnostiske metoder er at de kan brukes også der man ikke har kjennskap til hvordan modellen er utviklet.

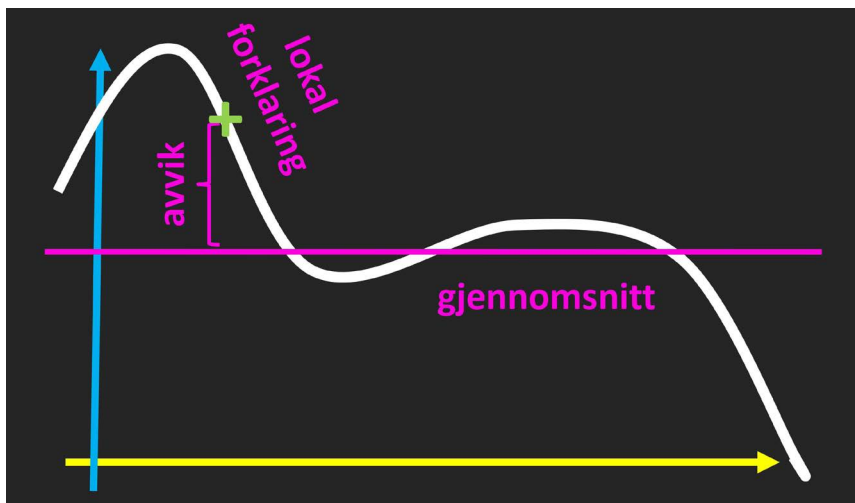
verdier av x . Den gir bare en tilnærmet forklaring for en liten del av modellens gyldighetsområde.



Figur 16: Lokal forklaring av kompleks modell ved hjelp av approksimasjon.¹³⁷

En approksimasjon basert på SHAP-metoden viser hvilke variabler som har virket inn på maskinlæringsmodellens enkeltprediksjon, og hvilken vekt hver av dem har hatt i det samlede resultatet. Det er dette som kalles *Shapley-verdier*. Summen av alle Shapley-verdiene er enkeltprediksjonen, for eksempel en sannsynlighet for at en individuell kunde vil misligholde et lån, eller sannsynligheten for at en sykmeldt person vil friskmeldes innen et visst tidsrom. Verdiene kan være positive og negative alt ettersom personens egenskaper trekker sannsynligheten opp eller ned i relasjon til gjennomsnittet.

¹³⁷ Kilde: Ribeiro et al. (2016) s. 1138.



Figur 17: Lokal forklaring ved bruk av SHAP-metoden ¹³⁸

Slike forklaringer kan være unøyaktige, fordi de nettopp forenkler modellens kompleksitet og gir en tilnærmet beskrivelse av hvordan maskinlæringsmodellen opptrer innenfor et begrenset område. Forklaringsmodellene kan derfor være ustabile og lite pålitelige. For enkelte inngangsvariabler kan forklaringen bli direkte feil eller misvisende.¹³⁹ I tillegg kommer at den versjonen av forklaringen som presenteres for brukeren antakelig er en forenklet utgave av approksimasjonen. Man må derfor gjøre en vurdering av hva som er et akseptabelt nivå av informasjonstap og feilmargin. Jeg ser nærmere på bruk av approksimasjoner til å forklare forvaltningsavgjørelser i avsnitt 5.4.1 og 5.4.3.

3.4.4 Forklaring som formidling og interaksjonsdesign

Som jeg sa i innledningen til dette kapittelet er forklarbarhet også noe man *tilfører* en modell gjennom formidling av forklaringer til en bruker. Forklaringene er ikke umiddelbart forståelige, men må presenteres på en måte som er tilpasset konteksten og mottakerens formål, forventninger og kompetansenivå.¹⁴⁰

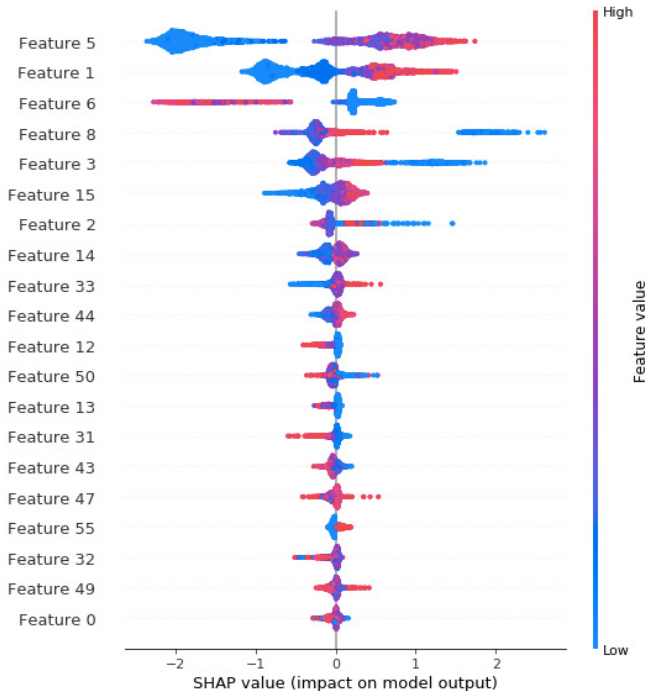
Et plott av Shapley-verdier for en enkeltprediksjon (det Beaudouin et al. kaller «raw information») kan for eksempel se ut som figur 18 nedenfor. Den er ikke

¹³⁸ Kilde: Løland (14.5.20). Det man forklarer er enkeltprediksjonen (det grønne plusstegnet) sitt avvik fra gjennomsnittet, for eksempel hvilke egenskaper som gjør at akkurat denne kunden har høyere sannsynlighet for å misligholde et lån enn gjennomsnittet.

¹³⁹ Rudin (2019) s. 208, Wachter et al. (2018) s. 48, Beaudouin et al. (2020) s. 21.

¹⁴⁰ Beaudouin et al. (2020) s. 9, The Royal Society (2019) s. 19.

intuitivt forståelig og må fortolkes og formidles til brukerne. Det kan gjøres grafisk ved hjelp av histogram, eller som tekst, og man bør velge bort de variablene som ikke har hatt særlig betydning for det samlede resultatet. Dersom det er tale om en forvaltningsavgjørelse, er det et rettslig spørsmål hvordan og med hvilket detaljeringsnivå forklaringen formidles. Her må man blant annet se hen til kravene i forvaltningsloven § 25 og til klart språk (se nærmere om dette i avsnitt 5.4.3).



Figur 18: Eksempel på plott av Shapley-verdier, som viser hvordan variablene har virket inn på enkeltprediksjonen.¹⁴¹

For å illustrere i mer detalj hva forklarbarhet som formidling kan innebære, vil jeg beskrive et av prosjektene til NAV, som er en modell for prediksjon av lengden på sykefravær. Den er ikke implementert foreløpig, men er i pilot.

¹⁴¹ Kilde: <https://datascience.stackexchange.com/questions/65307/how-to-interpret-shapley-value-plot-for-a-model>, oppslag 31.7.20.

3.4.5 Eksempel fra NAV: Forklaring av prediksjoner på sykefraværets lengde

Ved lengre sykefravær kaller NAV til dialogmøte 2 senest innen 26 uker. På møtet deltar den sykmeldte, arbeidsgiveren, NAV og eventuelt sykmeldende lege. Formålet med møtet er å legge en plan for det videre løpet og se på om det er mulig eller nødvendig med ytterligere tilrettelegging på arbeidsplassen.¹⁴²

I mange tilfeller faller behovet for møtet bort fordi den sykmeldte blir friskmeldt, og ressurser og tid går bort til unødvendig planlegging eller gjennomføring av møter. Maskinlæringsmodellen skal derfor brukes til å forutsi behovet i et gitt sykmeldingstilfelle, og den er trent på historiske data om sykmeldingers lengde. Prediksjonene gir beslutningsstøtte til veilederne som avgjør om de skal kalle inn til dialogmøte 2.

Hvis prediksjonen sier at brukeren mest sannsynlig vil være sykmeldt utover uke 28, taler det for at møtet blir nødvendig. Dersom prediksjonen derimot indikerer at personen vil ha et kortere sykmeldingsforløp kan veilederen kanskje avvente innkallingen. Meningen er at veilederen skal bruke sitt faglige skjønn i den endelige avgjørelsen, men at avgjørelsens fundament styrkes gjennom kunnskapen om sykefravær som finnes i dataene (se også figur 5 i avsnitt 2.6).

Modellen er todelt. Den består av et nevralt nett som gjør fritekstanalyse av opplysninger på legens sykmelding og en kompleks modell som analyserer tabulære data fra sykmeldinger.¹⁴³ Oppå sykmeldingsmodellen legges en forklaringsmodell som skal forklare enkeltprediksjonen. Den lokale forklaringen gir en approksimasjon av hvilke faktorer som gjorde at akkurat denne personen 'x' sin sykmelding blir predikert til varighet 'y'.¹⁴⁴

Utviklingsteamet har jobbet med forklarbarhet ved å ha med seg interaksjonsdesignere¹⁴⁵ og konsultere veiledere i en brukergruppe. Tilbakemeldingene deres ble brukt til å videreutvikle løsningen, i det NAV kaller en brukersentrisk og iterativ tilnærming. I den første versjonen av forklaring fikk veilederne kun vite at «Det er stor sannsynlighet for friskmelding innen 28 uker», vist ved en tidslinje (grønn linje i figur 19). Tilbakemeldingen fra veilederne var at denne for-

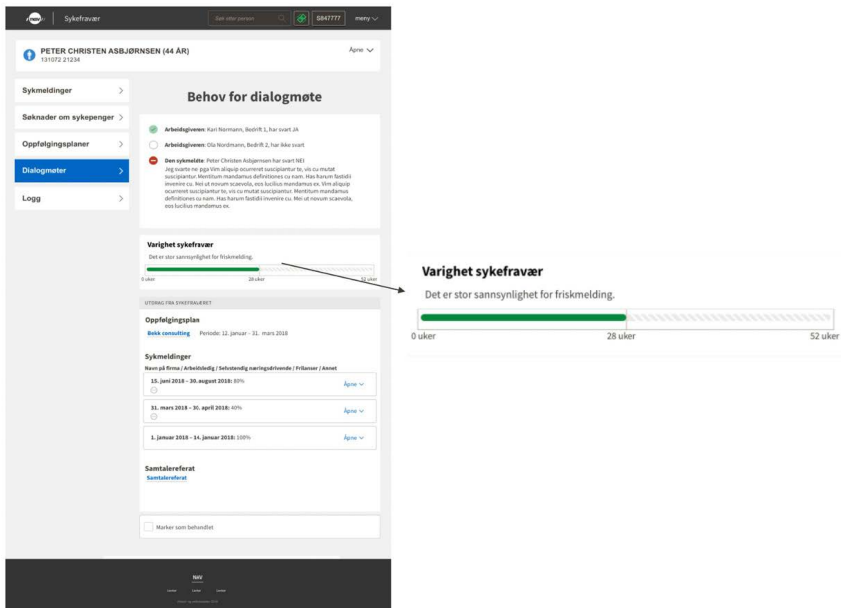
142 <https://www.nav.no/no/bedrift/oppfolging/sykmeldt-arbeidstaker/relatert-informasjon/slik-folger-du-opp-sykmeldte/>, oppslag 2.3.2020.

143 Det er tale om en *gradient boost* modell, som vil si at den er satt sammen av mange beslutningsstrær.

144 Sutterud og Topper (samtale 3.9.20).

145 Interaksjonsdesignere jobber med utforming av brukergrensesnittet til en systemløsning, med sikte på å få til gode brukeropplevelser.

klaringen ikke ga dem nok informasjon til at de kunne vurdere prediksjonen i lys av sin erfaring.¹⁴⁶

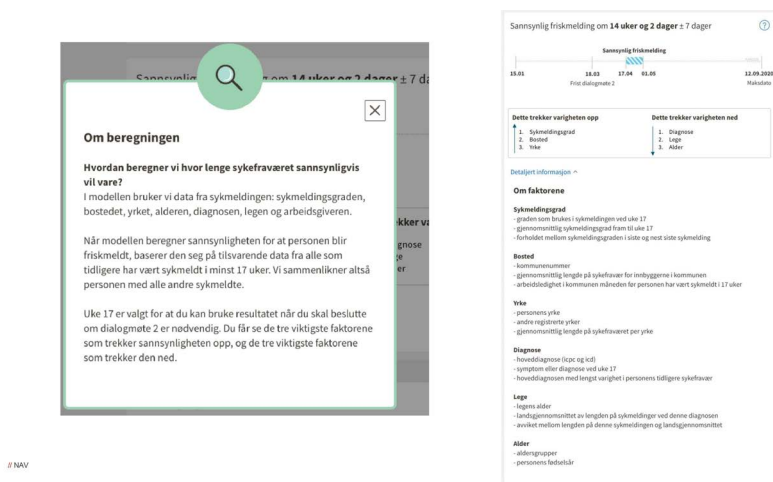


Figur 19: Forklaring versjon 1 (Bilde av veiledernes grensesnitt).¹⁴⁷

I neste versjon (figur 20) hadde forklaringen flere elementer: en sannsynlighet for friskmelding med en beregning av usikkerhet, hvilke faktorer som hadde trukket prediksjonen opp eller ned og en mulighet for å klikke seg inn på mer utdypende informasjon om faktorene og beregningene. Med denne varianten forsøkte de å dekke flere av behovene som veilederne meldte inn.

¹⁴⁶ Lyngstad (14.5.20).

¹⁴⁷ Kilde: Lyngstad (14.5.20)



Figur 20: Forklaring versjon 2, med utvidede forklaringsmuligheter.¹⁴⁸

Forklaringen av enkeltprediksjonen er basert på den beskrevne SHAP-metoden for approksimasjoner. Den er presentert ved å oppgi de tre faktorene som har betydd mest for prediksjonen i begge retninger (under «Dette trekker varigheten opp/ned» i figuren). Hver faktor er satt sammen av flere variabler, og ved å trykke på «Detaljert informasjon» får veilederen en spesifikasjon av dem. Under «Om beregningen» får veilederen opplysninger om grunnlaget for modellens beregninger og hvorfor uke 17 er valgt som tidspunkt for modelleringen.

Den lokale forklaringen blir på denne måten supplert med en form for global forklaring. Den er ikke veldig teknisk orientert, den sier for eksempel ingenting om hva slags algoritmer som er brukt eller hvordan de virker. Forklaringen er forsøkt tilpasset brukergruppens ønsker og behov.

Om dette er en tilstrekkelig forklaring for veilederne, og om det er en egnet fremgangsmåte for å komme frem til forklaringenes innhold er også rettslige spørsmål. Det samme er vurderingen av om modellen er et egnet verktøy som støtte for veiledernes beslutninger. Jeg kommer som nevnt tilbake til disse spørsmålene i kapittel 5.

Eksempelen viser hvordan forklaringer er kontekst- og behovsavhengige, og at brukere kan medvirke i hvordan forklaringer og begrunnelser utformes. «Bru-

¹⁴⁸ Kilde: Lyngstad (14.5.2020).

kere» vil ikke være begrenset til NAVs veiledere som i dette eksempelet, men alle aktører som er koblet til en forvaltningsprosess. Det kan være NAVs brukere (partene i saker om tjenester og ytelser), offentligheten, klageorgan, domstoler eller tilsynsorganer.

Dersom komplekse maskinlæringsmodeller blir tatt i bruk i saksbehandlingen vil formidling av forklaringer komme til å bli en viktig bestanddel i å sikre åpenhet, rettssikkerhet og kontroll med algoritmiske beslutningssystemer. Det kan innebære at forvaltningen må nærme seg kommunikasjonen med offentligheten og partene i en forvaltningssaker på en annerledes måte, der ulike visuelle virkemidler, brukerdialo og interaksjonsdesign får større betydning. Jeg kommer tilbake til dette poenget i avsnitt 6.3.

3.4.6 Forholdet mellom transparens og forklarbarhet

Transparens står i et beslektet, men noe uavklart forhold til forklarbarhet. Noen forfattere definerer det som tilgjengelig teknisk dokumentasjon om systemets design og utvikling. Informasjonen må ikke være publisert eller bearbeidet. Den må bare være tilgjengelig, for eksempel i forbindelse med at det skal føres tilsyn.¹⁴⁹

Transparens i denne forstand er ganske passivt. Det forutsetter at det finnes dokumentasjon om systemet, men ikke at den er bearbeidet med tanke på formidling til et bredere publikum uten teknisk kompetanse. Den ville derfor egne seg dårlig til å gi ut i forbindelse med et innsynskrav. Transparens er dermed en nødvendig, men ikke tilstrekkelig, betingelse for å ansvarliggjøre forvaltningsorganet og legge til rette for demokratisk kontroll. På dette punktet er det en parallell til forholdet mellom forvaltningens passive og aktive informasjonsplikt, som jeg kommer tilbake til i avsnitt 4.2.

I andre tilnærminger er transparens et mer direkte uttrykk for ansvarlighet hos aktøren som utvikler modellen og iverksetter systemet. Larsson og Heintz anser for eksempel at forklarbarhet er modellnært og teknisk, mens transparens bør handle om ansvarliggjøring og styring (governance) av kunstig intelligenssystemer. Transparens krever en bredere tilnærming, også faglig sett, der man ser på hvordan systemene virker i en rettslig og sosial kontekst.¹⁵⁰

149 Castelluccia og Le Métayer (2019 s. 40) definerer transparens som «the availability of the ADS code with its design documentation, parameters and learning dataset». Beaudouin et al. (2020 s. 8–9) skriver tilsvarende: «Transparency does not necessarily mean that the underlying information is easily comprehensible to humans. Transparency can simply mean that the information is made available as-is».

150 Larsson og Heintz (2020) s. 7–10.

Også Annany og Crawford argumenterer for at ansvarliggjøring krever systemperspektiv, og at transparens ikke oppnås gjennom å forklare algoritmens indre logikk. De formulerer det slik:

We suggest here that rather than privileging a type of accountability that needs to look inside systems, that we instead hold systems accountable by looking across them—seeing them as sociotechnical systems that do not contain complexity but enact complexity by connecting to and intertwining with assemblages of humans and non-humans.¹⁵¹

Transparens kan da innebære at man viser maskinlæringsmodellenes plass i statens prosesser: måten de inngår i myndighetsutøvelsen og spiller sammen med andre komponenter i beslutningssystemet, som faste algoritmer og menneskelige vurderinger. Det forutsetter en mer aktivt beskrivende, formidende informasjonsvirksomhet, som vil gjøre det enklere for offentligheten å stille forvaltningsorganet til ansvar. Dette tas opp igjen i avsnitt 4.4.2.

Det mer sosio-teknisk orienterte perspektivet vil også innebære en bevissthet om hvordan beslutningssystemet vil virke i sin kontekst: Hvordan kan man forvente at ulike aktører eller brukere vil forholde seg til modellen, og hvilke forklaringsbehov har de? Avsnitt 5.4.3 og 5.5. drøfter dette videre.

Hvis man igjen bruker CORELS-modellen fra avsnitt 3.2.3 som eksempel blir det viktig å spørre seg hva som skal være modellens rettslige funksjon i avgjørelser om kausjon og prøveløslatelse. Skal den være et grunnleggende utgangspunkt for dommerens skjønnsutøvelse, eller bare et frivillig supplement? Og hvordan passer den inn i den sosiale, organisatoriske og kulturelle konteksten i domstolene der den skal brukes? Har dommerne tillit til modellen, og har de nok kunnskap om dens potensiale og begrensninger til å forstå hvordan prediksjonen skal vektlegges i avgjørelsene?

3.5 Oppsummering: Hvordan skal forvaltningen forklare algoritmiske avgjørelser?

I dette kapitlet har jeg vist at det finnes flere måter å forklare en maskinlæringsmodell på, avhengig av modellens kompleksitet og det kontekstuelle behovet for forklaring. Jeg har også presentert ulike forklaringsbegreper, som lokal og global forklaring og transparens.

¹⁵¹ Annany og Crawford (2016) s. 2. Royal Society skriver tilsvarende: «Complex processes often surround human decision-making in critical domains, and a wider environment of accountability may need to develop around the use of AI». (Royal Society 2019 s. 25, min utheving).

Det er ikke entydig hvilken informasjon som vil inngå i en lokal eller global forklaring av en maskinlæringsmodell, eller hva det vil si at et system er transparent. Det vil derfor være behov for å tilpasse, utfylle eller presisere begrepene og velge forklaringsmetoder etter konteksten man befinner seg i.¹⁵²

Når konteksten er forvaltningsvirksomhet, kan det være behov for flere typer forklaringer som supplerer hverandre, og som samlet sett bidrar til å gjøre beslutningsprosessen tilstrekkelig åpen og rettssikker. For forvaltningen blir også selve *formidlingen* av forklaringene veldig viktig, fordi det er et grunnleggende krav at borgerne skal kunne forstå forvaltningens saksbehandling og grunnlaget for en avgjørelse som rettes mot dem.

Lokale forklaringer kaster altså lys over modellens enkeltprediksjoner *ex post*, i etterkant. De trengs for å begrunne avgjørelser for parten og for saksbehandleren som skal bruke prediksjonen som beslutningsstøtte i en enkeltsak.

Globale forklaringer trengs for å gi et overordnet bilde av hvordan modellen oppfører seg. De bidrar derfor til forutberegnelighet. Globale forklaringer kan gis uavhengig av enkeltprediksjonen, for eksempel i tilknytning til et innsynskrav etter offentleglova, som informasjon i et forhåndsvarsel eller som ledd i forvaltningens veiledning. De er også viktige for internt bruk, slik at etatens ansatte (som ledere og saksbehandlere) kan forstå modellen.

Modeller som er komplekse kan man ikke gi en helt dekkende eller uttømmende global forklaring på, fordi man ikke vet på forhånd hvordan de oppfører seg for alle inngangsvariabler (over hele modellens gyldighetsområde). Testing av modellen gjør at man kan avdekke hvordan den opptrer «i snitt» eller for utvalgte grupper. Denne typen informasjon bidrar til forutberegneligheten, og kan kalles *global forklaring i snever forstand*.

Overordnede beskrivelser av hvordan modellen typisk arter seg for ulike grupper er imidlertid ikke tilstrekkelig for å trygge offentligheten og private parter på at skjevheter og svakheter i data, i modelleringen eller implementeringen ikke gir seg utslag i diskriminerende, usaklige eller vilkårlige avgjørelser. For å kunne ha en begrunnet tillit til systemet og muligheten for å ettergå og kritisere modellens forutsetninger er det nødvendig med en mer *utvidet global forklaring*, en slags brukerveiledning som Beaudouin et al. har foreslått (avsnitt 3.4.1). Den må omfatte beskrivelser av hvordan data er håndtert og hvilke skritt som er tatt

152. For eksempel skriver Rudin at «Interpretability is a domain-specific notion, so there cannot be an all-purpose definition». (Rudin, 2019 s. 6).

for å unngå at skjevheter eller svakheter realiserer seg i diskriminerende eller vilkårlige avgjørelser.¹⁵³

Modellenes kompleksitet gjør at man ikke kan forvente av den enkelte borger eller part at de fullt ut setter seg inn i systemenes forutsetninger og virkemåte. Forklarbarhet kan derfor i mange tilfeller komme til å dreie seg om å innfri et krav om dokumentasjon, som viser at forvaltningsorganet etterlever regler for utvikling, anvendelse og kontroll av beslutningssystemet. Det betyr også at forklarbarhet, systemkrav og kontroll henger nært sammen.¹⁵⁴

I tillegg er det behov for et forklaringsbegrep som er mer systemorientert, som jeg var inne på i forrige avsnitt om transparens. Muligheten for både forståelse og ansvarliggjøring krever at man ser hvordan maskinlæringsmodellen inngår i myndighetsutøvelsen sammen med andre komponenter i beslutningssystemet. Jeg er usikker på hva som vil være et dekkende begrep for dette. I noen sammenhenger kan et flytdiagram gjøre nytten, men med mulighet for mer utdypende dokumentasjon. Arkivlovutvalget bruker begrepet *systembeskrivelse*, men i en noe annen betydning (se omtale i avsnitt 4.2.2). Jeg velger allikevel å bruke dette som betegnelse på en forklaring som viser funksjonsfordelingen mellom maskinlæring, faste algoritmer og menneskelige vurderinger i et algoritmisk beslutningssystem.

I de neste to kapitlene vil jeg bruke disse begrepene når jeg ser på hvilke rettslige krav som vil gjelde til forklarbarhet av algoritmiske beslutningssystemer i velferdsforvaltning.

153 Artikkel 13, *Transparency and provision of information to users*, i EUs forslag til regulering av KI-systemer gir en pekepinn om hva som kan inngå i en slik global forklaring, og dette samsvarer godt med det som allerede er nevnt. I annet ledd står det at systemets bruker må få en lettfattelig bruksanvisning. Tredje ledd presiserer hvilken informasjon som skal inngå i denne, blant annet en beskrivelse av systemets formål, hvor treffsikkert det er, risikomomenter, treffsikkerheten for spesifikke grupper, hvilke data som inngår og informasjon om trenings- og testdata. Forslaget bruker «system» i samme betydning som jeg bruker «modell» her i kapitlet. «Users» sikter som nevnt til dem som forholder seg direkte til systemet, og ikke subjektene (partene i forvaltningssaker). (EU-kommisjonen 2021).

154 Artikkel 11 *Technical documentation* i EUs forslag til regulering av KI systemer krever at det foreligger oppdatert dokumentasjon på at systemet overholder lovgivningen, og at den gjøres tilgjengelig for blant annet nasjonale tilsynsmyndigheter. (EU-kommisjonen 2021).

4 Forklarbarhet som forutsetning for et velfungerende demokrati

4.1 Innledning

Når maskinlæringsmodeller anvendes som ledd i myndighetsutøvelse, må de kunne underkastes demokratisk kontroll. Borgerne må kunne forstå forvaltningens saksbehandling og ha mulighet for å kreve innsyn i forvaltningens systemer. Det vil derfor være et demokratisk problem dersom algoritmiske beslutningssystemer utvikles og implementeres på en måte som unndrar seg offentlighet.¹⁵⁵ Spørsmålet i dette kapitlet er dermed hvordan perspektiver på forklarbarhet og transparens kan bidra til overholdelse av demokratiske prinsipper og grunnverdier i en algoritmisk forvaltning.

I avsnitt 4.2 presenteres grunnleggende rettsregler og prinsipper som skal fremme og beskytte demokratiet. Åpenhet og forståelighet er en forutsetning for en opplyst offentlighet (deliberativt demokrati) og tilrettelegging for demokratisk deltakelse (deltakerdemokrati). Avsnitt 4.3 beskriver regler for dokumentasjon og innsyn som grunnlag for å ansvarliggjøre forvaltningen, og hvordan disse kan tilpasses automatiserte saksbehandlingssystemer. I avsnitt 4.4 drøfter jeg regler og tiltak for dokumentasjon, innsyn og medvirkning ved algoritmisk forvaltning.

4.2 Behovet for en åpen og forståelig forvaltning

4.2.1 Ytrings- og informasjonsfrihet

Grunnloven § 100 og EMK artikkel 10 forankrer ytringsfriheten og informasjonsfriheten som grunnleggende betingelser for et demokratisk samfunn. Grunnloven § 100 første og annet ledd lyder:

Ytringsfrihet bør finne sted.

Ingen kan holdes rettslig ansvarlig for å ha meddelt eller mottatt opplysninger, ideer og budskap med mindre det lar seg forsvare holdt opp imot ytringsfrihetens begrunnelse i sannhetssøken, demokrati og individets frie meningsdannelse. Det rettslige ansvar bør være foreskrevet i lov.

¹⁵⁵ Dette ser man mange eksempler på internasjonalt, som NGOen Algorithmwatch rapporterer om (<https://algorithmwatch.org/en/>).

EMK artikkel 10 nr. 1 lyder:

Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises.

Ytrings- og informasjonsfriheten bygget på en tanke om at individene trenger en viss grad av kompetanse for å være autonome borgere. Et levende demokrati forutsetter at meningsutveksling skjer i det offentlige rom, og at borgerne har den nødvendige kunnskap til å innta standpunkter og delta i en meningsutveksling. Det vil også gi et bedre beslutningsgrunnlag at ulike standpunkter kan komme til uttrykk. Dette kan kalles det kommunikative eller *deliberative* aspektet ved demokratiet.

Ytringsfrihetskommisjonen fremhevet at det borgerne trenger å forstå er «de krefter og prosesser som former de ytre betingelser for våre liv». Makten må ha et ansikt, og derfor er det viktig å ha åpenhet rundt politiske avgjørelser og forvaltningens myndighetsutøvelse. Demokrati, åpenhet og kontroll henger derfor tett sammen.¹⁵⁶

4.2.2 Offentlighetsprinsippet og den passive informasjonsplikten

Etter Grunnloven § 100 femte ledd har enhver «rett til innsyn i statens og kommunenes dokumenter». Begrensninger i denne retten kan bare fastsettes i lov ut fra «hensyn til personvern og av andre tungtveiende grunner». Bestemmelsen er et uttrykk for offentlighetsprinsippet. Den omtales gjerne som den passive informasjonsplikten fordi realiseringen av den beror på at borgerne ber om innsyn.¹⁵⁷

Nærmere regler om offentlighet for forvaltningens dokumenter er gitt i offentleglova. Loven skal bidra til å realisere Grunnloven § 100. Det fremgår av formålsbestemmelsen i § 1:

Formålet med lova er å leggje til rette for at offentleg verksemd er open og gjennomiktig, for slik å styrkje informasjons- og ytringsfridommen, den demokratiske deltakinga, rettstryggleiken for den enkelte, tilliten til det offentlege og kontrollen frå ålmenta.

Hovedregelen om innsynsrett er gitt i offentleglova § 3:

¹⁵⁶ NOU 1999:27 s. 22–23 og 95.

¹⁵⁷ NOU 2019: 5 s. 243.

Saksdokument, journalar og liknande register for organet er opne for innsyn dersom ikkje anna følgjer av lov eller forskrift med heimel i lov. Alle kan krevje innsyn i saksdokument, journalar og liknande register til organet hos vedkommande organ.

Bestemmelsen er formulert som en regel om tilgang til informasjon. Det følger av ordlyden at retten til innsyn knytter seg til eksisterende dokumenter, slik at det ikke er noen plikt for forvaltningsorganet å opprette et nytt dokument for å besvare et innsynskrav eller bearbeide informasjonen før den gis ut.¹⁵⁸ Et gjennomgående hensyn i forarbeidene til offentleglova er at forvaltningen i minst mulig grad skal belastes med merarbeid i forbindelse med innsynskrav.¹⁵⁹

Hensynet bak Grunnloven § 100 og offentleglovas formålsbestemmelse tilsier at det må kunne kreves at informasjonen som gjøres tilgjengelig er forståelig for den alminnelige borger. Spørsmålet om mulig behov for tilpasning av informasjonen ser ikke ut til å være særskilt problematisert i forarbeidene til ytringsfrihetsbestemmelsen i Grunnloven. Ytringsfrihetskommisjonen setter et likhetstegn mellom tilgang til informasjon (transparens eller gjennomskiktighet) og muligheten for å forstå prosesser i samfunnet:

En slik gjennomskiktighet er en forutsetning for at vi skal kunne forholde oss til disse krefter på en fornuftig måte, og dermed kunne nyte den karakteristiske frihetsfølelse det bør gi å leve i et demokrati.¹⁶⁰

Den passive innsynsretten må imidlertid sees i sammenheng med den aktive informasjonsplikten, som forutsetter at forvaltningen inntar en mer forklarende og formidlende rolle.

4.2.3 Aktiv informasjonsplikt

Statens aktive informasjonsforpliktelse følger av Grunnloven § 100 sjette ledd. Her står det: «Det påligger statens myndigheter å legge forholdene til rette for en åpen og opplyst offentlig samtale».

Den aktive informasjonsplikten supplerer den passive informasjonsplikten i femte ledd. Hensikten er å oppfordre til at staten aktivt går ut med informasjon om sider ved sin virksomhet som berører innbyggerne, som Ytringsfrihetskommisjonen beskriver det:

158 Et unntak fra det er § 9 om innsyn i databaser, hvor et innsynskrav kan innebære «ei samstilling av opplysningar som er elektronisk lagra i databasane til organet».

159 For eksempel: NOU 2003:30 s. 134, 217 og 233.

160 NOU 1999:27 s. 23.

For å bidra til kvalitet og mangfold i den offentlige debatt vil staten i mange sammenhenger være forpliktet til en aktiv informasjonsvirksomhet, både om egen og private virksomhet. Det innebærer at staten av eget tiltak må gå ut med informasjon til allmennheten, eller sørge for at andre rettssubjekt pålegges en aktiv informasjonsplikt.¹⁶¹

Plikten er imidlertid ikke presisert i lovgivning. Forvaltningslovutvalget har vurdert om en bestemmelse om generell aktiv informasjonsplikt bør lovfestes i forvaltningsloven. Utvalget fremhever som en fordel at det kan «bidra til å aktualisere den passive informasjonsplikten»¹⁶². Man kan gjøre folk oppmerksomme på forhold det kan være interessant å undersøke nærmere og be om innsyn i, som de ellers ikke ville fått kjennskap til. Utvalget foreslår ikke en slik bestemmelse, fordi utvalget vurderer det slik at den ville blitt for vag og veiledningsplikten langt på vei vil ivareta informasjonsbehovet.¹⁶³

En viss konkretisering av informasjonsplikten finnes i veilederen Statens kommunikasjonspolitikk. Veilederen gir ikke borgerne noe rettskrav på informasjon, men sier hvilke prinsipper som bør være førende for forvaltningsorganenes informasjon til offentligheten. Målet med den er at innbyggerne skal få «korrekt og klar informasjon om sine rettigheter, plikter og muligheter, ha tilgang til informasjon om statens virksomhet og inviteres til å delta i utformingen av politikk, ordninger og tjenester».¹⁶⁴

Informasjonen skal være tilpasset målgruppen og formulert i et klart språk. Den skal også nå alle berørte. For å innfri det kan det noen ganger være tilstrekkelig å gjøre informasjon tilgjengelig, mens det i andre sammenhenger kan være nødvendig å gå mer målrettet til verks. Spesielt gjelder det dersom informasjonen angår innbyggernes plikter og rettigheter. Staten skal gå aktivt ut med informasjon fremfor å vente på eventuelle innsynskrav der det er en forutsetning for at innbyggerne skal kunne delta aktivt i demokratiet.¹⁶⁵

4.2.4 Deltakerdemokrati og brukermedvirkning

Nært forbundet med det deliberative demokratiet er idealet om *deltakerdemokratiet*, som omtales i en rekke offentlige utredninger og dokumenter. Tanken er at borgerne skal medvirke til den politiske prosessen utover det å avgi sin stemme ved valg, og at det kan styrke demokratiet og tilliten i samfunnet.¹⁶⁶

161 NOU 1999:27 s. 250.

162 NOU 2019: 5 s. 245.

163 NOU 2019: 5 s. 245–246.

164 Fornyings- og administrasjonsdepartementet (2009).

165 Fornyings- og administrasjonsdepartementet (2009).

166 NOU 2006: 7, NOU 2003: 19 (Makt- og demokratiutredningen), St.meld. nr. 19 (2008–2009) (Forvaltningsmeldingen).

Ifølge Dahl kan man snakke om bredde og dybde i deltakelsen. Bredde viser til at flere trekkes inn i beslutningsprosesser, at alle interesser er representert og at alle hensyn kommer frem. Det kan øke forståelsen for de demokratiske prosessene og det aktuelle forvaltningsområdet. Dybden handler om det kvalitative: at man sikrer bedre beslutningsgrunnlag, mer rettferdig avveining mellom ulike gruppers interesser og behov og at beslutningene faktisk gjennomføres. Avgjørelsene kan dermed bli bedre utredet og få større legitimitet.¹⁶⁷

I forvaltningslovgivningen finnes det eksempler på bestemmelser som krever eller oppfordrer til medvirkning og høring. Et eksempel er plan- og bygningsloven § 5-1 første ledd, hvor det står at den som fremmer et planforslag skal legge til rette for medvirkning. Etter annet ledd har kommunen et spesielt ansvar for å legge til rette for at alle grupper og interesser får mulighet til å delta eller bli hørt.

I forlengelse av deltakerdemokrati finner man også ulike varianter av brukermedvirkning. Dette er særlig relevant for utforming og gjennomføring av kommunal tjenesteyting. Et eksempel fra velferdsretten er pasient- og brukerrettighetsloven § 3-1 om «pasientens eller brukerens rett til medvirkning». Her står det blant annet at tjenestetilbudet så langt det er mulig skal utformes i samarbeid med pasienten eller brukeren, og at medvirkningen «skal tilpasses den enkeltes evne til å gi og motta informasjon».

En bestemmelse om konsultasjon finnes dessuten i GDPR artikkel 35 om personvernkonsekvensvurdering. I 35 (9) står det at: «Dersom det er relevant skal den behandlingsansvarlige innhente synspunkter på den planlagte behandlingen fra de registrerte eller deres representanter (...)». Eksempler på mulige metoder for innhenting kan være brukerpanel og brukerundersøkelser. Hva som er en egnet fremgangsmåte, vil avhenge av konteksten.¹⁶⁸

4.2.5 Krav til klart språk

Ny lov om språk (språklova) trer i kraft 1.1.2022. Den har en egen bestemmelse om klarspråk i § 9:

Offentlege organ skal kommunisere på eit klart og korrekt språk som er tilpassa målgruppa.

Formålet er at borgerne skal kunne finne, forstå og nyttiggjøre seg informasjonen de får. «Språk» viser ifølge forarbeidene ikke bare til tekst, men også «inn-

¹⁶⁷ Dahl gjengitt i Engberg (2003) s. 64, NOU 2006: 7 kap. 4.1.3, NOU 2003: 19 kap. 2.3.

¹⁶⁸ Artikkel 29-arbeidsgruppen (2017b).

hald, struktur og visuelle element».¹⁶⁹ Det vil si at alle virkemidler eller teknikker for å formidle et budskap vil være omfattet av kravet til klart språk.

Forvaltningslovutvalget har også foreslått en bestemmelse om klart språk i utkastets § 8 annet ledd, som sier at forvaltningsorganet skal bruke et språk som er «klart, presist og så vidt mulig tilpasset mottakeren». Utvalget ser klart språk som en forutsetning for å oppfylle forpliktelser til veiledning, utredning og begrunnelse (fvl §§ 11, 17 og 25), og for tilliten mellom borger og forvaltning.¹⁷⁰

Med disse bestemmelsene legges et tydelig ansvar på forvaltningen i å uttrykke seg forståelig. I begge bestemmelsene fremheves forvaltningens plikt til å tilpasse kommunikasjonen. Hva som skal regnes som tilstrekkelig «klart» vil bero på konteksten og borgerens eller partens informasjons- og forklaringsbehov.

4.3 Dokumentasjon og innsyn i digital forvaltning

Et fremtredende trekk ved den digitaliserte forvaltningen er bruken av automatiserte saksbehandlingssystemer. Automatisert rettsanvendelse reiser noen spørsmål om løsninger for dokumentasjon og innsyn, som jeg drøfter i det følgende.

4.3.1 Dokumentbegrepet og arkivplikten

Saksdokumenter er den grunnleggende informasjonsenhet som innsyns- og arkivregler er koblet til. «Saksdokument» er definert i offentleglova § 4 første ledd som «en logisk avgrenset informasjonsmengde som er lagret på et medium for senere lesing, lytting, fremvisning eller overføring».

Dokumentbegrepet er teknologinøytralt, og uavhengig av lagringsmedium og format. Det avgjørende er at informasjonen kan identifiseres og gjenfinnes. At det er et saksdokument for organet betyr etter § 4 annet ledd at det har kommet inn til eller er opprettet av organet selv og at det angår organets virksomhet. Dokumentet anses som opprettet dersom organet har sendt det fra seg eller det er ferdigstilt.

Forvaltningens informasjon må lagres og systematiseres for at den skal være tilgjengelig for samfunnet. Det betyr at arkivlovgivningen også er et viktig mid-

¹⁶⁹ Prop. 108 L (2019–2020) s. 119.

¹⁷⁰ NOU 2019: 5 s. 177.

del for å realisere formålet om en åpen forvaltning og en opplyst offentlig sam-
tale.¹⁷¹

Arkivlova § 6 etablerer arkivplikten for offentlige organer. Bestemmelsen lyder:

Offentlege organ pliktar å ha arkiv, og desse skal vera ordna og innretta slik at doku-
menta er tryggja som informasjonskjelder for samtid og ettertid.

Bestemmelsen er lite spesifikk, men forarbeidene understreker at det er tale om
et *toleddet informasjonsbegrep* for arkiv. Det er både dokumentene som sådan og
dokumentene sett i sammenheng som har informasjonsverdi:

Dokumentet inngår vanlegvis i ei lang kjede av dokument, og kvart einskilt dokument
gjev som regel berre avgrensa informasjon. Derimot kan kjeda av dokument sett i
samanheng gje både utfyllande informasjon om den konkrete saka som er behandla,
og informasjon om korleis saka er vorten til. Såleis kan ein tala om eit toleda informas-
jonsomgrep for arkiv; for det første den informasjonen kvart einskilt dokument kan
gje, og for det andre den informasjonen arkivet samla kan gje om både saka, saksføre-
buinga og grunnlaget for avgjerda.¹⁷²

Når man tar vare på dokumentene og kjeden de inngår i, oppnås sporbarhet og
etterprøvbarehet for forvaltningens avgjørelser. Det gir mulighet for å forstå for-
valtningens prosesser og kontrollere myndighetsutøvelsen.

Arkivlovutvalget peker på at den digitale transformasjonen i offentlig forvalt-
ning setter andre krav til dokumentasjon av virksomheten:

Mye informasjon oppstår og oppbevares i fagsystemer, databaser og programkode som
ikke passer med en naturlig språklig forståelse av ordet dokument. Denne utviklingen
gjør det ikke mindre viktig at forvaltningen dokumenterer sin virksomhet og system-
atiserer og lagrer dokumentasjonen slik at den kan gjøres tilgjengelig i samtid og etter-
tid.¹⁷³

Forvaltningsprosesser har ikke alltid de samme naturlige dokumentkjeder som
arkivlovens forarbeider omtaler. Tradisjonelt kunne man altså få den nødven-
dige prosess- og kontekstinformasjon ved å se i saksmappen på dokumentasjo-
nen av transaksjonene i saken. Det får man ikke i en digital prosess hvor etatene
baserer seg på å dele data og hvor informasjonen bare gir mening der og da, i

171 NOU 2019: 9 s. 47 og 87.

172 Ot.prp. nr. 77 (1991–92) s. 6.

173 NOU 2019: 9 s. 88.

konteksten hvor dataene inngår. En søknad eller et vedtak kan for eksempel basere seg på data som er lagret forskjellige steder, i ulike databaser eller hos ulike etater.¹⁷⁴ Arkivbehovet er det samme, men det kan tilsi at man trenger andre løsninger for å sikre at informasjonen er tilgjengelig og meningsfull.

Spørsmålet er derfor hvordan man gitt nye digitale forvaltningsprosesser kan sikre at informasjonen både er tilgjengelig og at den er meningsfull for offentligheten. For å svare på det er det først nødvendig å si noe om egenskapene ved den digitale informasjonen for deretter å vurdere hvordan den kan dokumenteres.

4.3.2 Dokumentasjon av automatiserte saksbehandlingssystemer

Når rettsanvendelsen skal automatiseres ved hjelp av faste algoritmer, transformeres lovt tekst til programkode i en prosess som kalles systemdrevet fortolkning. Systemdrevet fortolkning innebærer at man tar stilling til rettsspørsmål på forhånd, før systemet implementeres og sakene tas til behandling. Dette står i motsetning til den saksdrevne fortolkningen som skjer i manuell enkeltsaksbehandling.¹⁷⁵

Ved bruk av slike automatiserte beslutningssystemer vil det ikke være tilstrekkelig å dokumentere enkeltsaken, fordi det ikke gir den nødvendige kontekst og bakgrunn for å forstå og kontrollere rettsanvendelsen. For å få et dekkende bilde av «saka, saksførebuinga og grunnlaget for avgjerda»¹⁷⁶, er det nødvendig med dokumentasjon som viser de forutsetninger og tolkningsvalg som er tatt av systemets utviklere. Schartum kaller dette for dokumentasjon av beslutningssystemets rettslige innhold, og definerer det som «skriftlig materiale som blir til som del av utviklingsarbeidet, og som er egnet til å gjengi og forklare de av systemets egenskaper og funksjoner som er rettslig relevante». Forvaltningslovutvalgets betegnelse på det samme er «rettslig kravspesifikasjon».¹⁷⁷

Det finnes imidlertid ikke noen uttrykkelig lovbestemmelse som forplikter et forvaltningsorgan til å dokumentere et beslutningssystems rettslige innhold. Schartum påpeker at det å lage systemløsninger uten at det er tilgjengelig forståelig dokumentasjon om det rettslige innholdet vil være i strid med prinsipper om forsvarlig saksutredning og saksbehandling. Han understreker likevel at

174 NOU 2019: 9 s. 153–155, 159.

175 Schartum (2018) s. 276–278.

176 Ot.prp. nr. 77 (1991–92) s. 6.

177 Schartum (2018) s. 276–278, NOU 2019: 5 s. 259.

mangelen på tydelige bestemmelser kan være et hinder for åpenhet i forvaltningen og borgernes rettssikkerhet.¹⁷⁸

Forvaltningslovutvalget og Arkivlovutvalget har foreslått bestemmelser om dokumentasjon av automatiserte og delautomatiserte saksbehandlingssystemer, som er ment å ivareta dette. Forvaltningslovutvalgets utkast § 12 ledd lyder:

Forvaltningsorganet skal dokumentere det rettslige innholdet i automatiserte saksbehandlingssystemer. Dokumentasjonen skal offentliggjøres, hvis ikke annet følger av lov eller særlige hensyn taler mot det.¹⁷⁹

Arkivlovutvalgets forslag til bestemmelse gir en mer presis beskrivelse av hva som skal dokumenteres. Arkivlovutvalgets utkast § 10 lyder:

(1) Virksomheter som helt eller delvis automatiserer rettsanvendelse, skal dokumentere

- a) hvilke datatyper som anvendes
- b) hvilke kilder som benyttes for disse datatypene
- c) hvilke behandlingsregler som er utledet av rettsreglene og som er styrende for vedtakene, og
- d) hvilke endringer som er gjort av systemet på bakgrunn av vedtak av lover og forskrifter og andre politiske vedtak.

(2) Kongen kan pålegge at det utarbeides dokumentasjon som gir grunnlag for rekonstruksjon av kjørbare programkode, etter at systemet er tatt ut av drift.

Forvaltningslovutvalgets forslag er primært rettet mot å dokumentere transformasjonsprosessen fra lovtekst til kode i faste algoritmer, mens Arkivlovutvalgets bestemmelse er videre og mer fleksibel med hensyn til beslutningssystemets innretning. Bokstav c) i Arkivlovutvalgets forslag er motsatsen til rettslig kravspesifikasjon i Forvaltningslovutvalgets forslag. I tillegg kreves det dokumentasjon av hvilke data som inngår og hvor de er hentet fra, systemendringer og informasjon som gjør det mulig å rekonstruere vedtak i ettertid. Denne bestemmelsen vil derfor også være mer egnet for systemer hvor også maskinlæring inngår. § 10 må dessuten sees i sammenheng med § 9, som sier at det må gis systembeskrivelser for virksomhetens informasjonssystemer. Systembeskrivelsen skal inneholde den dokumentasjon som er nødvendig for at systemet skal kunne driftes på en betryggende måte. Det vil for eksempel være «bruksrutiner, datamodell og systemegenskaper» (utkastets § 9 første ledd).

178 Schartum (2015) s. 202. Han skriver også: «Dersom myndighetsutøvelsen kapsler seg inn i beslutningssystemer, kan dette hevdes å være en trussel mot åpenheten i offentlig forvaltning og dermed også mot folks mulighet for å ivareta egne interesser». (Schartum, 2018 s. 279).

179 NOU 2019: 5 s. 20.

4.3.3 Formidling og forståelighet

Automatisering gjør også spørsmålet om forståelighet og formidling av dokumentasjonen mer påtrengende. Å legge frem programkode som svar på et innsynskrav etter offentleglova § 3 vil gi liten mening for de fleste. Det kan også være problematisk med tanke på kravet til klarspråk (kommende språklav § 9).

Forvaltningslovutvalget legger til grunn at den rettslige kravspesifikasjonen som pålegges etter § 12 vil være i pseudokode¹⁸⁰ eller annet «semi-formelt uttrykk», og at den som utgangspunkt vil være lesbar, forståelig og presis nok til å gi inntrykk av hvordan systemet virker. I enkelte tilfeller kan det likevel være nødvendig å supplere med beskrivelser i naturlig språk. Ettersom dokumentasjonen skal offentliggjøres vil det dessuten overflødiggjøre en del innsynskrav.¹⁸¹

Arkivlovutvalget skriver også at dokumentasjonen må utformes på en forståelig måte, og kan for eksempel formidles som pseudokode, brukerhistorier, flytdiagrammer eller prosessbeskrivelser.¹⁸²

Systemoffentlighet kan være et supplement til innsynsrett, som kan være en god måte å formidle et systems virkemåte til borgerne. Med systemoffentlighet sikter Schartum et al. til at forvaltningen offentliggjør beslutningssystemet med programkoden som uttrykker gjeldende rettsregler. En variant av det er ulike former for «kalkulatorer», hvor brukerne kan simulere rettsanvendelsen og se hva utfallet i saken vil bli. Som et eksempel nevner de Lånekassens nedbetalingskalkulator.¹⁸³

4.3.4 Oppsummering

En rett til en åpen og forståelig forvaltning følger av Grunnloven, offentlighetsprinsippet og idealene om det deliberative og deltakende demokrati. Den passive informasjonsplikten i offentleglova § 3 medfører et rettskrav på informasjon til borgeren. Det kreves imidlertid en viss kompetanse og aktiv interesse hos borgerne for å formulere og spesifisere innsynskrav. Innsynsretten gjelder dessuten for eksisterende saksdokumenter, slik at det i utgangspunktet ikke er noen plikt for forvaltningsorganet til å tilpasse eller forklare informasjonen i forbindelse med et innsynskrav.

Dersom det er uklart hvilken dokumentasjon som finnes, og den bare finnes i et teknisk og lite forståelig format, vil ikke innsynsretten nødvendigvis være til-

180 Pseudokode er en beskrivelse av programmeringen i et språk som nærmer seg naturlig norsk og som kan forstås av personer som ikke er programmeringskyndige.

181 NOU 2019: 5 s. 266–267, Schartum et al. (2017) s. 102.

182 NOU 2019: 9 s. 274.

183 Schartum et al. (2017) s. 103.

strekkelig for å oppfylle formålet om en åpen og kritisk offentlig samtale om forvaltningens virksomhet og systemer. Dette er en problemstilling som gjør seg gjeldende i forbindelse med digital forvaltning og automatisert saksbehandling, og som også er aktuell ved eventuell implementering av algoritmiske beslutningssystemer i forvaltningen.

Forvaltningslovutvalget og Arkivlovutvalgets utkast til bestemmelser om dokumentasjon av automatiserte saksbehandlingssystemer gir bedre forutberegnelighet og etterprøvbarehet enn det som følger av gjeldende lovbestemmelser, men Arkivlovutvalgets forslag er det som i størst grad kan bidra til at systemene dokumenteres og at dokumentasjonen blir tilgjengelig for offentligheten i en forståelig form. Jeg kommer mer inn på dette i avsnitt 4.4.1.

Den passive informasjonsplikten suppleres av den aktive informasjonsplikten, som kan bidra til å gjøre kjent temaer eller dokumenter det kan være aktuelt å kreve innsyn i. Den aktive informasjonsplikten er imidlertid lite konkretisert utover Grunnloven § 100 sjette ledd. Det enkelte forvaltningsorgan velger i stor grad hvordan den skal oppfylles, og det blir et informasjonsfaglig spørsmål der statens informasjonspolitikk er retningsgivende. Systemoffentlighet kan imidlertid være en hensiktsmessig måte å oppfylle den aktive informasjonsplikten. Schartum fremhever det som et godt supplement til dokumentoffentlighet.¹⁸⁴

Jeg vil nå gå mer konkret inn på innsyn og medvirkning ved algoritmisk forvaltning.

4.4 Hvordan kan innsyn og medvirkning sikres i algoritmiske beslutningssystemer?

Algoritmisk forvaltning kan utfordre betingelsene for det deliberative og deltakende demokratiet. Én utfordring er relatert til algoritmiske beslutningssystemers synlighet og tilgjengelighet. Spørsmålet er om offentligheten får kjennskap til systemene, og om innsynsretten er praktisk gjennomførbar og gir borgerne forståelig informasjon de kan nyttiggjøre seg. I avsnitt 4.4.1 ser jeg på hvordan algoritmiske beslutningssystemer bør dokumenteres, mens avsnitt 4.4.2 behandler formidling av dokumentasjon til offentligheten, enten i forbindelse med innsynskrav eller utøvelse av den aktive informasjonsplikten.

En annen utfordring dreier seg om vurderinger, forutsetninger og antakelser som nedfelles i algoritmene. En maskinlæringsmodell som fra utsiden kan

184 Schartum (Innlegg på eForvaltningskonferansen 15.3.21).

fremstå som objektiv og basert på harde fakta utledet av data, er i realiteten en refleksjon av normer, verdier og politikk som har gjort seg gjeldende under systemets design og utvikling.¹⁸⁵ Det er problematisk hvis politiske veivalg og verddivurderinger lukkes inne i et beslutningssystem uten at problemstillingene kan gjøres til gjenstand for offentlig debatt. Dette er ikke noe som vil løses gjennom innsynsrettigheter, men ved å legge til rette for åpne og deltakende prosesser med interesserte borgere og berørte brukere. Avsnitt 4.4.3 demonstrerer nytten og behovet for medvirkning (deltakelse) gjennom et eksempel, og viser til aktuelle rettslige grunnlag for medvirkning.

4.4.1 Modellene og beslutningssystemene må dokumenteres

Gjeldende lovgivning er som nevnt under avsnitt 4.3.2 lite konkret med hensyn til dokumentasjonskrav av automatiserte og delautomatiserte saksbehandlingssystemer. Fravær av tydelige bestemmelser er et problem fordi det hindrer allmennoffentlighet og ansvarliggjøring og kontroll av forvaltningen. Det kan for eksempel gjøre det vanskelig for borgerne å formulere og konkretisere innsynskrav.

Poenget kan illustreres med Kauns analyse av to innsynskrav i et automatisert saksbehandlingssystem for sosialhjelp i Trelleborg kommune i Sverige.¹⁸⁶ Studien har en viss overføringsverdi til norske forhold. I det første eksempelet Kaun omtaler, ba en journalist om innsyn i saksbehandlingssystemets kildekode. Kommunen avslo under henvisning til at algoritmen var et «produkt» og ikke et offentlig dokument i henhold til Tryckfrihetsförordningen. Journalisten klaget saken inn for forvaltningsdomstolen, som avgjorde at det skulle gis innsyn. I begrunnelsen var domstolen tilbakeholden med å si at kildekode i utgangspunktet omfattes av definisjonen på offentlig dokument. I stedet ble det lagt vekt på at koden tilhørte kommunen og at ingen part ville være skadelidende om det ble gitt innsyn. Til slutt sendte konsultentselskapet som hadde utviklet systemet kildekode (men klarte samtidig å inkludere personsensitive opplysninger om Trelleborgs sosialhjelpsmottakere).¹⁸⁷

Det andre eksempelet gjaldt en fagforening for offentlig ansatte som ikke spesifiserte innsynskravet utover å si at de ønsket «meningsfull informasjon» om hvordan systemet fungerte. Da utviklet innsynskravet seg til en slags forhandling med kommunen, og etter flere møter ble de enige om at det fagforeningen

¹⁸⁵ Allhutter et al. (2020).

¹⁸⁶ Systemet er basert på faste algoritmer, ikke maskinlæring. Det er en enkel beslutningstremodell som sjekker opplysninger mot bl.a. Skatteverket. Systemet behandler fornyelser, ikke første gangs søknader. (Kaun, 2021).

¹⁸⁷ Kaun (2021) s. 10. Det er lite sannsynlig at det ville blitt konklusjonen i en norsk domstol, med tanke på det vidt formulerte og teknologinøytrale dokumentbegrepet i offentliglova og arkivlova.

hadde behov for var et skjermbilde av saksbehandlernes brukergrensesnitt. Kommunen fulgte imidlertid ikke opp, og da det var gått ett år klaget fagforeningen til Justititeombudsmannen. Kommunen valgte til slutt å besvare innsynskravet med et skjermbilde av algoritmens beslutningstrær.¹⁸⁸

Studien viser at det kan være avstand mellom offentlighetsprinsippets intensjon og den praktiske virkeligheten. Uklare regler og rutiner koblet med dårlig systemforståelse hos både forvaltning, publikum og domstoler kan medføre tungvinte prosesser som ikke nødvendigvis fører frem.

Jeg vil derfor se nærmere på hvordan algoritmiske beslutningssystemer bør dokumenteres med sikte på ansvarliggjøring, sporbarhet og borgernes mulighet for å utøve innsynsrettigheter. Jeg tar utgangspunkt i henholdsvis systemdokumentasjon (beskrevet i avsnitt 2.4.1) og systembeskrivelser (introdusert i avsnitt 3.5).

Systemdokumentasjon er grunnlaget for å oppfylle arkivplikten og informasjonsplikten, og for at en ekstern aktør skal kunne forstå og ettergå modellen. Systemdokumentasjon kan omfatte informasjon om utvalg og håndtering av data, treningsprosessen, kildekode, vektorer, hyperparametere og ulike forutsetninger som er lagt til grunn. Det må også føres logger som viser hvordan modellen er endret gjennom læring eller oppdatering. Selv om gjeldende lovverk altså setter få konkrete krav til systemdokumentasjon, vil dette endre seg med vedtakelsen av ny arkivlov og lovgivning om KI-systemer fra EU.

Utkast til arkivlov § 9 og § 10 (omtalt i 4.3.2) vil delvis ivareta behovet for dokumentasjon, men jeg tror ikke bestemmelsene er tilstrekkelig rettet mot å dokumentere modellutviklingsprosessen. I henhold til § 9 første ledd bokstav c skal virksomheten dokumentere «hvilke behandlingsregler som er utledet av rettsreglene og som er styrende for vedtakene». Dersom behandlingsreglene er utledet ved hjelp av en maskinlæringsmodell trenger man også dokumentasjon på modellens utvikling. Uten slik dokumentasjon mangler trinnet som gir den nødvendige kontekst og bakgrunn for rettsanvendelsen, og det mangler et tilstrekkelig grunnlag for å ansvarliggjøre forvaltningsorganet for systemutviklingen og avgjørelsene.

EU-kommisjonens lovforslag inneholder, slik jeg forstår det, omfattende krav til dokumentasjon av modellutviklingen. Forslaget angår imidlertid ikke model-

188 Kaun (2021) s. 11.

lenes implementering i en rettsanvendelsesprosess.¹⁸⁹ Det kan derfor se ut til at den foreslåtte lovgivningen ikke er egnet til å sikre nødvendig dokumentasjon av en rettsanvendelsesprosess som integrerer maskinlæring.

Når det gjelder *systembeskrivelser* sa jeg i avsnitt 3.5 at de er nødvendige for å vise hvordan systemet er innrettet og utøver myndighet, og at de vil være viktige med tanke på å ansvarliggjøre forvaltningsorganet. En systembeskrivelse kan dessuten bidra til å klargjøre for offentligheten om systemets avgjørelser er hel- eller delautomatiserte. Tvilen rundt anvendelsen av GDPR artikkel 22 (se avsnitt 5.2), underbygger at det er et slikt behov. Systembeskrivelser kan videre bidra til å gjøre offentligheten oppmerksom på systemets eksistens og at kunstig intelligens inngår i beslutningsprosessen. Det er viktig med tanke på å aktualisere den passive informasjonsplikten.¹⁹⁰

Det er dermed flere argumenter for å lovfeste en plikt til å utarbeide systembeskrivelser som del av systemets dokumentasjon. I alle tilfelle kan det forankres i forvaltningens aktive informasjonsplikt.

4.4.2 Åpenhet krever aktiv formidling

I avsnitt 3.4.6 viste jeg til at litteraturen om algoritmiske beslutningssystemer bruker transparens i to betydninger, både i en passiv forstand som dreier seg om ren tilgjengeliggjøring av systemdokumentasjon, og i en systemorientert og formidlende forstand hvor formålet er å ansvarliggjøre aktøren (forvaltningsorganet) og vise hvordan systemet virker i sin kontekst. Det er nærliggende å assosiere denne distinksjonen med forholdet mellom forvaltningens passive og aktive informasjonsplikt.

Et passivt innsyn, der informasjon tilgjengeliggjøres i ubearbeidet format, vil først og fremst ha nytteverdi for miljøer med særskilt teknisk kompetanse, som for eksempel tilsynsmyndigheter. For et bredere publikum innebærer offentlighetsprinsippet, hensynene bak offentleglova og kravet om klarspråk i kommende språklav § 9 at forvaltningen må være mer målrettet og proaktiv i sin informasjonsvirksomhet om algoritmiske beslutningssystemer. Informasjonen

¹⁸⁹ I lovforslaget står det at leverandører av systemer definert som høyrisiko, må legge frem teknisk dokumentasjon som viser at systemet etterlever de krav som gjelder (art. 16, jf. art 11). Dette spesifiseres nærmere i vedlegg IV nr. 1. Første ledd krever en generell beskrivelse av systemet (blant annet dets formål, og en brukerveiledning). Annet ledd krever mer detaljert beskrivelse av systemets komponenter og utviklingsprosessen. Blant annet skal det opplyses om systemets generelle logikk, algoritmevalg, hva systemet optimerer for, parametere og deres relevans og forutsetninger som er tatt. Det skal også angis hva som er kildene til data og hvordan de er rensket og merket, og hva slags treningsmetoder som er benyttet. (EU-kommisjonen 2021).

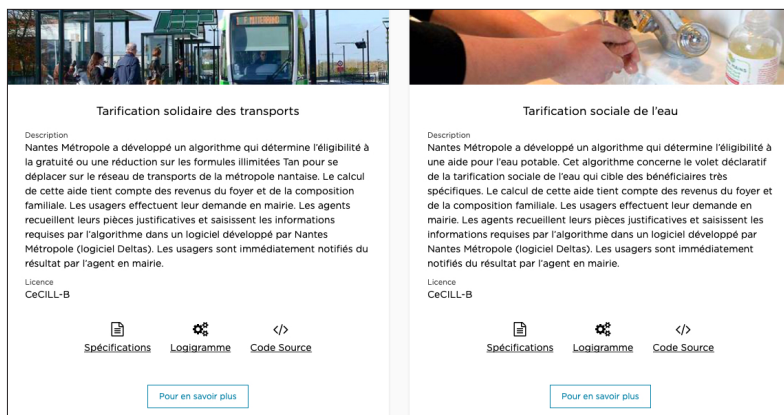
¹⁹⁰ Jf. formuleringen i NOU 2019: 5 s. 245.

må velges ut og formidles på en måte som gjør at borgerne kan forstå og nyttiggjøre seg den (se også avsnitt 4.3.3). Jeg vil i det følgende omtale noen tiltak som kan bidra til en mer åpen og forståelig algoritmisk forvaltning.

Det ene tiltaket er å *gjøre systemdokumentasjonen forståelig*. Arkivlovutvalget uttaler som nevnt om dette at dokumentasjon av automatiserte og delautomatiserte systemer i henhold til § 10 kan formidles som pseudokode, brukerhistorier, flytdiagrammer eller prosessbeskrivelser (se avsnitt 4.3.3). Jeg antar at det kan la seg gjennomføre også for maskinlæringsmodeller i algoritmiske beslutnings-systemer. Det kan nok være hensiktsmessig å begrense mengden av informasjon som tilpasses på denne måten, for eksempel ved at man velger ut de mest sentrale dokumentene og vurderingene.

Åpenhet kan også oppnås ved ulike varianter av *systemoffentlighet*. En brukervennlig løsning kan være å publisere modellen med et grensesnitt som gjør at borgerne selv kan teste den ut. De kan simulere utfall for ulike inngangsvaria-bler og se hvordan modellen oppfører seg.¹⁹¹

Et annet alternativ kan være å publisere eller tilby *brukerveiledninger* og *systembeskrivelser* som viser hvordan modellene er utviklet og beslutningssystemet er innrettet, og modellene utviklet (se også avsnitt 3.5).

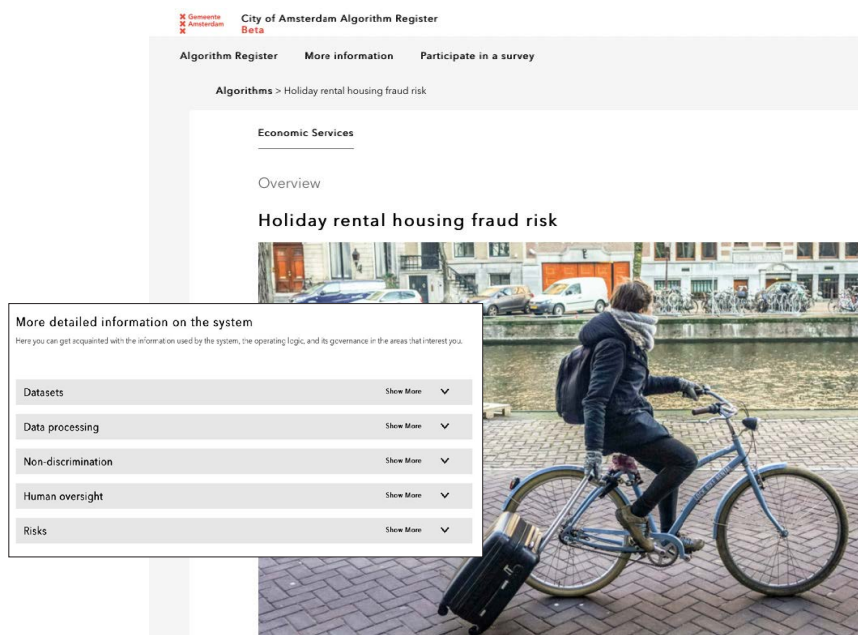


Figur 21: Nantes Métropoles åpne algoritmer for å beregne rabatt på offentlig transport og vann.¹⁹²

191 Sandis (6.7.20), Schartum et al. (2017) s. 103, Edvardsen og Engehagen (samtale 2.10.20). Man kan også se for seg slike løsninger til intern bruk for kompetanseheving blant ansatte.

192 Kilde: https://data.nantesmetropole.fr/pages/algorithmes_nantes_metropole/, oppslag 18.10.20

Et eksempel på en systembeskrivelse finnes på Nantes storbyområde sine nettsider. Der er det publisert informasjon om to automatiserte systemer for å behandle søknader om reduserte priser på offentlig transport og vann. Ved siden av en overordnet beskrivelse av systemet finnes mer detaljert informasjon publisert under overskriftene «Spesifikasjoner», «Flytdiagram» og «Kildekode». Under Spesifikasjoner finner man algoritmens regler, og hvordan de anvendes i tekst og pseudokode. Flytdiagrammet illustrerer hvilke trinn algoritmen går igjennom for å foreta prisberegningen, mens kildekoden er algoritmens programkode. Se figur 21, som er et skjermbilde fra Nantes Métropoles nettsider.



Figur 22: Amsterdams red-flag modell for å fore kontroll med ferieutleie av boliger.¹⁹³

Et annet eksempel finner man i Amsterdams register over algoritmiske beslutningssystemer som brukes i lokal forvaltning. Ett av disse er et «red flag»-kontrollsystem for å fange opp brudd på regler om utleie av leiligheter til turister. Skjermbildene i figur 22 gir et inntrykk av hvilken informasjon som er tilgjengelig for den som slår opp i registeret.

¹⁹³ Kilde: <https://algorithmeregister.amsterdam.nl/en/holiday-rental-housing-fraud-risk/>. Oppslag 6.12.20.

Under *Overview* får man en kort omtale av systemets formål og virkemåte. Under *More detailed information on the system* får man en beskrivelse av hvilke variabler som inngår, hvordan algoritmen brukes til å fatte avgjørelser, hvordan potensielle skjevheter er håndtert, risikoaspekter ved systemet og hva som er gjort for å unngå diskriminering. Publikum kan også gi tilbakemeldinger og komme med spørsmål.¹⁹⁴

Informasjonen som gis her kan man karakterisere som en utvidet global forklaring, jf. begrepsdefinisjonene i avsnitt 3.5.

Det vil være tillitvekkende at slik informasjon er tilgjengelig, og det kan bidra til å øke kunnskapsnivået og bevisstheten om algoritmiske beslutningssystemer i befolkningen.

Utover tiltakene omtalt ovenfor kan det bli nødvendig å vurdere hvordan man ivaretar behov for mer inngående kontroll gjennom bruk av innsynsretten. Her har jeg i tankene media, interessegrupper og frivillige organisasjoner som besitter eller kan skaffe den nødvendige tekniske kompetansen for å utføre kontroll på vegne av sine medlemmer eller interessene de representerer. Spørsmålet er om et innsynskrav i systemdokumentasjonen vil være tilstrekkelig for denne typen aktører.

Problemstillingen reises i et eksempel beskrevet av Amariles, i en sak om det algoritmiske beslutningssystemet APB (Admission Post Bac) iverksatt av franske utdanningsmyndigheter. Systemet skulle matche søkere til studieplass i henhold til vilkårene i lov om høyere utdanning. Formålet var å oppnå effektivitet og upartiskhet i tildelingen av studieplasser. Da det ble uttrykt bekymring om hvorvidt plassene på de mest populære studiene ble tildelt på en rettferdig måte, ba en elevorganisasjon om innsyn i algoritmen for å finne ut hvordan systemet virket.¹⁹⁵

Elevorganisasjonen fikk først ikke svar fra Utdanningsdepartementet. Organisasjonen tok da rettslige skritt, og fikk medhold i at kildekode var et offentlig dokument og gjenstand for allmennoffentlighet. Utdanningsdepartementet friga da kildekode i en PDF-fil, fordi det ellers kunne være en fare for at søkerens personopplysninger ble lekket. Dette var imidlertid ikke tilstrekkelig til å forstå hvordan systemet fattet avgjørelsene. Det ble derfor planlagt en *hackathon*

194 Venturebeat (28.9.20) «Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI», <https://venturebeat.com/2020/09/28/amsterdam-and-helsinki-launch-algorithm-registries-to-bring-transparency-to-public-deployments-of-ai/>, oppslag 9.11.20.

195 Amariles (2020) s. 284–286.

for å finne ut av det.¹⁹⁶ Før hackathonen fant sted ble imidlertid systemet satt til side av det det franske datatilsynet (CNIL). Årsaken var at systemet stred imot forbudet mot helautomatiserte algoritmiske avgjørelser i lov nr. 78-17 (det som nå er GDPR artikkel 22).¹⁹⁷

Forvaltingsorganet kan altså ikke oppfylle sin informasjonsplikt bare ved å gi ut bare kildekode, fordi det ikke er tilstrekkelig til å forstå hvordan et algoritmisk beslutningssystem basert på lærende algoritmer virker. I avsnitt 3.3.2 om etterprøvbarehet sa jeg at en grunnleggende form for innsyn kan gjennomføres ved at det gis tilgang til systemets kildekode og vektorer, alternativt ved at man kopierer modellen på et gitt virkningstidspunkt. Denne formen for innsynskrav vil ikke utløse noe behov for å gi ut datasett med personopplysninger brukt til trening av modellen.

Dersom modellen skal rekonstrueres, eller man skal ettergå dens forutsetninger i sin helhet, vil det være nødvendig med de fulle datasettene som ble brukt for å trene modellen. En slik omfattende gjennomgang bør forbeholdes offentlige tilsyns- eller revisjonsmyndigheter, fordi man ikke kan gi ut personopplysninger.

Jeg er usikker på hva som skal til av dokumentasjon for å dekke behovet til eksterne aktører som skal drive kontroll med forvaltningen, og nøyer meg med å vise til problemstillingen. Forvaltningen må i alle fall legge til rette for egnede løsninger for å besvare innsynskrav innenfor rammen av personvernet.

4.4.3 Medvirkning for mer representative modeller

Systembeskrivelser, systemoffentlighet og innsyn i systemdokumentasjon vil ikke alltid være tilstrekkelig fra et demokratiperspektiv. Når det skal implementeres systemer med stor velferdsmessig betydning, som vil ha konsekvens for grupper som ikke er godt representert der beslutningene tas eller som involver vurderinger som potensielt er kontroversielle, kan det være nødvendig å legge til rette for mer deltakende prosesser. Det kan bidra til at modellene bygger på akseptable og representative data, at uenighet om forutsetninger og indikatorer er belyst, og at personer som blir direkte berørt har fått uttale seg.

Den prediktive analysemodellen *Allegheny Family Screening Tool (AFST)* er et eksempel på hvordan det er jobbet med åpenhet og medvirkning i utviklingsfasen av et algoritmisk beslutningsstøttesystem. AFST brukes av barnevernet (Child Services) i Allegheny fylke i delstaten Pennsylvania. Algoritmens predik-

¹⁹⁶ Et hackathon er en samling hvor utviklere samarbeider innenfor en begrenset tidsramme, vanligvis om å utvikle kode for å løse et konkret problem. I dette tilfellet var formålet å bruke koden de hadde fått tilgang til for å finne ut hvordan systemet matchet studieplass med søker.

¹⁹⁷ Amariles (2020) s. 284–286.

sjon skal hjelpe saksbehandleren til å vurdere om en bekymringsmelding skal lede til nærmere undersøkelser, eller om den kan legges til side.

Utvelgelsesprosessen var opprinnelig innrettet uten beslutningsstøtte. Saksbehandlerne hadde tilgang til mye data når de fikk en bekymringsmelding å ta stilling til, men det var vanskelig å finne frem til de relevante faktorene under tidspress. Det førte til at utvelgelsen ofte ikke traff godt, slik at det ble opprettet saker i situasjoner der det ikke viste seg nødvendig mens alvorlige tilfeller av omsorgssvikt og mishandling ikke ble fanget opp.

På grunn av disse svakhetene valgte barnevernmyndighetene å utvikle en prediktiv analysemodell. Algoritmen som ble utviklet har 100 kriterier, og henter data fra 8 databaser innenfor kriminalomsorg, velferdsforvaltning, rusomsorg, hjemløshet og psykiatri. Prediksjonen har form av en skår som indikerer sannsynligheten for at barnet vil bli tatt ut av hjemmet i løpet av de neste to årene.

Utviklingsprosessen har blitt beskrevet som gjennomslutning og deltakende. Det er fullt innsyn i variablene, som er gjennomgått og kvalitetssikret av akademikere og offentlig ansatte. Det har vært åpne høringer der ulike grupper av interessenter har deltatt, som advokater, eksperter på barnevern, foreldre og tidligere fosterbarn. I materialet som ligger på Internett, beskriver de hvilke vurderinger som er gjort ved definisjon av variabler og valg av metoder. Det er også utført en etisk konsekvensvurdering.

En evaluering viste blant annet at systemet førte til at flere saker ble opprettet på korrekt grunnlag (det man kan kalle ekte positive, se avsnitt 2.4), uten at det totale antallet saker henvist for oppfølging økte. Samtidig økte andelen hvite barn blant de ekte positive, mens andelen barn fra svarte familier gikk ned. Modellen førte dermed til bedre demografisk paritet (likebehandling på tvers av befolkningsgrupper) sammenlignet med den tidligere rutinen.

Eksempelen viser at medvirkning kan gjøre modellene mer treffsikre og relevante for sitt formål, og dermed bidra til avgjørelser mindre preget av bias. Deltakelse kan også bedre forståelsen for modellen og forankre den hos brukerne og de berørte. At modellens muligheter og begrensninger forstås av saksbehandlerne har også betydning for kvaliteten på avgjørelsene de tar, som jeg kommer tilbake til i avsnitt 5.4.3. Selv om AFST berømmes for åpenheten har systemet og

medvirkningen fra fagfolk og berørte grupper har modellen også problematiske sider, som har blitt kritisert.¹⁹⁸

Flere tar til orde for at det er nødvendig med mekanismer for medvirkning eller høring for å bidra til at systemene er forståelige og ansvarlige.¹⁹⁹ EU høynivå-ekspertgruppe på kunstig intelligens mener at interessenter som blir direkte eller indirekte berørt bør høres både i forkant av, og i løpet av, systemets livssyklus.²⁰⁰

Som omtalt i avsnitt 4.2.4. har forvaltningsretten ulike regler om medvirkning, høring og samhandling, som kan tilpasses utviklingen av algoritmiske beslutningssystemer. Ett slikt grunnlag er GDPR artikkel 35 om personvernkonsekvensvurdering. En personvernkonsekvensvurdering foretatt i forbindelse med et algoritmisk beslutningssystem kan kalles algoritmisk konsekvensvurdering (*algorithmic impact assessment*201). Når berørte parter og fageksperter involveres i forbindelse med en konsekvensvurdering (artikkel 35 9. ledd) kan det gi kvalitativt bedre systemer med bredere forankring, samtidig som at medvirkningen fører til større forståelse for hvordan teknologien og modellen virker.²⁰²

4.4.4 Oppsummering

I dette kapittelet har temaet vært forklarbarhet og transparens som bidrag til å ansvarliggjøre forvaltningsorganet og forhindre at beslutningssystemer iverksettes uten demokratisk forankring eller mulighet for offentlighetens innsyn.

Det finnes regler og prinsipper som skal ivareta demokratiet gjennom informasjon og medvirkning, men mine drøftinger har vist at algoritmisk forvaltning kan gjøre det nødvendig med noen tekniske og regulatoriske tilpasninger.

I avsnitt 4.4 har jeg for det første argumentert for at algoritmiske beslutningssystemer bør dokumenteres ved hjelp av dokumentasjon fra modellutviklingsprosessen og dessuten gjennom systembeskrivelser som viser hvordan systemet utøver myndighet i samspill med andre komponenter (som menneskelig skjønn og faste algoritmer). Dette vil kreve at regelverket tilpasses.

198 <https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx>, oppslag 13.1.21, New York Times 2.1.18 «Can an Algorithm Tell when Kids are in Danger? »

199 Amariles (2020) s. 287

200 EUs høynivåekspertgruppe på kunstig intelligens (2019) pkt. 82.

201 Begrepet *algorithmic impact assessment* brukes litt ulikt av forskjellige forfattere. Reisman et al. (2018) ser på en slik konsekvensvurdering som en offentlig prosess med høring, til forskjell fra en vanlig personvernkonsekvensvurdering i henhold til GDPR.

202 Kaminski og Maglieri (2019).

Arkivlovutvalget har foreslått en bestemmelse om dokumentasjon av systemer for hel- og delautomatisert rettsanvendelse (§ 9), men den krever ikke dokumentasjon av *hvordan* behandlingsreglene er utledet av rettsreglene, noe som vil være nødvendig dersom automatiseringen skjer ved hjelp av en maskinlæringsmodell.

Det finnes ikke foreløpig noe spesifikt rettslig grunnlag for å kreve at forvaltningsorganet utarbeider systembeskrivelser, men ettersom det kan påvises et slikt behov (det vil også bli tydelig i eksemplene som omtales i kapittel 5) bør også det lovfestes. Uansett bør forvaltningsorganet se det som del av sin aktive informasjonsplikt.

Algoritmisk forvaltning krever en mer proaktiv holdning til informasjonsplikten, fordi informasjon om modeller og beslutningssystemer i de aller fleste tilfeller må velges ut og formidles på en egnet måte, også i forbindelse med innsynskrav etter offentleglova. Jeg presenterte noen tiltak for tilpasning og formidling, herunder mer forståelig systemdokumentasjon, systemoffentlighet og tilgang til brukerveiledninger og systembeskrivelser. Jeg reiste også et spørsmål jeg ikke har forutsetning for å svare på, nemlig om ulike samfunnsaktører (som organisasjoner og interessegrupper) vil ha tilstrekkelig mulighet for å føre kontroll med forvaltningens systemer gjennom innsynsretten. Dersom kontrollen krever at man rekonstruerer modellen ved hjelp av tilgang til treningsdata som inneholder personopplysninger, vil personvernet utgjøre en begrensning. Dersom det kan være tilstrekkelig med kildekode og vektorer supplert med dokumentasjon på at systemet etterlever lovfestede krav, burde det ikke by på problemer.

Avslutningsvis sa jeg at medvirkning bør etterstrebes ved utvikling av systemer med potensielt store velferdsmessige konsekvenser for borgerne. Medvirkning kan gjøre maskinlæringsmodellene mer representative i dobbel forstand, ved å sikre at variablene som inngår i modelleringen er relevante for formålet og ved at ulike interesser blir hørt.

Jeg går nå over til å behandle rettssikkerhet i algoritmisk velferdsforvaltning, der maskinlæringsmodeller bidrar til avgjørelser rettet mot enkeltpersoner.

5 Forklarbarhet som forutsetning for en rettssikker velferdsforvaltning

5.1 Innledning

Algoritmiske beslutningssystemer som fatter eller tilrettelegger grunnlaget for forvaltningsavgjørelser må overholde forvaltningsrettens rettssikkerhetsgarantier. Rettssikkerhetsgarantiene skal bidra til riktige og rettferdige avgjørelser, og beskyttelse av enkeltindividet i møte med forvaltningen.

Bruk av maskinlæringsmodeller kan gjøre det vanskelig å få innsyn i eller forstå grunnlaget for en avgjørelse, noe som utfordrer retten til kontradiksjon. Maskinlæringsmodellene kan dessuten være trent på eller benytte store datasett med personopplysninger i avgjørelsene. Det kan utgjøre en risiko for privatliv, personvern og retten til ikke-diskriminering. I dette kapittelet er derfor spørsmålet hvordan begreper og metoder for å fremme forklarbarhet kan bidra til rettssikre algoritmiske avgjørelser.

Avsnitt 5.2 behandler forklarbarhet i lys av retten til privatliv og personvern, og tar spesielt for seg GDPRs bestemmelser om åpen og rettferdig behandling og forbudet mot helautomatiserte avgjørelser. Jeg utleder hvilke krav til forklarbarhet som skal stilles til kontrollsystemer i velferdsforvaltningen basert på SyRI-dommen.

Avsnitt 5.3 går nærmere inn på hvilke forklaringer som er nødvendige for å ivareta retten til kontradiksjon. Manglende innsyn og begrunnelse kan være i strid med prinsippet om åpen og rettferdig behandling. Jeg bruker Datatilsynets vurdering av IB sin karakteralgoritme som eksempel.

Avsnitt 5.4 er en analyse av to lokale forklaringsmetoders egnethet for forvaltningsavgjørelser. Analysen tar utgangspunkt i den tekniske gjennomgangen i kapittel 3.

I avsnitt 5.5 er temaet algoritmisk diskriminering, som er en særlig rettssikkerhetsrisiko når maskinlæringsmodeller integreres i forvaltningens saksbehandling. Forklarbarhet handler da om å dokumentere at det er brukt egnede metoder og tiltak for å avdekke og håndtere slagside (bias).

5.2 Grunnleggende rettslige krav om ivaretagelse av privatliv og personvern ved algoritmisk forvaltning

Retten til privatliv er beskyttet av Grunnloven § 102 og EMK artikkel 8. Sistnevnte lyder:

1. Everyone has the right to respect for his private and family life, his home and his correspondence.
2. There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

Grunnloven § 102 må tolkes i lys av EMK art. 8, og Høyesterett har lagt til grunn at det på tilsvarende måte som for EMK art. 8 (2) kan skje inngrep i privatlivet dersom tiltaket har hjemmel, ivaretar et legitimt formål og er forholdsmessig.²⁰³

EMD har gjentatte ganger understreket at «privatliv» må tolkes vidt; det kan ikke avgrenses absolutt. I henhold til EMDs praksis er beskyttelse av personopplysninger sentralt for ivaretagelsen av retten til privatliv. Det å lagre data som kan knyttes til noens privatliv anses i seg selv som innblanding (interference), men om det har skjedd en krenkelse av artikkel 8 kommer an på konteksten: måten opplysningene har blitt samlet inn, lagret og behandlet og hvordan resultatene brukes.²⁰⁴ Brudd på personvernlovgivningen kan derfor samtidig innebære en krenkelse av EMK art. 8.²⁰⁵

5.2.1 GDPR og kravene til åpenhet, innsyn og informasjon til den registrerte

Nærmere regler for håndtering av personopplysninger er gitt i personopplysningsloven, som inkorporerer EUs generelle personvernforordning (GDPR), jf. personopplysningsloven § 1. GDPR har blant annet som siktemål å gi den registrerte innsikt i og kontroll over når og hvordan personopplysninger behandles. GDPR gir den registrerte rett til å be om innsyn, retting og sletting og rett til å protestere eller reservere seg.

²⁰³ Rt. 2014 s. 1105 (28).

²⁰⁴ EMD: Case of S. and Marper v. United Kingdom, avsnitt 67, EMD: Case of Satakunnan Markkinapörssu Oy and Satamedia Oy v. Finland, avsnitt 136.

²⁰⁵ I SyRI-saken omtalt nedenfor, tolket retten EMK art. 8 annet ledd i lys av GDPRs prinsipper for beskyttelse av personopplysninger, og viste til at det ikke var grunn til å anta at minimumsbeskyttelsen av privatlivet er mindre omfattende enn det som følger av GDPR (avsnitt 6.41 i dommen).

GDPR artikkel 5 første ledd nedfeller prinsipper for behandling av persondata. Disse er: lovlighet, rettferdighet og åpenhet, formålsbegrensning, dataminimering, riktighet, lagringsbegrensning og integritet og konfidensialitet. Prinsippene danner utgangspunkt for mer spesifikke bestemmelser.

Det er spesielt prinsippene om lovlighet, rettferdighet og åpenhet i bokstav a som har betydning for spørsmålet om forklarbarhet. Jeg går derfor litt nærmere inn på innholdet i disse og koblingen til de bestemmelsene som regulerer rett til informasjon og innsyn for den registrerte.

I prinsippet om lovlighet ligger at det må finnes et rettslig grunnlag for å behandle personopplysningene. Disse grunnlagene er opplistet i artikkel 6 første ledd. For eksempel kan det være utøvelse av offentlig myndighet som den behandlingsansvarlige er pålagt, jf. bokstav e. Behandlingen må i alle tilfeller tilfredsstillende et krav om nødvendighet, med unntak for tilfeller der det er gitt samtykke fra den registrerte, jf. bokstav a.

Med rettferdighet og åpenhet menes blant annet at den registrerte må bli informert om at personopplysningene vil bli behandlet, formålet med behandlingen og eventuelle risikoer for uheldige konsekvenser. Formålet må være innenfor det den registrerte med rimelighet kan forvente. Den registrerte har krav på en lett-fattelig forklaring, i tekst eller gjennom visualisering, hvordan opplysningene vil bli behandlet. Krav til kommunikasjons utforming er nærmere presisert i artikkel 12 nr. 1.

Av fortalens punkt 39, 58, 60 og 71 følger det også at det må være klart hvilke risikoer som er forbundet med behandlingen, hvilke beskyttelsesmekanismer som er på plass og hvilke rettigheter den registrerte har. Det må gjennomføres tekniske og organisatoriske tiltak som forhindrer feil og usaklig eller diskriminerende forskjellsbehandling.

Som nevnt i avsnitt 2.3.1 har systemer som lager og anvender profiler vært en særlig kilde til bekymring for personvernet. Profilerings innebærer at det genereres nye personopplysninger, som personene det gjelder ofte ikke kjenner til eller har forutsetninger for å forstå grunnlaget for. Dette er særlig aktuelt dersom man benytter komplekse maskinlæringsmodeller i profileringen. Profilerings som grunnlag for forskjellsbehandling medfører også en risiko for urettferdighet og diskriminering. Ved siden av prinsippet om åpenhet er artikkel 13-15 om rettigheter til informasjon og innsyn og artikkel 22 om automatiserte avgjørelser sentrale for ivaretagelse av personvernet ved bruk av algoritmiske beslutningssystemer.

Artikkel 13 og 14 gir den registrerte rett til diverse informasjon om formål, lagring, mulighet for å klage (med videre) i forbindelse med innsamling av personopplysninger, mens artikkel 15 gir den registrerte rett til å be om innsyn i lagrede personopplysninger, kilden til opplysninger, hvordan opplysningene skal behandles, informasjon om rett til å klage med mer. Artikkel 13-15 inneholder alle en rett til å bli informert dersom personopplysningene skal inngå i helautomatiserte avgjørelser, herunder profilering.

Artikkel 22 første ledd slår fast at:

Den registrerte skal ha rett til ikke å være gjenstand for en avgjørelse som utelukkende er basert på automatisert behandling, herunder profilering, som har rettsvirkning for eller på tilsvarende måte i betydelig grad påvirker vedkommende.

Første ledd er formulert som en rettighet for den registrerte, men tolkes ofte som et forbud som retter seg mot den behandlingsansvarlige.²⁰⁶ At avgjørelsen «på tilsvarende måte påvirker vedkommende» skal ifølge Artikkel 29-arbeidsgruppen forstås som at den har langvarige eller permanente konsekvenser for den registrerte, at det kan påvirke den registrertes adferd, valg eller situasjon i betydelig grad, eller fører til eksklusjon eller diskriminering.²⁰⁷

Artikkel 22 omhandler *helautomatiserte* avgjørelser, slik at anvendelsesområdet snevres inn betraktelig. Systemer hvor prediksjoner fra en maskinlæringsmodell bare gir beslutningsstøtte vil altså ikke være omfattet.

Det kan være vanskelig å avgjøre om det er fattet en «avgjørelse» og om den skal karakteriseres som helautomatisert eller ikke. Artikkel 29-arbeidsgruppens føring er at det menneskelige elementet må være reelt, og ikke ha blitt innlemmet i beslutningsprosessen med tanke på å omgå artikkel 22. Det menneskelige bidraget må bestå i en vurdering av alle relevante data og kunne ha innflytelse på utfallet av saken.²⁰⁸ Dette var et omstridt spørsmål i SyRI-saken som jeg omtaler nedenfor i avsnitt 5.2.3. I den saken tok ikke retten endelig stilling til om artikkel 22 kom til anvendelse.²⁰⁹

²⁰⁶ Bygrave (2019) s. 252.

²⁰⁷ Artikkel 29-arbeidsgruppen (2017) s. 22. Artikkel 29-arbeidsgruppen var EUs rådgivende organ i personvernspørsmål frem til ikrafttreddelsen av GDPR i 2018. Gruppen har publisert en rekke veiledere om tolkning og anvendelse av GDPRs bestemmelser. Det er nå Personvernrådet (EDPB) som fyller funksjonen.

²⁰⁸ Artikkel 29-arbeidsgruppen (2017) s. 21.

²⁰⁹ Avsnitt 6.60 i dommen.

Annet ledd bokstav lister opp tre alternative grunnlag for unntak fra forbudet. Det er bokstav b) som er det mest relevante unntaksgrunnlaget for offentlig forvaltning. For at automatisert saksbehandling basert på profilering skal være lovlig, må det være uttrykkelig hjemlet i lov og det må være mekanismer som sikrer de registrertes rettigheter og interesser.²¹⁰

Tredje ledd gjelder rettssikkerhetsgarantier for den registrerte i tilfeller der det likevel er tillatt med automatiserte avgjørelser etter annet ledd bokstav a) og c). Da må den behandlingsansvarlige sørge for egnede tiltak som ivaretar de registrertes rettigheter og interesser. Som et minimum må den registrerte få mulighet til «menneskelig inngripen fra den behandlingsansvarlige, til å uttrykke sine synspunkter og til å bestride avgjørelsen».

Fjerde ledd angår automatiserte avgjørelser basert på sensitive personopplysninger, såkalte «særlige kategorier av personopplysninger» i henhold til artikkel 9 første ledd. Det omfatter blant annet helseopplysninger. I disse tilfellene er kravene strengere. Det kreves uttrykkelig samtykke fra den registrerte, eller at behandlingen er nødvendig av hensyn til viktige allmenne interesser.

5.2.2 Gir GDPR en rett til forklaring på avgjørelser basert på maskinlæring?

Artikkel 22 tredje ledd har stått sentralt i en diskusjon de siste årene, om hvilken rett til forklaring av algoritmiske avgjørelser som følger av GDPR. Tanken er at dersom man kan utlede en slik rett, så vil det være et viktig bidrag til etterrettelighet og ansvarlighet i algoritmiske beslutningssystemer fremover.

Etter sin ordlyd gir 22 (3) gir først og fremst en rett til å kreve at et menneske trekkes inn i avgjørelsesprosessen («human-in-the-loop») og en mulighet til å klage. Forklaring nevnes ikke i artikkel 22, men det fremgår uttrykkelig av fortalens punkt 71, hvor det blant annet står at (min utheving):

Under alle omstendigheter bør en slik behandling ledsages av nødvendige garantier som bør omfatte spesifikk informasjon til den registrerte og rett til menneskelig inngripen, til å uttrykke sine synspunkter, *til å få en forklaring på avgjørelsen som er truffet etter en slik vurdering*, og til å protestere mot avgjørelsen.

210 Basert på en tolkning av artikkel 22 (1) som et forbud mot helautomatiserte avgjørelser, har Forvaltningslovutvalget foreslått en bestemmelse som skal gi hjemmel for automatisert saksbehandling hvor personopplysninger inngår, jf. 22 (2) bokstav b. Utkastets § 11 lyder: «Kongen kan gi forskrift om at forvaltningsorganet på bestemte sakområder kan treffe avgjørelser ved helautomatisert saksbehandling. Avgjørelser som er lite inngripende overfor den enkelte, kan treffes ved helautomatisert behandling uten hjemmel i forskrift.». Flertallet mente det var tilstrekkelig for å overholde GDPR å gi en overordnet hjemmel i forvaltningsloven til lite inngripende saker, men antok at behovet for klar hjemmel ville skjerpes ved for eksempel mer avansert profilering. (NOU 2019: 5 s. 262–263.)

I rettsteorien er det ulike oppfatninger om hvilken rett til forklaring som følger av GDPR på dette punktet. Noen hevder fortalens punkt 71 indikerer at tredje ledd innebærer en rett til forklaring av helautomatiserte avgjørelser *ex post*, det vil si en lokal forklaring av en enkeltprediksjon etter at en avgjørelse er truffet.²¹¹

Wachter et al. mener på sin side at artikkel 22 (3) ikke medfører en rett til lokal forklaring (definert som “the logic or rationale, reasons, and individual circumstances of a specific automated decision”).²¹² De viser til at fortalen ikke er rettslig bindende, og at den behandlingsansvarlige ikke kan pålegges en plikt gjennom en uklar formulering i fortalen. Lovgivningsprosessen underbygger at det heller ikke var hensikten.²¹³

Det er også andre bestemmelser som kan sies å gi en rett til forklaring på automatiserte avgjørelser, nemlig artikkel 13 (2) f, 14 (2) g og 15 (1) h om den behandlingsansvarliges plikt til å gi informasjon når det er samlet inn data og den registrertes rett til å be om innsyn. I alle tre bestemmelsene står det at den behandlingsansvarlige skal gi informasjon om:

forekomsten av automatiserte avgjørelser, herunder profilering, som nevnt i artikkel 22 nr. 1 og 4, og, i det minste i nevnte tilfeller, relevant informasjon om den underliggende logikken samt om betydningen og de forventede konsekvensene av en slik behandling for den registrerte

Artikkel 29-arbeidsgruppen uttaler at artikkel 13 (2) f og 14 (2) g innebærer at den registrerte må gjøres oppmerksom på at det blir tatt en avgjørelse basert på en profil, og at den behandlingsansvarlige må forklare på en enkel og tydelig måte hvordan avgjørelsene fattes.²¹⁴

Wachter et al. påpeker at forklaringen her må dreie seg om å gi en beskrivelse av «system functionality», fordi det ennå ikke er fattet noen avgjørelse. Dataene er bare samlet inn.²¹⁵ Systemfunksjonalitet oppfatter jeg som tilsvarende en global forklaring; en overordnet beskrivelse av hvordan systemet fungerer, uavhengig av enkeltprediksjonene. (Wachter et al. definerer det som: “the system’s requirements specification, decision trees, pre-defined models, criteria, and classification structures; »).²¹⁶

211 Goodman og Flaxman (2016), Kaminski og Maglieri (2019).

212 Wachter et al. (2017) s. 6.

213 Wachter et al. (2017) s. 11–12.

214 Artikkel 29-arbeidsgruppen (2017) s. 16.

215 Wachter et al. (2017) s. 15–16.

216 Wachter et al. (2017) s. 6.

Om artikkel 15 (1) h mener Artikkel 29-arbeidsgruppen at den krever at behandlingsansvarlige finner en enkel måte å forklare bakgrunnen for avgjørelsen («the rationale behind, or the criteria relied on in reaching the decision»). Den skal ikke involvere komplekse utlegninger eller gjengivelse av algoritmen. Forklaringen må likevel være tilstrekkelig til at den registrerte kan forstå utfallet.²¹⁷ Bygrave støtter at bestemmelsen innebærer en rett til forklaring *ex post*, fordi han mener det er nødvendig for å realisere muligheten for å klage. Det følger dessuten av kravene til den behandlingsansvarlige i art. 5 (2) og 24.²¹⁸

Wachter et al. mener imidlertid at heller ikke innsynsretten etter artikkel 15 (1) h innebærer noen *ex post* rett til forklaring av en enkeltprediksjon, og viser til at ordlyden («forventede konsekvenser») peker frem i tid mot en mulig avgjørelse. Det betyr at det også her er systemfunksjonaliteten, modellens globale virkemåte, som den behandlingsansvarlige pålegges å forklare. Det kreves bare en helt overordnet beskrivelse av systemets logikk.²¹⁹

Når det gjelder 22 (4) om særlige kategorier av personopplysninger, bemerker Edwards og Veale at den fører til en «catch-22»-situasjon, fordi man ikke kan vite om en avgjørelse baserer seg på sensitive data uten at man har innsyn i hvilke data algoritmen baserer seg på. Man må litt enkelt sagt ha en forklaring for å vite om man har krav på forklaring.²²⁰

Gjennomgangen her viser at GDPR gir de registrerte krav på en global forklaring (i hvert fall i snever forstand) av maskinlæringsmodeller som benyttes til å fatte automatiserte avgjørelser, mens det er uenighet om man direkte kan utlede noen rett til forklaring av modellens enkeltprediksjoner. Ettersom artikkel 22 viser seg å ha et begrenset nedslagsfelt og inneholder vilkår som byr på en del tvil, kan prinsippet om åpenhet i artikkel 5 (1) a få større betydning for retten til forklaring av algoritmiske beslutningssystemer.²²¹ SyRI-saken, som omtales straks nedenfor, er et eksempel på det.

Diskusjonen får ikke så stor praktisk betydning for tilfeller der maskinlæringsmodeller skal benyttes til å fatte avgjørelser i offentlig forvaltning. En rett til lokal forklaring vil uansett følge av forvaltningslovens regler om begrunnelse, veiledning og partsinnsyn. Den har likevel interesse og overføringsverdi, fordi de rettslige vurderingene foretas med utgangspunkt i de tekniske realitetene ved maskinlæringsmodeller.

217 Artikkel 29-arbeidsgruppen (2017) s. 25.

218 Edwards og Veale (2017) s. 52, Bygrave (2019) s. 254.

219 Wachter et al. (2017) s. 17, s. 29–30.

220 Edwards og Veale (2017) s. 50.

221 For eksempel: Gil González og de Hert (2019) s. 618.

Kaminski og Maglieri påpeker for øvrig at man i juridisk teori etter hvert har beveget seg i retning av å se på forklaring ikke bare i forbindelse med de individuelle rettighetene i GDPR. Reglene for systemkontroll, som for eksempel konsekvensvurdering, involverer også ulike redegjørelser, beskrivelser og konsultasjoner knyttet til systemet. Når man ser disse samlet får man lagvise forklaringer som utfyller hverandre («multi-layered explanations») og som bidrar til ansvarlighet og etterrettelighet.²²² (Dette ble også omtalt i avsnitt 4.3.3.)

5.2.3 SyRI-saken: Et algoritmisk beslutningssystem i strid med retten til privatliv og personvern

Den nederlandske SyRI-saken er en god illustrasjon på flere problemer ved bruk av maskinlæringsmodeller i forvaltningsbeslutninger. SyRI (Systeem Risico-Indicatie) er et algoritmisk beslutningssystem som er utviklet av nederlandske myndigheter for å avdekke svindel med skatt, sosialhjelp og trygdeytelser. SyRI er en form for prediktiv analyse, der en risikomodell identifiserer personer med forhøyet risiko for svindel basert på ulike indikatorer. Systemet ble implementert som prosjekter rettet mot utvalgte distrikter eller bydeler, der to eller flere offentlige etater kunne inngå samarbeid om å koble sine brukerdata mot hverandre.²²³

En koalisjon bestående av en fagforening, sivilsamfunnsorganisasjoner og to privatpersoner (omtalt i dommen som NJCM et al.) gikk til søksmål mot staten for å få systemet stoppet. Deres påstand var at SyRI var i strid med retten til privatliv og personvern, jf. blant annet EMK art. 8 og GDPR, fordi det fungerte som en svart boks med minimalt med innsyn for offentligheten og de berørte.²²⁴ Saksøkerne mente også at SyRI ble anvendt på en måte som diskriminerte og stigmatiserte enkelte grupper, fordi prosjektene var rettet mot fattige og sosialt belastede områder. Dette ble støttet av FNs spesialrapportør på ekstrem fattigdom i hans *amicus curiae*.²²⁵

Saken omtales her fordi den belyser hvilke krav til forklarbarhet som kan bli stilt til algoritmiske beslutningssystemer som benyttes i velferdsforvaltning, fortrinnsvis til kontrollformål, med grunnlag i GDPR artikkel 22 og EMK artikkel 8. Saken har fått stor oppmerksomhet internasjonalt, og blir sett på som en prøvesak for algoritmiske kontrollsystemer da det er få eksempler på at et slikt system er trukket for retten. Dommen ble avsagt 5. februar 2020 av Rechtbank Den Haag (førsteinstansen) og ble ikke anket. I etterkant har nederlandske trygdemyndigheter lagt ned SyRI-systemet.

²²² Kaminski og Maglieri (2019) s. 3.

²²³ Avsnitt 3.1 og 4.5 i ECLI:NL:RBDHA:2020:1878, Rechtbank Den Haag 5.2.20.

²²⁴ Avsnitt 5.1.

²²⁵ FNs spesialrapportør på ekstrem fattigdom (26.9.19).

SyRI hadde sin lovmessige forankring i SUWI-loven²²⁶ og SUWI-dekretet, omtalt i dommen som SyRI-lovgivningen.²²⁷ Lovgivningen definerte SyRI som «en teknisk infrastruktur med tilhørende prosedyrer som brukes til å koble og analysere data i et sikkert miljø for å generere risikorapporter».²²⁸ En risikomodell ble definert som en modell med forutbestemte indikatorer på forhøyet risiko for misbruk av velferdsytelser eller brudd på skatte- og arbeidslovgivningen. En indikator ble definert som «enhver opplysning som sannsynliggjør en bestemt omstendighet». At indikatorene var forutbestemte, var i henhold til prinsippet om *velg før du samler* (select before you collect). Dette skulle forhindre vilkårlig innsamling av data og etterlevelse av GDPRs krav til formålsbegrensning og dataminimering.²²⁹ SUWI-dekretet inneholdt også en opplisting av hvilke data som kunne behandles i systemet. Det var i alt 17 overordnede datakategorier, om alt fra sysselsetting til utdanning, skatt, eiendom, trygdeytelser, gjeldsbyrde, bolig med mer.²³⁰

Under SyRI var de samarbeidende etater felles behandlingsansvarlige etter GDPR artikkel 26. Analysen foregikk ved to adskilte instanser. Databehandler var Stichting Inlichtingenbureau (IB), som var særskilt opprettet for formålet. IB hadde ansvar for å systematisere og pseudonymisere²³¹ dataene og prosessere dem i risikomodellen. Dersom modellen predikerte forhøyet risiko, ble saken flagget, de-pseudonymisert og sendt videre til en analyseenhet hos Inspectie SZW²³². Der bestemte de hvilke saker som skulle ettergås nærmere i et såkalt endelig risikoutvalg, blant annet basert på en utluking av falske positive. Denne delen av prosessen hadde altså en «human-in-the loop». For sakene til videre undersøkelse ble det opprettet en risikorapport som kunne lagres i inntil 2 år. Risikorapportene ble sendt til etatene i prosjektet, og i noen tilfeller politi og påtalemyndighet. For de flaggede sakene som ikke skulle følges opp videre måtte dataene slettes innen 4 uker. (Figur 23 nedenfor er et flytdiagram som gir oversikt over prosessen.)²³³

226 SUWI står for structuur uitvoeringsorganisatie werk en inkomen (arbeids- og inntektsstrukturimplementeringsloven). Loven, som er fra 2001, og skulle blant annet bidra til økt sysselsetting og redusert avhengighet av sosiale stønader. (<http://www.inspires-research.eu/innovativesocial-policy/22-Implementation-structure-work-and-income-%28SUWI%29>, oppslag 7.6.21).

227 Avsnitt 4.1.

228 Avsnitt 3.1.

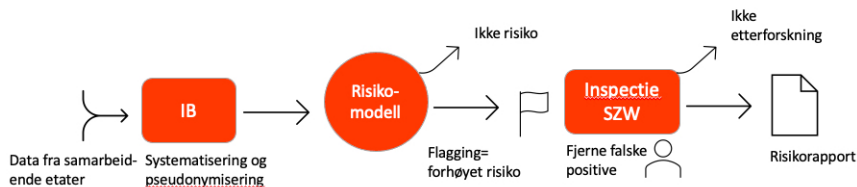
229 Avsnitt 4.23.

230 Avsnitt 4.17.

231 Pseudonymisering betyr at personopplysningene ikke lenger kan knyttes til en bestemt registrert uten bruk av tilleggsopplysninger, i dette tilfellet ved hjelp av en «key file» som IB opprettet.

232 Inspectie SZW er tilsynsmyndigheten under Sosial- og arbeidsdepartementet, som blant annet bistår kommunene i saker med trygdebedrageri.

233 Avsnitt 4.15, 4.28–4.31.



Figur 23: Modell av SyRI beslutningssystem

Myndighetene var pålagt gjennom SyRI-lovgivningen å annonsere oppstarten av et prosjekt offentlig. Kunngjøringen skjedde på Internett i en base for offentlige kunngjøringer. Det var imidlertid ikke noe krav om at befolkningen i de aktuelle områdene skulle informeres på individuelt grunnlag. Det ble opprettet et register for risikorapporter som det var mulig å kreve innsyn i, men personene det gjaldt ble ikke informert om at det var generert en risikorapport. Det ble heller ikke gitt informasjon til personer som fikk sine data analysert uten at det ble generert noen risikorapport. Bare dem som ble gjenstand for nærmere undersøkelser eller etterforskning på grunnlag av en risikorapport fikk underretning.²³⁴

Retten konsentrerte prøvingen om EMK art. 8, og viste til at beskyttelse av privatlivet innbefatter overholdelse av GDPRs prinsipper for beskyttelse av personopplysninger.²³⁵

Partene var enige om at innhenting og bruken av personopplysninger under SyRIs prosjekter innebar en innblanding i privatlivet, jf. EMK art. 8 første ledd. Det første spørsmålet for retten var derfor omfanget og alvorlighetsgraden av innblanding. Dette var krevende, fordi det var uklart hvordan SyRI-systemet egentlig virket. Saksøkerne hevdet at det var en dyplærende algoritme som analyserte stordata av ukjent opphav på en ustrukturert og vilkårlig måte. Staten sa på sin side at det var tale om en enkel beslutningstremodell som analyserte «eksisterende faktiske data» (bestående feitelijke gegevens) fra et begrenset utvalg offentlige etater med henblikk på å finne uregelmessigheter. Staten ønsket imidlertid ikke å dokumentere dette nærmere, fordi det kunne føre til at systemet ble manipulert og ikke lenger ville virke etter sin hensikt.²³⁶

²³⁴ Avsnitt 4.26 og 6.54.

²³⁵ Avsnitt 6.21.

²³⁶ Avsnitt 6.45–6.53.

Retten la etter en bevisvurdering til grunn at SyRI ikke genererte nye risikoprofiler under implementering, men at risikomodellen i seg selv var basert på profilering. Ettersom lovgivningen ikke stengte for at nye risikomodeller og indikatorer kunne bli utviklet i fremtiden, konkluderte retten med at SyRI var «på linje med» dyplærende, autonome algoritmer i utviklingsfasen, men ikke i implementeringsfasen.²³⁷

Partene var også uenige med hensyn til om SyRI var omfattet av GDPR artikkel 22 om automatiserte avgjørelser basert på profilering. Saksøkerne mente at det å opprette en risikorapport i seg selv var en automatisert enkeltavgjørelse i artikkel 22 sin forstand, fordi det å fjerne falske positive ved Inspectie SZW ikke utgjorde noen reell, meningsfull menneskelig involvering. Staten argumenterte med at det ikke var en automatisert avgjørelse. Den viste til at en risikorapport i seg selv ikke kunne få noen rettslig virkning overfor den registrerte, fordi den bare var utgangspunkt for nærmere undersøkelser. Uansett, hevdet staten, var de påkrevde garantier på plass for å ivareta personvernet.²³⁸

Retten mente at måten dataene ble koblet i SyRI var profilering i henhold til definisjonen i GDPR artikkel 4, men den tok ikke endelig stilling til om artikkel 22 kom til anvendelse. Selv om en risikorapport ikke førte til noen direkte rettslige konsekvenser for den registrerte, ville den uansett utgjøre et tyngende inngrep i privatlivet.²³⁹

For at et inngrep i privatlivet skal være lovlig i henhold til EMK 8 (2), må det ha et rettsgrunnlag som er tilstrekkelig tilgjengelig og klart til at borgerne kan forutberegne sin rettsstilling. I tillegg må det være på plass garantier (safeguards) som forhindrer misbruk og vilkårlighet. Hva slags garantier som er nødvendig kommer an på inngrepets art og omfang, slik at vurderingen av lovmessighet i stor grad flyter sammen med forholdsmessighetsvurderingen.²⁴⁰ Retten tok derfor ikke uttrykkelig stilling til om lovgivningen var tilstrekkelig klar og forutberegnelig, men gikk rett på vurderingen av nødvendighet og forholdsmessighet.²⁴¹

237 Avsnitt 6.51–6.54. Dette er litt uklart, men jeg forstår det som at retten fant at det var en modell basert på veiledet læring, også med tanke på at det var et krav om å velge kriterier før data ble samlet inn. Det var altså ikke en modell som lette etter anomalier i dataene på en måte som ikke var forhåndsdefinert (såkalt ikke-veiledet læring). Modellen var heller ikke dynamisk, det vil si at den ikke fortsatte å lære basert på nye data når den først var implementert i et prosjekt. (Se også avsnitt 2.2 for beskrivelse av læringsmetoder.)

238 Avsnitt 6.57–6.58.

239 Avsnitt 6.55–6.60.

240 Avsnitt 6.68–6.72, EMD: P.G. and J.H. v. UK, EMD: S. and Marper v. UK avsnitt 99.

241 Avsnitt 6.72.

Under kravet til nødvendighet slo retten fast at det aktuelle alternativet i dette tilfellet var «the economic well-being of the country», da det dreide seg om det nederlandske samfunnets økonomiske interesse i å sikre skatteinntekter og forhindre utbetaling av velferdsytelser til personer som ikke har krav på dem. SyRIs formål var dermed legitimt og svarte på et presserende samfunnsmessig behov (pressing social need). Anførselen fra saksøkerne om at man i stedet kunne ha strengere dokumentasjonskrav og grundigere saksbehandling som tiltak mot misbruk ble tilbakevist.²⁴²

Retten innledet forholdsmessighetsvurderingen med å si at staten har et særlig ansvar for balanseringen av interesser når den benytter ny teknologi til å samle og analysere persondata på denne måten. Det skyldes blant annet at det vil være vanskelig for den enkelte registrerte å overskue konsekvensene for sitt privatliv. Det er også en risiko for en nedkjølingseffekt, ettersom borgerne vil være mindre interessert i å dele sine data hvis de ikke stoler på at myndighetene respekterer deres privatliv.²⁴³

I avsnittene 6.87-6.106 trekker retten frem og drøfter en rekke mangler ved SyRI-lovgivningen, som samlet førte til en konklusjon om at det ikke var tilstrekkelige med garantier på plass til å rettferdiggjøre inngrepet i de registrertes privatliv. Disse avsnittene er særlig relevante for spørsmålet om krav til forklarbarhet.

For det første inneholdt ikke SyRI-lovgivningen informasjon om hvordan risikomodellen ble utviklet eller virket. Det var først i forbindelse med rettsaken at staten opplyste at risikomodellen besto av risikoindikatorer, koblinger av data og en terskelverdi. Det var ingen dokumentasjon tilgjengelig på hvordan man hadde kommet frem til indikatorene, eller hvordan de var kvalitetssikret. Staten hevdet at det var empirisk belegg for at de valgte indikatorene hadde reell prediktiv verdi, men dette ble ikke nærmere underbygget.²⁴⁴

SyRI-lovgivningen var også taus om hvilke spesifikke indikatorer modellen bygget på. Det var derfor umulig å vite hvilke faktiske omstendigheter som i et konkret tilfelle kunne føre til en konklusjon om forhøyet risiko. Det at staten under hovedforhandlingen hadde gitt noen eksempler på potensielle indikatorer var ikke tilstrekkelig.²⁴⁵ Terskelverdien ble heller ikke oppgitt, det vil si den numeriske grensen som avgjorde om en enkeltperson ville bli predikert med forhøyet risiko eller selektert ut.

²⁴² Avsnitt 6.73–6.77.

²⁴³ Avsnitt 6.84–6.85 og 6.5.

²⁴⁴ Avsnitt 6.87–6.88.

²⁴⁵ Avsnitt 6.87–6.88.

At det ikke var noen slik forutberegnelighet med hensyn til hvordan modellen virket eller informasjon til de registrerte, gjorde at de verken kunne ta til motmæle dersom det ble opprettet en risikorapport eller ha mulighet for å konstatere om en flagging hadde skjedd på korrekt grunnlag.²⁴⁶

Videre var det ikke konkretisert i lovgivningen hvilke data som inngikk i konkrete prosjekter. Selv om de tillatte datakilder var angitt i SUWI-dekretet, var det potensielle datagrunnlaget så vidt at det skulle vært spesifisert innenfor rammen av hvert prosjekt for å skape forutberegnelighet. Lovgivningen var dermed i strid med prinsippene om formålsbegrensning og dataminimering.²⁴⁷

Mangelen på gjennomsiktighet i lovgivningen gjorde det dessuten umulig å ta stilling til om risikoen for stigmatisering og diskriminering av beboere i nabolagene som ble tatt ut til prosjektene hadde blitt tatt høyde for og nøytralisert.²⁴⁸

Det institusjonelle oppsettet var ikke tilstrekkelig til å oppveie for manglende innsyn i modellens virkemåte og utfall. Blant annet burde en uavhengig, ekstern instans vurdert det enkelte prosjekt helhetlig, med tanke på hvilke data som skulle inngå. Det hadde blitt foretatt en personvernkonsekvensvurdering (DPIA) i forbindelse med vedtakelsen av SyRI-lovgivningen, men retten mente det burde skjedd i forkant av hvert enkelt SyRI-prosjekt, jf. GDPR artikkel 35 (1). Staten hadde videre gitt for sparsomt med opplysninger om hvordan Inspectie SZW organiserte seg internt og hva slags metoder som ble benyttet til å komme frem til det endelige utvalget av saker. Retten kunne derfor ikke vurdere effekten av at det var en «human-in-the loop» i denne delen av analyseprosessen.²⁴⁹

I denne saken ser man hvordan vurderingen av forholdsmessighet etter EMK art. 8 (2) forankres i prinsippet om transparens (åpenhet) i GDPR art. 5 første ledd bokstav a. Retten fant det ikke nødvendig å gå inn på enkeltbestemmelser i GDPR.²⁵⁰

Momentene som ble trukket frem i avsnitt 6.86 følgende viser at retten la svært mye i prinsippet om åpenhet, og at det i stor grad samsvarer med de ulike aspekter ved forklarbarhet som jeg beskrev i avsnitt 3.5.

For det første omfatter åpenhet offentlighet eller mulighet for innsyn i hvordan modellen virker: hvilke algoritmetyper som benyttes, hvordan modellen er trent

246 Avsnitt 6.90.

247 Avsnitt 6.97–6.98.

248 Avsnitt 6.93–6.94.

249 Avsnitt 6.94, 6.99, 6.101, 6.103–6.105.

250 Avsnitt 6.107.

og testet, hvordan terskelverdier er satt og hvilke data som inngår. (Datautvalget ble for øvrig også knyttet til prinsippene om dataminimering og formålsbegrensning i artikkel 5.) Dette er en global forklaring i utvidet forstand.

De registrerte har for det andre kontradiktoriske rettigheter når deres personopplysninger behandles av systemet. De har rett til å informeres om behandlingen, til å forstå modellen og grunnlaget for utfallet i sin sak. De må ha nok informasjon til å kunne ta stilling til om deres opplysninger blir behandlet korrekt av systemet, for eventuelt å kunne kreve retting eller å protestere. Med andre ord må de registrerte kunne forstå modellen på et overordnet (globalt) nivå, og de trenger lokal forklaring av enkeltprediksjonen i sin sak.

På det institusjonelle planet er det nødvendig med informasjon om hvordan beslutningssystemet er innrettet, hva som er spillet mellom automatiserte og menneskelige vurderinger og hvilke mekanismer som ivaretar kontroll. Det er altså nødvendig med en form for systembeskrivelse som viser hvordan systemet fatter avgjørelsene. Dette er spørsmål som også vil inngå i en personvernkonsekvensvurdering, jf. GDPR art. 35.

Retten foretok en samlet vurdering, som tilsier at det ikke hadde vært nødvendig for staten å innfri på alle disse punktene. Uavhengig kontroll eller tilsyn kunne kanskje kompensert for manglende offentlighet om hva de presise indikatorene i modellen var. Det kan derfor være mulig med kontrollsystemer uten full offentlighet om hvilke egenskaper som kan føre til at saker flagges.

Avgjørelsen indikerer også at det er mer betenkelig fra et privatliv- og rettssikkerhetssynspunkt å benytte modeller som er basert på ikke-veiledet læring eller modeller som er dynamiske og fortsetter å lære og tilpasse seg etter at de er satt i produksjon. Det vil være viktig for forutberegnelighet, etterprøvnbarhet og likebehandling at parameterne (som her: indikatorer på svindel) er tydelig definert på forhånd.

SyRI ble ikke vurdert opp mot nederlandsk forvaltningsrett, selv om det kunne vært naturlig med tanke på at retten pekte på mangelen på kontradiktoriske partsrettigheter. Dette kan skyldes måten saksøkerne valgte å ramme inn saken, men kan også skyldes at retten ikke fant det nødvendig.

Det er en viss spenning i synet på hvor langt personvernet rekker med hensyn til det materielle innholdet i avgjørelser. I denne saken kom domstolen til at det hører til prinsippet om åpen behandling og retten til privatliv at den registrerte kan få den nødvendige informasjon til å vurdere riktigheten i avgjørelsen og

kunne klage. Dette er også i tråd med Artikkel 29-arbeidsgruppens vurdering. EU-domstolen har imidlertid lagt et mer begrenset syn til grunn og i en del saker sagt at personvernet er prosessuelt, og ikke materielt. Det gjelder innsamlingen og behandlingen av data og den registrertes kontroll over det, men ikke utfallet av behandlingen. Eventuelle klager på avgjørelsens riktighet må håndteres ved at man klager etter den aktuelle sektorlovgivningen.²⁵¹

5.2.4 Oppsummering

Avsnitt 5.2 har altså vist at det hersker uklarhet med hensyn til betydningen av GDPR artikkel 22 for retten til forklaring av algoritmiske enkeltavgjørelser. Det ser ut til at det er enighet om at 13 (2) f, 14 (2) g og 15 (1) h innebærer en rett til en global forklaring i snever forstand av modellens virkemåte, men retten gjelder ved automatiserte avgjørelser og det ser ut til at det byr på vanskeligheter å avgjøre om et system fatter hel- eller bare delautomatiserte avgjørelser. SyRI-dommen bidrar heller ikke til noen klargjøring på dette punktet, fordi retten unnlot å ta endelig stilling til partenes anførsler om spørsmålet.

SyRI-dommen taler derimot for at GDPRs prinsipp om åpen og rettferdig behandling i artikkel 5 (a) kan strekkes svært langt med tanke på å stille krav til forklarbarhet i algoritmiske beslutningssystemer. Det gjelder særlig dersom systemet behandler store mengder personopplysninger i profiler og / eller benyttes til kontrollformål hvor rettssikkerhet og kontradiksjon for de registrerte er sentralt.

Dommen har ingen nevneverdig rettskildemessig vekt i norsk sammenheng, men den er likevel viktig fordi den er et eksempel på en detaljert redegjørelse for hvordan et algoritmisk beslutningssystem er innrettet. Den gir en rettslig analyse av mange aspekter ved systemet: Dets utvikling, anvendelse, konsekvensene for de registrerte samt den institusjonelle konteksten. Dersom NAV skulle vurdere å iverksette et tilsvarende system måtte et eventuelt behov for å holde tilbake forklaringer fra offentligheten og de registrerte oppveies med kompenserende tiltak som eksternt og uavhengig kontroll og betryggende saksbehandling for personer som trekkes ut for kontroll.

Jeg vil nå se mer spesifikt på retten til kontradiksjon ved algoritmiske avgjørelser.

251 Wachter og Mittelstadt (2019) s. 36.

5.3 Retten til kontradiksjon ved algoritmiske avgjørelser

En viktig side ved retten til kontradiksjon er partens mulighet for å ettergå og bestride forvaltningens beslutningsgrunnlag og vurderinger.

GDPR inneholder som beskrevet i avsnitt 5.1.2 flere bestemmelser som skal ivareta den registrertes prosessuelle rettigheter i forbindelse med behandling av personopplysninger. Hvor langt rettighetene strekker seg med hensyn til det materielle innholdet i avgjørelsene er det som sagt i forrige avsnitt uenighet om.

Forvaltningsloven er ikke på samme måte som GDPR formulert med tanke på automatiserte systemer eller algoritmisk forvaltning. Den er i hovedsak tilpasset menneskelig, dokumentbasert enkeltsaksbehandling. Reglene er imidlertid utslag av et grunnleggende prinsipp om at parten har rett til å forstå grunnlaget for en avgjørelse med tanke på å ivareta sine interesser i saken. Reglene om veiledning, partsinnsyn og begrunnelse i §§ 11, 18 og 23–25 ivaretar ulike aspekter ved dette grunnhensynet. Avhengig av hvordan et beslutningssystem er innrettet, hvilket tidspunkt informasjonen skal gis, hvor inngripende avgjørelsen er og prediksjonens betydning for den samlede avgjørelsen, vil det være påkrevd med lokale og globale forklaringer, så vel som systembeskrivelser (jf. også avsnitt 3.5)

Jeg har ikke eksempler på at et algoritmisk beslutningssystem har blitt vurdert opp mot forvaltningslovens regler, men saken om avgangskarakterer for IB-elevene fra 2020 er et eksempel fra norsk kontekst som blant annet angår retten til partsinnsyn og begrunnelse. Karaktersetting er avgjørelser med stor velferdsmessig betydning for de berørte, og muligheten for kontradiksjon er derfor viktig.

5.3.1 Algoritmisk karakterfastsettelse ved IB og Datatilsynets intervensjon

Våren 2020 ble det ikke avholdt skriftlig eksamen på grunn av koronaepidemien. For elevene ved IB utgjør eksamenskarakteren normalt en stor andel av sluttkarakteren i faget. For å kompensere for det manglende vurderingsgrunnlaget utviklet IB en form for algoritmisk beslutningssystem for å sette karakterene. I systemet inngikk tre komponenter som til sammen ga sluttkarakteren. Disse var: karakterene på elevens skriftlige innleveringer, lærerens prediksjon for eksamenskarakter for eleven og en justerende faktor basert på skolens historiske resultater. Det var den tredje komponenten («school context») som konflikten dreide seg om.

IB-stiftelsens administrasjon (IBO – International Baccalaureate Organization) opplyste om metoden for karakterfastsettelse på sine nettsider i mai 2020. Her

understreket IBO de at det var tale om en «awarding model» og ikke en «computer-based algorithm».²⁵²

I en senere redegjørelse fra juli 2020 opplyste IBO at skolekontekst-komponenten besto av en beregning av en «unik faktor» for hvert fag på hver enkelt skole, som ble brukt til å justere karakterene. Den unike faktoren var et mål på graden av samsvar mellom karakter på innleverte arbeider og eksamenskarakter for tidligere årskull ved skolen.²⁵³ Tanken var altså at man kunne korrigere for systematiske forskjeller i karaktergivning mellom skoler, slik at resultatet for elevkullet samlet sett ville bli mer rettferdig. Systemet førte til at mange elever følte seg urettferdig behandlet, fordi de fikk lavere karakterer enn de hadde forventet uten at det ble gitt noen begrunnelse. Flere elever ba om innsyn i sine data og logikken bak beslutningssystemet, noe de ikke fikk.

Datatilsynet engasjerte seg i saken, og ba i brev til IBO 20.7.20 om en nærmere redegjørelse for systemet, med en beskrivelse av hvilke data som inngikk, deres vektning og utregningene. De ble også bedt om å uttale seg om hvorvidt karakterfastsettelsen ble rammet av GDPR art. 22 og om de anså at behandlingen var i overensstemmelse med prinsippene om åpenhet og rettferdighet i artikkel 5 (1) (a), dataminimering (c) og riktighet (d).

Datatilsynet gikk videre med saken, og forhåndsvarslet IBO 7.8.20 om at de vurderte å fatte vedtak med krav om retting. I sin vurdering viste Datatilsynet til at karakterer må anses som personopplysninger i henhold til GDPR art. 4 (1). I strid med det IBO selv hadde anført, la Datatilsynet til grunn at systemet involverte en algoritme som utførte profilering, jf. art. 4 (4). Når historiske data brukes til å sette karakterer, tilskrives elevene egenskaper basert på tidligere årskull fremfor at de gis en reell individuell vurdering.

Systemet ble så vurdert opp mot prinsippene om rettferdighet, åpenhet og riktighet, jf. GDPR art 5 (1) bokstav a og d. Under rettferdighet og riktighet sa Datatilsynet at elevene hadde rett til å få sine data behandlet på en måte som var i tråd med deres forventninger. Deres forventning var at de ville bli vurdert basert på sine kvalifikasjoner, ikke resultatene til tidligere elever ved skolen. Det var da også tvil om de fastsatte karakterene var riktige når de delvis var basert på en prediksjon om eksamenskarakter basert på historiske data. Å trekke inn

252 <https://www.ibo.org/globalassets/news-assets/coronavirus/assessment-model-letter-may-2020-en.pdf>, oppslag 28.4.21.

253 <https://ibo.org/news/news-about-the-ib/awarding-may-2020-results-further-information>, oppslag 30.4.21.

skolekontekst representerte dessuten vektlegging av elevenes bosted, som medførte en risiko for forskjellsbehandling basert på sosioøkonomisk status.²⁵⁴

Behandlingen var videre ikke i tråd med prinsippet om åpenhet, fordi elevene ikke hadde fått innsyn i modellen. En avgjørelse om karakterfastsettelse har stor betydning for elevenes fremtid, hvilket tilsier skjerpede krav til åpenhet. IBO skulle derfor ha gitt innsyn i detaljer om modellen og faktorenes vektning («concrete details about the model and the weighting of factors therein»).

Datatilsynet konkluderte etter dette med at behandlingen av elevenes personopplysninger ikke var tilstrekkelig åpen, rettferdig eller korrekt, og at IB måtte ta skolekomponenten ut av karakterfastsettelsen.

I sitt tilsvare 21.8.20 gjorde IBO blant annet gjeldende at Datatilsynet ikke var rett myndighet i saken, men det britiske Information Commissioner's Office (ICO). IBO mente dessuten at GDPR ikke var et egnet rettslig rammeverk, fordi prosessuelt riktig og rettferdig databehandling er noe annet enn materielt riktig og rettferdig karaktersetning. Her ser man igjen at uenigheten om GDPRs rekkevidde gjør seg gjeldende.²⁵⁵

Etter hva jeg kan se er det ikke fattet noe vedtak i saken, og jeg vet ikke om det skyldes at Datatilsynet kom til at de likevel ikke hadde kompetanse eller om det var fordi IBO trakk algoritmen og benyttet en annen kalkuleringsmetode. Uansett er Datatilsynets analyse interessant som et eksempel på en rettslig vurdering av manglende forklarbarhet, i betydningen innsyn og begrunnelse, i algoritmiske avgjørelser. Den egner seg også til å demonstrere de to nivåene av global forklaring og hvordan de henger sammen med muligheten for kontradiksjon, kontroll og ansvarliggjøring av aktøren som iverksetter systemet (se også avsnitt 3.5).

Dersom man legger til grunn IBOs opplysning om at det var en unik faktor per fag per skole ville modellen oppføre seg likt for alle elevene på en enkeltskole. Dermed får man en global modell, og global og lokal forklaring sammenfaller (se også avsnitt 3.4.1 og figur 12).

254 En slik effekt ble rapportert for den britiske ekvivalenten, Ofquals kalibreringsmodell for A-level karakterer. Den endte med å bli trukket, slik at standpunktkarakterene ble gjeldende. Når skolens historiske resultater fra tidligere år relativt sett fikk større betydning i karakterfastsettelsen, falt det uheldig ut for sterke elever fra sosialt belastede områder. De fikk sine standpunktkarakterer nedjustert i større grad enn elever fra skoler med høyere snittresultat, ikke minst privatskoler. (The Guardian 19.8.20.)

255 Jf. Wachter og Mittelstadt (2019 s. 6) og avsnitt 5.1.3 om SyRI.

En global forklaring i mer snever forstand ville innebære å gi elevene en begrunnelse som opplyste hva den unike faktoren var, og hvordan de tre komponentene hadde blitt vektet i den samlede karakteren. Med denne informasjonen kunne elevene selv ha kontrollert sine resultater. Dersom de oppdaget feil i utregningene, kunne de klaget på det grunnlaget. Det er dette man vanligvis forbinder med kontradiksjon, og det skulle elevene hatt som et minimum.

I denne saken var det også bekymring om algoritmen gjorde at elevens bosted eller valg av skole kunne påvirke karakteren på en måte som var usosial eller som kunne bidra til å forsterke sosiale forskjeller. Datatilsynet bestred dessuten selve grunnpremisset for modellen: at tidligere årskulls resultater kan inngå i beregningen av enkeltelevens slutt karakterer.

Dersom tilliten til beslutningssystemet svikter eller det er grunn til å betvile modellens forutsetninger, er det nødvendig med en mer utvidet global forklaring som dokumenterer de vurderinger og valg som er tatt ved systemets utvikling. I dette tilfellet skulle IBO ha gjort tilgjengelig en nærmere begrunnelse for hvorfor de anså tidligere årskulls eksamensresultater som relevante og riktige data å bruke. De skulle også dokumentert at risikoen for utilsiktede skadelige virkninger, som diskriminering, vilkårlighet eller annen skjevbehandling, hadde blitt tilbørlig hensyntatt og håndtert.

Eksempelet viser også at det er viktig for tilliten at beslutningssystemets innretning blir gjort offentlig i form av en systembeskrivelse. I dette tilfellet førte manglende åpenhet til tvil og usikkerhet om karakterfastsettelsen var en helautomatisert prosess, eller om det var en «human-in-the-loop» som foretok en kvalitetssikring til slutt.

På samme måte som domstolen i SyRI-saken innebærer Datatilsynets anvendelse av prinsippet om åpen og rettferdig behandling utstrakte krav til beslutningssystemets forklarbarhet. Det blir også sagt direkte jo viktigere avgjørelsen er for den registrerte, desto viktigere blir det med åpenhet (det vil si: at systemet er forklarbart).

Saken er også et eksempel på at GDPR artikkel 22 er vanskelig å håndtere, fordi man ved systemer som er lite tilgjengelige eller transparente er prisgitt informasjonen som den behandlingsansvarlige gir. Datatilsynet ba om en redegjørelse for IBOs syn på om systemet var omfattet av artikkel 22, men da de ikke fikk noe svar på det punktet lot de spørsmålet ligge.

Jeg kommer tilbake til dette eksempelet i avsnitt 5.5 om algoritmisk diskriminering. I det følgende vil jeg foreta en vurdering av lokale forklaringsmetoders nytteverdi for algoritmiske forvaltningsavgjørelser.

5.4 Betydningen av lokale forklaringsmetoder

I SyRI- og IB-sakene ble det ikke gitt noen forklaring på sakens utfall for de berørte, som var én av grunnene til at disse systemene ikke ble ansett som tilstrekkelig åpne og rettferdige. Det er imidlertid teknisk mulig å gi individuelle forklaringer, også fra komplekse maskinlæringsmodeller. I avsnitt 3.4.2 beskrev jeg noen metoder for å forklare enkeltprediksjoner, og i avsnitt 3.4.3 gikk jeg i mer detalj inn på hvordan en approksimasjon med Shapley-verdier utføres.

Jeg skal nå se nærmere på om slike lokale forklaringsmetoder, i samspill med andre tiltak som gode systembeskrivelser og globale forklaringer, vil være egnet som bidrag til rettssikkerhet og kontradiksjon i algoritmisk forvaltning. Perspektivet er både på automatiserte avgjørelser og beslutningsstøtte (se også avsnitt 2.6 og figur 4 og 5). I det første tilfellet skal forklaringen inngå i en begrunnelse som må overholde forvaltningslovens krav, mens den i det andre tilfellet skal bidra til forsvarlig utredning og avveining av sakens momenter. Jeg begrenser drøftelsen til approksimasjoner og kontrafaktiske forklaringer som er metoder fremhevet i den tekniske litteraturen.

5.4.1 Lokale approksimasjoner og vedtaksbegrunnelser

Forvaltningsloven § 24 første ledd sier at enkeltvedtak skal ha begrunnelse. Det vil også gjelde for en avgjørelse fattet av en algoritme. Begrunnelsens mest sentrale funksjon er å bidra til at parten forstår sakens utfall og kan akseptere og innrette seg etter avgjørelsen. Den må derfor være utformet på en måte som oppleves relevant og meningsfull for parten.

Approksimasjoner er ikke umiddelbart tilgjengelige og forståelige i sin ubearbeidede form; de må formidles til parten i saken. Som vist i avsnitt 3.4.5 i eksempelet fra NAV, kan lokale forklaringer tilpasses konteksten og partens behov. Ved hjelp av interaksjonsdesign kan man utforme begrunnelser som kan være mer interaktive eller klikkbare, hvor det er mulig å få forklaringer på ulike aspekter ved modellen eller prediksjonen. Algoritmiske begrunnelser kan dermed tilfredsstille krav til klarspråk (se avsnitt 4.2.5), og kanskje bidra til å utvikle forvaltningsrettens perspektiv på hvordan begrunnelser kan se ut.

Begrunnelsen må videre inneholde den informasjon parten trenger for å konstatere at vedtaket er riktig, eller for å kunne påvise feil som kan rettes opp og kanskje lede til omgjøring. Den er utgangspunktet for å formulere en klage, og for vurderingen som foretas av klageorgan eller domstol. Begrunnelsen er derfor sentral for partens mulighet for å utøve sin rett til kontradiksjon og for kontrollen med forvaltningen.

For avgjørelser som inneholder skjønnbaserte vurderinger gjelder forvaltningsloven § 25 tredje ledd som et minimumskrav. Tredje ledd første punktum sier at «De hovedhensyn som har vært avgjørende ved utøving av forvaltningsmessig skjønn, bør nevnes.» Ordlyden indikerer at det ikke er nødvendig å omtale alle momenter av betydning, men dette kravet er nyansert og skjerpet gjennom rettspraksis. Begrunnelsen må vise at alle relevante momenter er vurdert, og dessuten må den tilpasses vedtakets betydning. Jo mer inngripende vedtak, jo mer detaljert begrunnelse kreves.²⁵⁶ At de viktigste momentene i vurderingen skal med i begrunnelsen gjør at parten kan vurdere om forvaltningen har holdt seg innenfor det saklige og rimelige i skjønnsutøvelsen, eller om den har overskredet sin myndighet.²⁵⁷

Den Shapley-baserte metoden er en lokal approksimasjon som gir en tilnærmet eller omtrentlig påvisning av de faktorer som har virket inn på prediksjonen og deres relative vekt. En matematisk beregnet approksimasjon kan potensielt være en presis refleksjon av hva som har vært utslagsgivende for utfallet, og den minner om strukturen i en forvaltningsbegrunnelse. Forklaringens presisjons- og detaljnivå kan dessuten reguleres gjennom antallet variabler som tas inn i begrunnelsen som gis til parten. Det kan tale for at approksimasjoner kan fungere som del av begrunnelsen for et vedtak.

Approksimasjoner har imidlertid noen utfordrende sider, som kan gjøre det problematisk å benytte dem i automatiserte vedtaksbegrunnelser. For det ene er det en risiko for at approksimasjonen for noen enkelttilfeller ikke stemmer, det vil si at den forenklete funksjonen er for unøyaktig til at den gjenspeiler den underliggende modellen. Da vil det heller ikke være samsvar mellom begrunnelsen og de variablene som reelt sett har spilt inn i avgjørelsen.²⁵⁸

256 Graver (2015) s. 459, med henvisning til Rt. 2000 s. 1056 og Rt. 1981 s. 745. Man kan ha som generelt utgangspunkt at velferdsforvaltningens vedtak er viktige og inngripende for parten. Ofte gjelder det muligheten for å kunne forsørge seg selv (Kane 2018 s. 315).

257 Den ulovfestede myndighetsmisbrukslæren angir rammene for forvaltningsskjønnet. Den omhandler utenforliggende hensyn, vilkårlighet, grov urimelighet og usaklig forskjellsbehandling. (NOU 2019: 5 s. 545.) Innenfor velferdsretten utgjør også det velferdsrettslige minstemål en skranke for forvaltningens skjønn.

258 Wachter et al. (2018) s. 12.

Approximasjoner reiser også spørsmål knyttet til saklighet og relevans. Det er ikke nødvendigvis fullstendig samsvar mellom prediktiv relevans og rettslig relevans. Man kan tenke seg egenskaper som har prediktiv verdi og som kan bidra til å øke modellens treffsikkerhet dersom de tas inn i modelleringen, men som vil bli oppfattet som lite saklige om de skulle inngå som momenter i en automatisert begrunnelse.²⁵⁹

Regler om dataminimering (GDPR) og et relevanskrav i den kommende KI-lovgivningen fra EU²⁶⁰ vil kanskje gjøre at problemet ikke er så uttalt, men jeg er usikker på om det vil løse det. Dersom man går for langt i å begrense datamengden gjennom saklighets- og relevansvurderinger vil jo det også svekke modellens potensiale. Noe av hensikten med teknologien er å oppdage sammenhenger i dataene som man ikke så for seg på forhånd.

Et annet problem for automatiske begrunnelser basert på approximasjoner kan være stedfortredende variabler (proxy-variabler, se forklaring i avsnitt 3.3.4). Dersom de avgjørende variablene egentlig er stedfortredende for en annen egenskap, vil den Shapley-baserte begrunnelsen ikke gjenspeile hva som er den egentlig utslagsgivende egenskapen. Selv om avgjørelsen er riktig innholdsmessig, kan begrunnelsen gi inntrykk av at det er lagt vekt på utenforliggende hensyn.²⁶¹

Det kan hende disse problemene er overdrevet og kan håndteres. Det vil for eksempel være en sammenheng mellom kvaliteten på arbeidet som er gjort i modellutviklingen og kvaliteten på forklaringene. Dersom det er gjort et godt arbeid for å rense opp i data og sikre at de er relevante og godt merket, vil det også gi mer logiske og forståelige forklaringer til brukeren.

Basert på min forståelse kan det likevel se ut til at lokale approximasjoner er betenkelige å bruke i automatiserte vedtaksbegrunnelser til parten. Det kan innebære at man velger mer intuitivt forståelige modeller, hvor det er mulig å følge algoritmens «resonnement» uten at man trenger å ty til egne forklaringsmodeller (se også avsnitt 3.2.2). Komplekse modeller bør kanskje forbeholdes

259 Oswald (2018) s. 10, Sutterud og Topper (samtale 3.9.20).

260 Artikkel 10 (3) 1. punktum i EU-kommisjonens forslag til regulering av KI-systemer lyder for eksempel: «Training, validation and testing data sets shall be *relevant*, representative, free of errors and complete.» (min utheving). I den nasjonale strategien for kunstig intelligens står det i tilknytning til prinsippet om dataminimering at «[U]tgangspunktet må være at virksomheten skal velge ut et relevant utvalg av data, og en datamengde som er tilstrekkelig stor.» (Kommunal- og moderniseringsdepartementet 2020 s. 57.)

261 Cobbe (2018) s. 25, Sutterud og Topper (samtale 3.9.20).

avgjørelser hvor begrunnelse ikke er påkrevd, eller hvor prediksjonen utgjør en mindre komponent i avgjørelsen.

5.4.2 Kontrafaktiske forklaringer og vedtaksbegrunnelser

Kontrafaktiske forklaringer er som nevnt mye omtalt i litteraturen om forklarbarhet. Dette er en forklaringstype som er beregnet på å informere brukeren om hva som må endres for å få et ønsket utfall.

Wachter et al. fremhever kontrafaktiske forklaringer som et godt alternativ til approksimasjoner. De er basert på perturbasjoner og er enkle å få til teknisk (se avsnitt 3.3 for en forklaring av perturbasjoner). De mener kontrafaktiske forklaringer kan hjelpe brukeren til å forstå og eventuelt bestride avgjørelsen, eller gjøre endringer med tanke på å oppnå et ønsket utfall i fremtiden.²⁶²

Jeg mener dette standpunktet ikke kan gjelde for forvaltningsavgjørelser. Den kontrafaktiske forklaringen er ikke noen begrunnelse i forvaltningsrettslig forstand. Den er ikke innrettet for å opplyse om hvilke momenter som har vært avgjørende i vurderingen eller for å rettferdiggjøre et resultat, men for å si hvilke egenskaper parten mangler. Selv om man i en og samme avgjørelse kan få flere kontrafaktiske forklaringer som belyser ulike sider ved utfallet, blir dette biter av informasjon som parten selv må sette sammen til en antakelse om hva som samlet sett er grunnlaget for avgjørelsen. Det vil ikke gi et godt utgangspunkt for å oppdage opplysninger som må korrigeres eller vurderinger som kan være gale. Det vil også gjøre det vanskelig å utforme en godt begrunnet klage.

Kontrafaktiske forklaringer kommer i konflikt med kontradiktoriske rettigheter og innholdskravet i forvaltningsloven § 25, og kan derfor ikke benyttes til en automatisk generert begrunnelse for vedtak. De har likevel betydelig nytteverdi, fordi de er spisset inn mot forhold som parten kan endre på. Slik sett er de brukervennlige. Man kan se for seg denne type begrunnelse for avgjørelser som ikke kan påklages eller som supplement til en begrunnelse som for øvrig tilfredsstiller forvaltningslovens krav.

5.4.3 Lokale forklaringer i forbindelse med beslutningsstøtte

Når maskinlæringsmodellen benyttes som beslutningsstøtte i en saksbehandlingsprosess må behovet for forklarbarhet, så lenge det ikke finnes mer spesifikke regler, vurderes i henhold til ulovfestede krav til forsvarlig saksbehandling eller forsvarlig forvaltning.

²⁶² Wachter et al. (2018) s. 13 flg. Artikkelen har tittelen «*Counterfactual Explanations without Opening the Black Box*» som henspiller på at det ikke er nødvendig å prøve å forklare modellens indre logikk («matematisk transparens») når det finnes en metode for å belyse modellens utfall som er mer forståelig og brukervennlig.

Forsvarlighet kan i Gravers fremstilling både vise til forvaltningens fremgangsmåte og innholdet i avgjørelsene. Saken må utredes tilstrekkelig slik at alle vesentlige sider er belyst, og de ulike momentene må vektes mot hverandre på en forsvarlig måte før avgjørelsen fattes.²⁶³

Boe mener dessuten at det gjelder et alminnelig krav til forsvarlig forvaltning, som blant annet omfatter måten virksomheten er organisert og oppgavene fordeles og løses. Også krav til de ansattes kompetanse inngår her.²⁶⁴

Flere juridiske forfattere har uttrykt bekymring for at algoritmisk støttede beslutningsprosesser medfører en risiko for at saksbehandlere legger overdreven vekt på prediksjonene og dermed lar modellene overstyre sin erfaring og fagkunnskap. Det omtales som *automation bias* og *rubber stamping*. Det kan også være en fare for at krav om effektivitet i saksbehandlingen gjør at tilbøyeligheten til å gjøre selvstendige vurderinger eller supplere grunnlaget med andre utredningsskritt reduseres.²⁶⁵ Et annet mulig problem er at saksbehandleren følger prediksjonene der de bekrefter egen oppfatning av saken, men ser bort fra dem dersom det er motstrid.²⁶⁶ Man kan se på dette som potensielle feil ved skjønnsutøvelsen hvor de ulike hensynene ikke vurderes og avveies forsvarlig i den samlede avgjørelsen, eller alternativt som et utslag av at beslutningssystemet ikke er innrettet forsvarlig.²⁶⁷

Risikoen for slik svikt i skjønnsutøvelsen kan motvirkes ved at saksbehandleren har god forståelse av maskinlæringsmodellen og er generelt fortrolig med teknologien. Relevant informasjon vil antakelig være kjennskap til hvilke variabler som inngår, modellens formål og begrensninger. Dersom modellens treffsikkerhet er lavere for enkelte grupper vil det også være relevant kunnskap. Det kan tilsi at saksbehandleren må tilskrive prediksjonen mindre vekt i saker som angår personer fra disse gruppene.

Momentene jeg trekker frem her er til dels sammenfallende med kompetansekrav til dem som skal stå for «Human Oversight» i artikkel 14 (4) i EU-kommisjonens forslag til regulering av høyrisiko KI-systemer. Det kan se ut som at

263 Graver (2015) s. 88.

264 Boe (2018) s. 155.

265 Allhutter et al (2020) beskriver uklare forventninger til saksbehandlerne i forbindelse med AMS-algoritmen for kategorisering av arbeidssøkende. Til å begynne med var modellens plass i organisasjonen og beslutningssystemet over hodet ikke tematisert. Det som senere ble kommunisert av forventninger var motstridende. På den ene siden skulle algoritmen være et tungt og objektivt argument i vurderingen, på den andre siden ble det sagt at den bare var en «second opinion», og at det var viktig at saksbehandlerne brukte sitt faglige skjønn til å utfylle modellen.

266 Olsen et al. (2019) s. 3, Algorithmwatch (2020) s. 53, Prabhu (samtale 5.1.21).

267 Oswald (2018) s. 14, Cobbe (2018) s. 9.

bestemmelsen er ment å gjelde for en mer teknisk rolle, men flere av de samme behovene gjør seg etter min oppfatning gjeldende for en saksbehandler som benytter en maskinlæringsmodell som beslutningsstøtte. Etter bokstav a) må denne personen forstå systemets muligheter og begrensninger (også med tanke på å fange opp feil), b) kjenne til risikoen for *automation bias*, c) ha forutsetninger for å fortolke systemets utfall basert på dets egenskaper, d) kunne avgjøre i et enkelttilfelle å se bort fra eller overstyre utfallet og e) kunne stoppe systemet.²⁶⁸

Et eksempel på en maskinlæringsmodell anvendt som beslutningsstøtte er NAVs modell for prediksjon av sykefraværets lengde, beskrevet i avsnitt 3.4.5. I prosessen med å finne frem til hensiktsmessige forklaringer fikk veilederne først bare presentert en omskriving av selve prediksjonen, som sa «Stor sannsynlighet for friskmelding innen uke 28» (figur 19). Dette var ikke tilstrekkelig for veilederne, antakelig fordi de manglet informasjon til å vurdere prediksjonen i lys av sin øvrige kjennskap til saken og generelle erfaring fra fagområdet. Manglende forklaring gjorde dermed at modellens nytteverdi som beslutningsstøtte ble begrenset.

I det neste forsøket fikk veilederne en SHAP-forklaring som viste vektingen av momenter i enkeltprediksjonen sammen med utfyllende informasjon om modellen (figur 20). De fikk da vite hvilke egenskaper eller faktorer som hadde vært utslagsgivende i det konkrete tilfellet, og denne informasjonen kunne holdes opp en mer generell forklaring om modellen, som beskrev dens formål og hvilke variabler som inngikk. Antakelig ville det også vært nyttig å vite noen mål på modellens treffsikkerhet for ulike diagnoser eller persongrupper. Med kombinasjonen av lokal og global forklaring økte man altså maskinlæringsmodellens potensiale for å supplere veilederens skjønnsutøvelse i enkeltsaker.

Dette understreker at lokale forklaringer, uten global eller kontekstuell forståelse av modellen, i de fleste tilfeller ikke vil være tilstrekkelig for forsvarlig skjønnsutøvelse.

Lokale approksimasjoner basert på Shapley-verdier fremstår her som en godt egnet lokal forklaringsmetode. De rettsikkerhetsmessige utfordringene som gjorde seg gjeldende for vedtaksbegrunnelser er mindre påtrengende, fordi saksbehandlerne vil kunne fange opp og korrigere for feil og vilkårlighet i prediksjonene dersom de har en overordnet forståelse av modellens virkemåte og begrensninger. Forklaringen vil også sette veilederne i stand til å begrunne en beslutning om dialogmøteinnkalling overfor den sykmeldte, dersom det er behov for det.

²⁶⁸ EU-kommisjonen (2021).

Dette kan innebære at det kan være forsvarlig å benytte komplekse modeller til beslutningsstøtte i velferdsforvaltning. En av forutsetningene er at systemets brukere, som saksbehandlere og andre funksjoner i organisasjonen, har tilstrekkelig kompetanse til å fortolke og anvende modellens resultater.

Når det gjelder kontrafaktiske forklaringer, tviler jeg på at de kan fungere i en situasjon hvor modellen skal være et beslutningsgrunnlag og utgangspunkt for skjønnsutøvelse. De kan derimot bidra til å gi økt forståelse for modellen ved å vise hvordan den opptrer i ulike situasjoner, eller bidra til likebehandling gjennom å vise saksbehandleren hva utfallet ville blitt dersom man justerer inngangsvariablene. De kan også være nyttige i en veiledningssituasjon.

5.4.4 Menneskelige versus algoritmiske begrunnelser

Enkelte mener algoritmisk saksbehandling har et potensiale for å gi mer transparente og betryggende prosesser, uten den forutinntatthet, vilkårlighet og slurv som kan prege manuelle avgjørelser. Menneskelige beslutningstakere er også svarte bokser, både bevisst og ubevisst.²⁶⁹

For eksempel er det ikke nødvendigvis slik at rettsanvenderen identifiserer og vektet alle relevante momenter i tur og orden, for så å nå en konklusjon. En begrunnelse i en manuelt behandlet sak vil ofte formuleres etter at avgjørelsen er tatt. Når man rettferdiggjør i etterkant, kan momentenes utvalg og vekting bli annerledes enn det som reelt sett lå til grunn.²⁷⁰

Bias og forutinntatte holdninger kan også påvirke menneskelige vurderinger, og kan være vanskeligere å avdekke hos et menneske enn hos en algoritme. Mennesker kan i tillegg bli påvirket av tilfeldige faktorer, som sinnsstemning, vær eller hvor lang tid som har gått siden siste måltid. Kahneman et al. mener slik «støy» er en betydelig kilde til vilkårlighet i avgjørelser. Ettersom dette er ubevisste mekanismer, vil de heller ikke synes i begrunnelsen.²⁷¹

Slike betraktninger gjør at Olsen et al. advarer mot å stille strengere krav til etterrettelighet og presisjon i algoritmiske begrunnelser enn det som er «the human standard» – den menneskelige standard som er etablert i forvaltningsretten.²⁷²

269 Coglianese og Lehr (2017) s. 1186, Castelluccia og Le Métayer (2019) s. 24, Olsen et al. (2019) s. 10, Oswald (2018) s. 6.

270 Oswald (2018) s. 6. Dette omtales i den juridiske metodelære som fasadelegitimasjonstesen. (Frøberg 2019 s. 511).

271 Kahneman et al. (2016).

272 Olsen et al. (2019) s. 10.

Jeg er usikker på hvor nyttig og relevant det er å sammenligne en menneskelig og en algoritmisk standard. Forvaltningslovens regler om begrunnelse er formulert med tanke på et menneskelig resonnement, der begrunnelsen skal gjenspeile en avveiningsprosess og rettferdiggjøre resultatet. Maskinlæringsalgoritmer foretar ikke avveininger eller resonnementer, de gjør sannsynlighetsberegninger og foretar slutninger basert på data. Avgjørelser som bygger på slike slutninger, kan da heller ikke begrunnes i forvaltningslovens egentlige forstand. De bør heller dokumenteres og forklares ved at man viser hvilke data som behandles og hvordan.

Et eksempel på en lovregel om forklaring av algoritmiske avgjørelser finnes i den franske bestemmelsen R311-3-1-2, som har hjemmel i loven om digital forvaltning (*Loi n°2016-1321 pour une République numérique*). Den sier at:

Forvaltningen, når den fatter en algoritmisk avgjørelse rettet mot en enkeltperson, på partens forespørsel, i en form som er forståelig, og uten å krenke lovpålagt taushetsplikt, må gi informasjon om:

- 1) graden og måten den algoritmiske behandlingen har bidratt til avgjørelsen
- 2) hvilke data som er behandlet og deres kilde
- 3) behandlingsparametrene, og der det er aktuelt, deres vektning i det konkrete tilfellet, og
- 4) hvilke operasjoner som inngikk i behandlingen ²⁷³

Bestemmelsen er interessant å se hen til fordi den er formulert med tanke på algoritmenes funksjonsmåte. Den sier uttrykkelig at forklaringer som forvaltningen gir til parten må være forståelige. Forklaringen må vise hvordan algoritmen inngår i avgjørelsen, i en form for systembeskrivelse. Det må også gis en forklaring på hvordan algoritmen virker og hvordan den er anvendt. Nr. 3 forstår jeg dithen at dersom modellen eller algoritmen er global, er det tilstrekkelig å angi parametrene. Hvis ikke må det også gis en lokal forklaring av hvordan de er vektet i avgjørelsen.

²⁷³ Dette er min omtrentlige oversettelse. Den originale ordlyden er:

« L'administration communique à la personne faisant l'objet d'une décision individuelle prise sur le fondement d'un traitement algorithmique, à la demande de celle-ci, sous une forme intelligible et sous réserve de ne pas porter atteinte à des secrets protégés par la loi, les informations suivantes :

1° Le degré et le mode de contribution du traitement algorithmique à la prise de décision ;

2° Les données traitées et leurs sources ;

3° Les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ;

4° Les opérations effectuées par le traitement. »

Dersom forvaltningsorganet innfrir på disse punktene vil parten få informasjon som tilsvarende en vedtaksbegrunnelse etter forvaltningsloven. Den viser hvordan systemet har fattet avgjørelsen, og den kan brukes til å utøve kontradiksjon og ansvarliggjøre beslutningstakeren. Eksempelet viser at det kan være en fordel med saksbehandlingsregler som er tilpasset algoritmisk forvaltning.

5.4.5 Oppsummering

I dette avsnittet har jeg vurdert to lokale forklaringsmetoder for komplekse maskinlæringsmodeller: approksimasjoner og kontrafaktiske forklaringer. Vurderingen ble foretatt både med henblikk på automatiserte avgjørelser og beslutningsstøtte.

Jeg kom til at approksimasjoner har en struktur og fleksibilitet som gjør dem egnet for en forvaltningsbegrunnelse, men at de kanskje ikke er tilstrekkelig gjennomgående presise til å regnes som rettssikre.

Kontrafaktiske forklaringer inneholder ikke den nødvendige informasjon til å kunne karakteriseres som en begrunnelse, men er brukervennlige og kan bidra til å belyse vurderinger og avgjørelser hvor begrunnelse ikke er påkrevd.

Jeg kom også til at approksimasjoner kan være egnet som forklaring dersom modellen gir beslutningsstøtte. Det krever imidlertid at saksbehandleren har god global forståelse av modellen og dens muligheter og begrensninger.

Det vil være viktig å utvikle forklaringsmetoder som er tilpasset forvaltningsrettslige begrunnelseskrav. Hvis ikke, må de mer komplekse modellene beholdes avgjørelser hvor begrunnelse ikke er påkrevd, eller hvor prediksjonen utgjør en mindre komponent i avgjørelsen.

Selv der lokale forklaringer er hensiktsmessige og nøyaktige kan de ikke stå på egne ben som begrunnelse til parten eller en forklaring til en saksbehandler. De må suppleres med forklaringer som beskriver modellen og beslutningssystemet.

Avslutningsvis reiste jeg spørsmålet om ikke algoritmiske avgjørelser bør ha et annet begrunnelseskrav enn menneskelige avgjørelser, som er mer i tråd med algoritmenes virkemåte.

I følgende avsnitt vil jeg gå inn på forklarbarhet i relasjon til bias og algoritmisk forskjellsbehandling. Fokuset flyttes da fra *ex post* forklaringer av enkeltprediksjoner til det globale nivået og systemutviklingen.

5.5 Forklarbarhet og algoritmisk forskjellsbehandling

Grunnloven § 98 forbyr diskriminering og usaklig forskjellsbehandling. Diskrimineringsforbudet er også nedfelt i mange lover og konvensjoner, og forbudet mot usaklig forskjellsbehandling setter en skranke for forvaltningens skjønnsutøvelse gjennom myndighetsmisbrukslæren.

Lovverket beskytter også mot det man kan kalle algoritmisk diskriminering: usaklig forskjellsbehandling eller diskriminering i forbindelse med automatisert behandling av personopplysninger.²⁷⁴ GDPR har bestemmelser som særlig retter seg mot dette, spesielt artikkel 22 og prinsippene om rettferdighet og åpenhet i artikkel 5 (1) (a), omtalt ovenfor i avsnitt 5.2.1. Diskriminerende avgjørelser basert på behandling av personopplysninger vil som nevnt også være i strid med retten til privatliv etter EMK art. 8.

I avsnitt 3.3.4 beskrev jeg noe av bakgrunnen for at diskriminering og usaklig forskjellsbehandling er særlige risikomomenter ved algoritmisk forvaltning. Dette ble også et tema i SyRI- og IB-sakene.

I SyRI-saken anførte saksøkerne at systemet kunne få en stigmatiserende og selvforsterkende effekt fordi prosjektene bare ble iverksatt i såkalte «problemområder» med høyere innvanderandel og lavere sosio-økonomisk status. Der- som det ble gjort funn i disse områdene, kunne det lære modellen til å legge ekstra vekt på sosio-økonomisk status i fremtiden. Retten var enig i at det forelå en risiko for utilsiktede skjevheter, fordi risikoprofilene var basert på (potensielt) store mengder med personopplysninger. Det var imidlertid ikke mulig å vurdere om denne risikoen var hensyntatt og nøytralisert, på grunn av manglende innsyn i dokumentasjon om systemet. Dette momentet bidro til konklusjonen om at SyRI-lovgivningen samlet sett ikke overholdt prinsippet om åpenhet i GDPR og retten til privatliv i EMK art. 8.²⁷⁵

I IB-saken uttalte Datatilsynet at bruken av data om skolekontekst i karakterfastsettelsen medførte en risiko for diskriminerende utfall. Det ble vist til at betydningen av skolekontekst kunne avhenge av mange faktorer, som for eksempel hvor lenge skolen hadde tilbudt IB, hvor mange elever det var ved skolen eller om skolen var lokalisert i et område med en høy andel av ressurssterke eller svakerestilte familier. Det var problematisk at det ikke var åpenhet om hvordan

²⁷⁴ Criado og Such (2020) s. 82.

²⁷⁵ SyRI-dommen, avsnitt 6.93 og 6.94.

komponenten var regnet ut eller om potensielle uheldige konsekvenser for sosial rettferdighet var vurdert.²⁷⁶

Slik jeg leser vurderingene i disse sakene krever prinsippene om åpenhet og rettferdighet en høyere grad av forklarbarhet i systemer hvor det foreligger en risiko for diskriminering.²⁷⁷ En slik risiko kan skyldes at modellen benytter store mengder personopplysninger, eller at den benytter data som fungerer som proxies (markører eller stedfortredende variabler) for sensitive eller beskyttede egenskaper.

Forklarbarhet vil i denne sammenheng utgjøre innsyn (der det er tilstrekkelig) eller dokumentasjon på at nødvendige tiltak er iverksatt for å avdekke og fjerne kilder til diskriminering og annen uønsket forskjellsbehandling. Forklarbarhet er dermed uløselig knyttet til systemkrav og kontroll av algoritmiske beslutningssystemer.

Systemkrav for å forhindre feil og diskriminering følger av GDPRs artikkel 25 om innebygd personvern og fortalens punkt 71. Fortalens punkt 71 sier i forbindelse med prinsippene om åpenhet og rettferdighet ved profilering og helautomatiserte avgjørelser, at den behandlingsansvarlige må bruke «egnete matematiske eller statistiske fremgangsmåter» og gjennomføre «egnete tekniske og organisatoriske tiltak» for å minimere risikoen for feil og diskriminering. Det må med andre ord benyttes metoder og kontrolltiltak som er tilpasset det aktuelle systemets risikopotensiale.²⁷⁸

Et særegent aspekt ved algoritmiske beslutningssystemer er at potensielle kilder til diskriminering og usaklig forskjellsbehandling kan gjøre seg gjeldende på ulike tidspunkter i modellens og systemets livssyklus. Mange er relatert til dataene selv, måten de er samlet inn, strukturert, merket og vasket, mens andre kan være forbundet med treningen, testingen eller måten systemet implementeres.²⁷⁹ Det tilsier at det er nødvendig med et bredt sett av tiltak gjennom hele modellen og systemets livssyklus: regler og rutiner for god datahåndtering, testing, vedlikehold og eksternt tilsyn eller kontroll.

276 Datatilsynets forhåndsvarsel til IBO 7.8.20.

277 En slik tenkning kan man også se gjenspeilet i EU-kommisjonens forslag til regulering av KI-systemer, ved at det er valgt en risikobasert tilnærming (EU-kommisjonen 2021).

278 EUs utkast til regulering av KI-systemer inneholder også diverse systemkrav er rettet mot å minimere risikoen for feil, skjevhet og vilkårlighet. F.eks. står det i pkt. 44 i Explanatory Memorandum: High quality training, validation and testing data sets require the implementation of appropriate data governance and management practices. Training, validation and testing data sets should be sufficiently relevant, representative and free of errors and complete in view of the intended purpose of the system.” (EU-kommisjonen 2021.)

279 Kaufmann (29.10.20).

Et annet poeng er at algoritmisk diskriminering og vilkårlighet er et utpreget gruppefenomen, og at det er uforutsigbart hvordan det arter seg.²⁸⁰ Det kan ramme beskyttede grupper man vanligvis har oppmerksomheten rettet mot (som konsekvens av historisk bias, for eksempel), men også grupper man vanligvis ikke assosierer med beskyttede egenskaper. De er bare en gruppe i kraft av at dataanalysen har avdekket statistisk korrelasjon mellom medlemmenes egenskaper, og ikke på grunn av at samfunnet oppfatter dem som å ha noen felles identitet eller felles trekk som bør være relevante for en avgjørelse.

Dette uforutsigbare elementet, sammen med modellenes kompleksitet og vanskelige tilgjengelighet gjør at algoritmisk diskriminering og usaklig forskjellsbehandling kan være vanskelig å påvise. Det blir derfor viktig med metoder for å avdekke skjevheter og forutse deres konsekvenser.

Med dette som bakgrunn vil jeg se litt nærmere på hva som kan være egnede matematiske eller statistiske fremgangsmåter og egnede tekniske og organisatoriske tiltak.

5.5.1 Avdekking av slagside med rettferdighetsverktøy

Rettferdighetsverktøy (ofte omtalt som fairness tools) er statistiske og matematiske analyseverktøy som kan brukes til å fange opp og måle diskriminerende tendenser i en modell. Criado og Such omtaler disse verktøyene som et førstelinjeforsvar mot diskriminerende og urettferdige algoritmiske avgjørelser. De kan både brukes til å fange opp diskriminerende tendenser i en modell som er implementert, og i tidligere faser med datahåndtering og trening.²⁸¹

Ved bruk av rettferdighetsverktøy prøver man å etterstrebe ulike former for *paritet*.²⁸² Det vil si at måten modellen tar feil skal være mest mulig lik på tvers av grupper (der slike ulikheter i feilraten kan innebære uønsket forskjellsbehandling). En stor fordel er at man da kan identifisere grupper som står i fare for å bli forskjellsbehandlet *før* modellen implementeres. Det gir mulighet for å agere i forkant ved å justere modellen. Rettferdighetsteknikker kan derfor bidra til at kontrollen med forvaltningsavgjørelser i større grad kan flyttes fremover i tid, til systemutvikling og -testing.²⁸³

280 Edwards og Veale skriver for eksempel at «algorithmic harms typically arise from how systems classify or stigmatise groups». (Edwards og Veale 2017 s. 22).

281 Criado og Such (2020) s. 8. Det skjer mye utvikling på dette området, og det finnes mange slike verktøy, enkelte av dem er tilgjengelige gratis online. (Lyngstad, innlegg på LDOs årskonferanse 29.10.20.)

282 Det finnes ulike mål, som for eksempel «accuracy rate parity», «misclassification rate» osv. Jeg går ikke i detalj om disse her.

283 Prabhu (samtale 5.1.21).

Rettferdighetsverktøy retter seg likevel bare mot modellens prediksjoner, og ikke beslutningssystemet som sådan. En tendens til skjevhet i maskinlæringsmodellen trenger ikke materialisere seg i avgjørelser som rettslig sett ville blitt karakterisert som diskriminering eller usaklig forskjellsbehandling. Dette er relative, komparative og kontekstuelle vurderinger man ikke kan tallfeste eller vurdere på forhånd – det blir en konklusjon man først kan trekke i det konkrete tilfellet etter at avgjørelsen er fattet.²⁸⁴

Rettferdighetsteknikker er heller ikke godt egnet til å avdekke verdier, ideologi og politikk som mer eller mindre bevisst nedfelles i modellen og beslutningssystemet, og som kan bidra til å opprettholde eller forsterke skjevheter i samfunnet. Det vil kreve en bredere tilnærming, som også analyserer systemet i sin kontekst.

5.5.2 Avdekking av slagside ved hjelp av sosio-teknisk analyse: AMS-algoritmen for oppfølging av arbeidssøkende

Som et eksempel på en slik kontekstuell analyse vil jeg omtale Allhutter et al. sin studie av et algoritmisk beslutningssystem implementert av Arbeitsmarktservice (tilsvarende NAV arbeid) i Østerrike.²⁸⁵ Systemet er også interessant å se på med tanke på at NAV ser for seg et lignende verktøy for segmentering av arbeidssøkende.²⁸⁶

Arbeitsmarktservice (AMS) innførte i 2020 et algoritmisk beslutningssystem i oppfølgingen av arbeidssøkende, omtalt som «AMS-algoritmen». Systemet profilerer arbeidssøkende og klassifiserer dem i tre kategorier basert på statistikk om hvem som erfaringsmessig kommer seg raskest i arbeid. Personer tilhørende gruppe A regnes for å ha gode utsikter og være selvgående, så disse tilbys lite oppfølging. Gruppe C har lavest skår. De anses for å ha dårlige utsikter for varig sysselsetting, og henvises til å få bistand fra andre etater. Gruppe B består av alle dem som ikke plasseres i A eller C. Det er denne gruppen som får mest oppfølging fra AMS, blant annet kompetansehevingstiltak. Avgjørelsen om hvilken gruppe man havner i kan derfor få betydelige velferdsmessige konsekvenser. Beslutningssystemet er delautomatisert, slik at en veileder skal vurdere model-

284 Wachter et al. (2020).

285 Allhutter et al. (2020).

286 Se figur 2 i avsnitt 2.5; «Trenger den arbeidsledige bistand fra NAV?».

lens prediksjon og ta en endelig avgjørelse om hvilken gruppe en arbeidssøker plasseres i.²⁸⁷

Systemets formål er innsparing på utgifter og antall ansatte, og optimering av ressursbruken. AMS skal bruke sine virkemidler slik at det får flest mulig tilbake i jobb på varig basis, som er deres viktigste styringsparameter. Da er det ikke kostnadseffektivt å bruke ressursene på gruppe A og C. Systemet blir gjerne fremhevet, blant andre av AMS sin direktør, som et verktøy for effektiv, objektiv og verdinøytral oppfølging.²⁸⁸

Allhutter et al. har som utgangspunkt at algoritmisk skjevhet er et sosio-teknisk fenomen. En sosio-teknisk analyse som deres har som formål å avdekke kilder til skjevhet og dekonstruere uuttalte forutsetninger og verdier som kodes inn i systemet gjennom blant annet utvalg og merking av variabler, fastsettelse av terskelverdier og valg av algoritmetype. Disse vurderingene er av mer teknisk karakter.

Systemet må også forstås i sin politiske og ideologiske kontekst. Allhutter et al. analyserte derfor hvordan systemet ble omtalt i mediedebatten og presentert i diverse offentlige dokumenter om arbeidsmarkedspolitikken. De ser AMS-algoritmen i lys av en neoliberalistisk dreining i oppfølgingen av arbeidssøkende. Neoliberal markedstenkning fører til en utpreget individualisering, der aktivisering, motivering og kompetanseheving av den enkelte blir sett på som et svar på det som i stor grad kan skyldes makroøkonomiske strukturer og svingninger. Det viser seg blant annet ved at AMS-algoritmen i stor grad benytter variabler knyttet til enkeltindividet i modelleringen, som eksempelvis kjønn, utdanning, helsestatus og ansettelseshistorikk, mens det er lite data relatert til økonomien, kun data om regionalt arbeidsmarked.²⁸⁹

Forskerne kom til at AMS-algoritmen lærte seg til å bruke arbeidsmarkedets konkurranselogikk i profileringen av arbeidssøkende, slik at grupper med en historikk med svakere tilknytning til arbeidsmarkedet hadde større sannsynlighet for å havne i gruppe C. Dermed kunne den bidra til å reproducere og forsterke de mekanismene som i utgangspunktet fører til at kvinner, ungdom og minoriteter har dårligere muligheter på arbeidsmarkedet. Systemet er altså ikke

287 Allhutter et al. (2020) s. 2. Dette er kategoriseringen i mer detalj: Gruppe A: > 66 % sannsynlighet for å komme seg i jobb med varighet på minimum 3 måneder innen 7 måneder, Gruppe C: < 25% sannsynlighet for å komme seg i jobb med varighet på minst 6 måneder i løpet av de neste 2 årene. Studien har også interessante betraktninger om saksbehandlerens rolle i relasjon til algoritmen og forholdet mellom prediksjon og skjønn, men det faller litt på siden å diskutere her.

288 Allhutter et al. (2020) s. 9 og 10.

289 Allhutter et al. (2020) s. 3.

bare et nøytralt og objektivt verktøy for effektive avgjørelser. Her er nærmere om enkelte av deres funn:

«Kvinner» ble definert som en enhetlig gruppe assosiert med dårligere jobbutsikter. Det førte til at også kvinner som i utgangspunktet hadde gode sjanser for å komme seg raskt i jobb (for eksempel på grunn av sine kvalifikasjoner) fikk en lavere skår enn de burde hatt.²⁹⁰ Variabelen «omsorgsoppgaver» bidro til også å trekke skåren ned, og denne variabelen var kun knyttet til kvinner.²⁹¹

Det er ikke tillatt etter diskrimineringslovgivningen å bruke etnisitet i modelleringen. I stedet ble «migrasjonsbakgrunn» benyttet, som også er forbundet med dårligere sjanser på arbeidsmarkedet. AMS var ikke åpne om hvordan denne variabelen var definert, men det er sannsynlig at den i stor grad korrelerer med etnisitet.²⁹²

Modellen vurderte en persons jobbmuligheter basert på arbeidsmarkedet i bostedsregionen. Men det ble ikke tatt høyde for at man kan søke jobber i andre regionale arbeidsmarkeder, for eksempel i naboregionen. En person bosatt i et område med svakere økonomi og dårligere arbeidsmarked ville dermed få en dårligere skår, uavhengig av muligheten for å forflytte seg. Bosted er kjent for å være en proxy for sosial klasse, og kunne også få den virkningen her.²⁹³

Studien illustrerer hvor komplisert det kan være å avdekke kilder til skjevhet og vurdere hvilke sosiale konsekvenser de kan få. En god sosio-teknisk analyse av et algoritmisk beslutningssystem krever tverrfaglig kompetanse innenfor teknologifag, samfunnsvitenskap og juss. Det betyr at innsyn i modellen, selv om informasjonen skulle være godt tilrettelagt, ofte ikke vil være tilstrekkelig til å kunne trygge potensielle parter og offentligheten på at et beslutningssystem ikke fører til usaklig forskjellsbehandling eller uønskede sosiale forskjeller.

Forklarbarhet, i betydningen åpen og rettferdig behandling av personopplysninger, innebærer dermed at det bør finnes tilgjengelig dokumentasjon på at gode datahåndteringsrutiner er fulgt, som at treningsdata har god representativitet og at man benytter rettferdighetsverktøy for å teste og overvåke systemet. Studien viser også at en algoritmisk konsekvensanalyse²⁹⁴ bør omfatte en form for tverrfaglig sosio-teknisk analyse, i hvert fall i tilfeller der man implementerer systemer med potensielt vidtrekkende velferdsmessige konsekvenser.

290 Allhutter et al. (2020) s. 2.

291 Allhutter et al. (2020) s. 2.

292 Allhutter et al. (2020) s. 8.

293 Allhutter et al. (2020) s. 8.

294 Jf. GDPR art. 35, Kaminski og Maglieri (2019).

5.5.3 Oppsummering

I dette avsnittet pekte jeg først på at algoritmisk diskriminering og forskjellsbehandling kan være vanskelig å avdekke på enkeltavgjørelsesnivå. Hensynet til rettssikkerheten tilsier derfor at man i størst mulig grad bør identifisere og håndtere kilder til skjevhet på systemnivå.

Det følger også av GDPR art. 25, jf. fortalens punkt 71 at den behandlingsansvarlige må iverksette nødvendige tiltak og fremgangsmåter for å avverge risikoen for diskriminerende avgjørelser. Dette kan være rettferdighetsverktøy komplementert med sosio-teknisk analyse eller andre varianter av algoritmisk konsekvensvurdering.

SyRI-dommen og Datatilsynets forhåndsvarsel viser at den behandlingsvarlige, for å overholde prinsippet om åpen og rettferdig behandling, må kunne dokumentere for offentligheten og sakens parter at de nødvendige tekniske og organisatoriske skritt er tatt for å minimere risikoen for feil og usaklig forskjellsbehandling. Det gjelder særlig i situasjoner der systemene behandler mye personopplysninger eller sensitive personopplysninger.

6 Avslutning

6.1 Oppsummerende punkter

I denne oppgaven har jeg vist at algoritmiske beslutningssystemer kan være en trussel mot rettssikkerhetsprinsipper og demokratiske verdier dersom de ikke er tilstrekkelig forklarbare for offentligheten og parten i en forvaltningssak.

Jeg har argumentert for et bredt perspektiv på forklarbarhet, som ikke bare tar for seg modellen og ulike metoder å belyse den, men også dens funksjon i et beslutningssystem eller saksbehandlingsprosess. Basert på den mer teknisk orienterte gjennomgangen i kapittel 3 kom jeg til at forklarbarhet bør sees på som et knippe med tiltak og egenskaper ved et algoritmisk beslutningssystem, som samlet kan bidra til at forvaltningsrettslige krav til åpenhet, forståelighet, etterprøvbarehet og ansvarlighet innfris.

I kapittel 4 viste jeg at retten til en åpen og forståelig forvaltning vil kreve en mer aktiv og formidlende informasjonsvirksomhet fra forvaltningens side. For å realisere offentlighetsprinsippet i Grunnloven § 100 og offentleglova § 3 og den aktive informasjonsplikten vil det være nødvendig med regler for hvordan algoritmiske beslutningssystemer skal dokumenteres, og eventuelt løsninger for å publisere dem.

Ved utvikling av systemer som kan være kontroversielle eller som vil ha stor velferdsmessig betydning, bør forvaltningen dessuten legge til rette for deltakende prosesser. Det kan forankres i GDPR artikkel 35 om personvernkonsekvensvurdering eller i tilpassede regler for algoritmisk konsekvensvurdering.

I kapittel 5 viste jeg hvordan lokale og globale forklaringer og systembeskrivelser må virke sammen for å ivareta retten til kontradiksjon, personvern og privatliv ved algoritmisk saksbehandling. Forvaltningsloven, GDPR og EMK artikkel 8 krever at parten i en forvaltningssak får forståelig informasjon om grunnlaget for en avgjørelse (hvilke data som inngår og hvordan de er behandlet av systemet) og mulighet for å bestride den.

Beslutningssystemer med maskinlæringsmodeller som lager profiler og utfører prediktiv analyse basert på personopplysninger kan medføre en risiko for vilkårlighet og usaklig eller diskriminerende forskjellsbehandling. Der det fore-

ligger en slik risiko skjerpes kravet til forklarbarhet, ved at parten har rett til å få vite hvordan risikoen har blitt vurdert og hensyntatt før og under systemets iverksettelse. Systemenes kompleksitet gjør det likevel ikke realistisk å ivareta dette gjennom partsinnsyn og begrunnelse rettet mot den enkelte part. Det bør heller håndteres gjennom systemkrav, og ved å tilgjengeliggjøre forståelig dokumentasjon som viser at disse kravene er etterlevd.

I kapittel 5 vurderte jeg også betydningen av lokale forklaringsmetoder, som er utviklet for å forklare enkeltprediksjoner fra komplekse modeller. Min gjennomgang tyder på at det er vanskelig å tilfredsstille forvaltningsrettslige krav til begrunnelse ved hjelp av lokale forklaringsmetoder. Det kan tilsi at man ikke benytter de mest komplekse modellene i automatiserte prosesser. Derimot kan slike metoder, fortrinnsvis approksimasjoner, være egnet til å forklare prediksjonen der den er grunnlag for eller supplerer skjønnet i en menneskelig avgjørelse. Saksbehandleren må i alle tilfelle forstå modellen på et globalt nivå for å kunne bruke prediksjonen på en forsvarlig måte i avgjørelsen.

Flere av eksemplene jeg har sett på i oppgaven har vist at det skaper usikkerhet dersom det er uklart hvordan systemet fatter avgjørelser. Hensynet til tillit og åpenhet gjør det nødvendig å formidle på en tydelig måte hvordan myndighetsutøvelsen skjer. Det kan innebære at man viser funksjonsfordelingen mellom automatiserte og menneskelige elementer i en systembeskrivelse. Systembeskrivelser vil også gjøre det enklere å stille forvaltningsorganet til ansvar.

6.2 Systemkrav og kontroll ved algoritmisk forvaltning

Gjennomgangen i kapittel 4 og 5 har demonstrert at algoritmisk forvaltning gjør det nødvendig med systemperspektiv og systemkrav. Dette er det lite tradisjon for i norsk forvaltningsrett.²⁹⁵ Forvaltningsloven regulerer først og fremst forholdet mellom den private parten og forvaltningen under behandlingen av en enkeltsak, den har ikke regler om utvikling og innrettelse av saksbehandlingssystemer.

GDPR har derimot flere systemkrav som har betydning for utvikling og implementering av algoritmiske beslutningssystemer. Ett eksempel er artikkel 25, som sier at de nødvendige tekniske og organisatoriske tiltak må være på plass for å ivareta de registrertes rettigheter og overholde GDPR.

²⁹⁵ Boe (2018) s. 164 skriver det slik: «Helt fra forvaltningsrettens barndom har det altoverskygende perspektivet vært forvaltningsakten, det enkeltstående vedtaket. Systemet har kommet i bakleksen».

EU-kommisjonens forslag til regulering av KI-systemer er i stor grad basert på systemkrav. Forslaget legger opp til omfattende systemkrav for aktører som utvikler eller bruker høyrisikosystemer. De pålegges for eksempel å ha et kvalitetskontrollsystem for å sikre regeletterlevelse (artikkel 17), et risikokontrollsystem (artikkel 9) og kompetansekrav til brukerne (artikkel 14, omtalt i avsnitt 5.4.3). Forståelighet er i seg selv et systemkrav, jf. artikkel 13, som blant annet sier at systemet må designes på en måte som gjør at brukerne forstår det og bruker det på riktig måte, og at graden av transparens må tilpasses systemets karakter og brukernes behov.²⁹⁶

Dette vil ha konsekvenser for hvordan kontrollen med forvaltningen utøves. Gjeldende mekanismer for kontroll med forvaltningsvedtak er tilpasset tradisjonell saksbehandling. Klagebehandling og domstolsprøving er innrettet for å fange opp og korrigere for menneskelige feil i enkeltavgjørelser. I algoritmiske beslutningssystemer er vilkårlige, gale og diskriminerende avgjørelser symptomer på systemsvikt, og de bør håndteres på systemnivå.

Ved algoritmisk forvaltning kan derfor kontrollen i større grad komme til å dreie seg om å sikre at systemkrav etterleves. Slik kontroll kan skje gjennom interne rutiner for testing og overvåkning, og eksternt gjennom tilsyn og revisjon. Dette var jeg også inne på i avsnitt 5.5 om algoritmisk forskjellsbehandling: Metoder for å avdekke potensielle skjevheter i utviklings- og testfasen gjør at kontrollen i større grad kan utføres på forhånd, før systemet er iverksatt og avgjørelsene fattes.

Regjeringen uttaler i sin strategi for kunstig intelligens at «tilsynsmyndigheter skal føre kontroll med at systemer basert på kunstig intelligens på sitt tilsynsområde opererer innenfor prinsippene for ansvarlig og pålitelig bruk av KI».²⁹⁷

Regjeringens KI-strategi er ikke konkret med hensyn til hvor inngående revisjon som skal utføres. Direktør for Datatilsynet Bjørn Erik Thon sa på LDOs årskonferanse 29.10.20 at han ikke trodde så mye på revisjon av selve algoritmene, men at de heller vil ha som funksjon å kontrollere dokumentasjon på at regelverket etterleves.

Det kan imidlertid også bli nødvendig med mer inngående revisjon av systemer. Som nevnt i avsnitt 3.3.2 vil maskinlæringsmodellen bare la seg ettergå fullstendig dersom tilsynsmyndigheten får tilgang til de fulle datasett som algoritmene ble trent på og har lært fra.

²⁹⁶ EU-kommisjonen (2021).

²⁹⁷ Kommunal- og moderniseringsdepartementet (2020) s. 6.

Riksrevisjonen har produsert retningslinjer for tilsyn med KI-applikasjoner i samarbeid med tilsvarende myndigheter i en del andre land. Her legges det opp til en mer inngående form for tilsyn, og det foreslås konkrete revisjonspunkter under blant annet overskriftene data, modellutvikling og prestasjon etter iverksetting. Revisjonen kan for eksempel gå ut på å trene modellen på nytt med mindre data, med færre personopplysninger blant variablene eller med enklere (mer forståelige) algoritmer, og se om den fortsatt presterer akseptabelt.²⁹⁸

6.3 Regler for algoritmisk saksbehandling

Ved siden av de regler knyttet til utvikling, overvåkning og kontroll som følger av GDPR og som vi kan forvente fra EUs lovgivning på kunstig intelligens-systemer, kan det være grunn til å vurdere egne regler for algoritmisk saksbehandling som supplerer forvaltningslovens regler. Her følger noen punkter som kan være aktuelle for nærmere utredning.

I avsnitt 3.4.5 (og senere i avsnitt 4.4.2) påpekte jeg at forvaltningen må finne gode metoder for å formidle egnede forklaringer på modeller og systemer til borgerne og parter i forvaltningssaker. Dette vil kreve tverrfaglighet, men juristenes funksjon vil være å passe på at rettslige krav til innhold og klarhet blir overholdt. I dette arbeidet blir det kanskje naturlig å bygge videre på erfaringer og mønsterpraksis, eller utvikle retningslinjer.

I avsnitt 5.3.4 omtalte jeg en fransk lovbestemmelse om algoritmiske forvaltningsavgjørelser, som viser at algoritmiske avgjørelser kan begrunnes (eller heller forklares) på en måte som er mer tilpasset algoritmers funksjonsmåte. Det kan også være at et mer tilpasset begrunnelseskrav kan gjøre det enklere å utvikle egnede forklaringsmetoder.

I forbindelse med at det i større grad vil gjelde systemkrav for saksbehandlings-systemer, kan det også være grunn til å utrede muligheten for systemklager, der man klager på maskinlæringsmodellen eller beslutningssystemets innretning.²⁹⁹

Man bør også se på den rettslige statusen til slutninger fra maskinlæringsmodeller som inngår i forvaltningens avgjørelser. De bør ikke anses som fakta underlagt det vanlige sivilprosessuelle kravet til sannsynlighetsovervekt, fordi de er basert på korrelasjon og sannsynlighet og ikke faktiske kjensgjerninger.

²⁹⁸ Riksrevisjonen (2020).

²⁹⁹ Cobbe (2018) s. 8 (se fotnote 101 for sitat). Boe (2018) er heller ikke fremmed for tanken om at det kan være mulig for en domstol å sette til side et saksbehandlingssystem fordi det ikke er forvarlig innrettet.

Det kan være man må stille krav til modellens relevans og prediksjonenes treffsikkerhet, utover det som ellers vil følge av systemkravene.³⁰⁰

Egenskapene ved de ulike læringsformene (veiledet, ikke-veiledet og forsterket, se avsnitt 2.3) tilsier at veiledet læring vil være den minst problematiske fra et forvaltningsperspektiv, fordi den gir en mer forutsigbar og kontrollert modellutviklingsprosess. Det vil antakelig også være enklere å regulere – og etterleve – krav til tydelig angivelse av formål, dataminimering og datarelevans når utgangsvariablene er definert på forhånd. (Spørsmålet kom opp i SyRI-saken, se avsnitt 5.2.3). Det kan derfor hende valg av læringsform vil bli gjenstand for regulering eller retningslinjer.³⁰¹

300 Wachter og Mittelstadt (2019), Greenstein (samtale 9.10.20).

301 Jeg kan ikke se at EUs forslag til regulering av KI-systemer gjør noen slik differensiering ved at det stilles ulike krav basert på læringsform. (EU-kommisjonen 2021).

7 Henvisninger

7.1 Litteratur

- Algorithmwatch (2020) *Automating Society 2020*, <https://automatingsociety.algorithmwatch.org>, oppslag 15.12.20.
- Amariles (2020) Amariles, David Restrepo, Algorithmic Decision Systems. Automation and Machine Learning in the Public Administration. I: Barfield, Woodrow (red.) *The Cambridge Handbook of the Law of Algorithms*. Cambridge: Cambridge University Press, doi:10.1017/9781108680844.015.
- Annany og Crawford (2016) Annany, Mike og Kate Crawford, Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*. 20 (3), 973-989, doi:10.1177/1461444816676645.
- Barocas og Selbst (2016) Barocas, Solon og Andrew D. Selbst, Big Data's Disparate Impact, *California Law Review*, 104, 671, DOI: 10.15779/Z38BG3.1.
- Beaudouin et al. (2020) Beaudouin, Valérie, Isabelle Bloch, David Bounie, Stéphan Cléménçon, Florence d'Alché-Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskyi og Jayneel Parekh, *Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach*. <https://ssrn.com/abstract=3559477>, oppslag 12.6.2020.
- Bing (1977) Bing, Jon, Automatiseringsvennlig lovgivning. *Tidsskrift for rettsvitenskap*, 1977, 195-229.
- Boe (2018) Boe, Erik Magnus, *Forsvarlig forvaltning*. Oslo: Universitetsforlaget, 2018.
- Brown et al. (2019) Brown, Anna, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin og Rhema Vaithianathan, *Toward Algorithmic Accountability in Public Services. A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services*. <https://doi.org/10.1145/3290605.3300271>.

- Burrell (2016) Burrell, Jenna, How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, juni 2016, doi:10.1177/2053951715622512.
- Bygrave (2019) Minding the Machine v2.0 The EU General Data Protection Regulation and Automated Decision- Making. I: I: Yeung, Karen og Martin Lodge (red.) *Algorithmic Regulation*. Oxford Scholarship Online: oktober 2019. DOI: 10.1093/oso/9780198838494.001.0001.
- Bygrave (2020) Bygrave, Lee A., *Machine Learning, Cognitive Sovereignty and Data Protection Rights with Respect to Automated Decisions*. University of Oslo Faculty of Law Research Paper No. 2020-35, SSRN: <https://ssrn.com/abstract=3721118>.
- Castelluccia og Le Métayer (2019) Castelluccia, Claude og Daniel Le Métayer, *Understanding algorithmic decision-making: Opportunities and challenges*. EPRS (European Parliamentary Research Service), doi: 10.2861/536131.
- Cobbe (2018) Cobbe, Jennifer, *Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making*. <http://dx.doi.org/10.2139/ssrn.3226913>.
- Coglianese og Lehr (2017) Coglianese, Cary og David Lehr, 2017. *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*. Faculty Scholarship. 1734. http://scholarship.law.upenn.edu/faculty_scholarship/1734, oppslag 6.8.20.
- Criado og Such (2019) Criado, Natalia og Jose M Such, Digital Discrimination. I: Yeung, Karen og Martin Lodge (red.) *Algorithmic Regulation*. Oxford Scholarship Online, oktober 2019, doi: 10.1093/oso/9780198838494.001.0001.
- Doshi-Velez et al. (2017) Doshi-Velez, Finale, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O’Brien, Kate Scott, Stuart Shieber, James Waldo, David Weinberger, Adrian Weller, Alexandra Wood, *Accountability of AI Under the Law: The Role of Explanation* (utkast). arXiv:1711.01134.

- Drtina (upubli-
sert) Drtina, Pernille Sandberg. Tekster om begrunnelsesk-
ravet i norsk forvaltningsrett. Ph. D. under arbeid.
- Eckhoff og Smith
(2014) Eckhoff, Torstein og Eivind Smith, 2014, *Forvaltningsrett*.
10. utgave. Oslo: Universitetsforlaget.
- Edwards og Veale
(2018) Edwards, Lillian og Michael Veale, Enslaving the Algo-
rithm: From a “Right to an Explanation” to a “Right to
Better Decisions”? *IEEE Security & Privacy* (2018) 16 (3),
46–54, doi:10.1109/MSP.2018.2701152.
- Edwards og Veale
(2017) Edwards, Lillian og Michael Veale, *Slave to the Algorithm*.
*Why a ‘right to an explanation’ is probably not the remedy
you are looking for*. [https://scholarship.law.duke.edu/cgi/
viewcontent.cgi?article=1315&context=dltr](https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=1315&context=dltr), oppslag
8.10.20.
- Engberg (2003) Engberg, Lars A. 2003, *Demokratisk ledelse i den helheds-
orienterede byfornyelse?* [https://www.researchgate.net/
publication/319141220](https://www.researchgate.net/publication/319141220), oppslag 23.02.21.
- Frøberg (2019) Frøberg, Thomas, Rettskildefagets avveiningsmodell –
kritikk og alternativer. I: Høgberg, Alf Petter og Jørn
Øyrehaugen Sunde (red.) 2019, *Juridisk metode og ten-
kemåte*. Oslo: Universitetsforlaget.
- Gil González og
de Hert (2019) Gil González, Elena og Paul de Hert, *Understanding the
legal provisions that allow processing and profiling of per-
sonal data—an analysis of GDPR provisions and princi-
ples*. <https://doi.org/10.1007/s12027-018-0546-z>.
- Goodman og
Flaxman (2016) Goodman, Bryce og Seth Flaxman, *European Union Reg-
ulations on Algorithmic Decision-Making and a “Right to
Explanation”*. arXiv:1606.08813 [stat.ML].
- Graver (2015) Graver, Hans Petter, 2015, *Alminnelig forvaltningsrett*. 4.
utgave. Oslo: Universitetsforlaget.
- Greenstein (2017) Greenstein, Stanley, *Our Humanity Exposed. Predictive
Modelling in a Legal Context*. Ph.d.-avhandling, Stock-
holms universitet., 2017. [http://su.diva-portal.org/smash/
get/diva2:1088890/FULLTEXT01.pdf](http://su.diva-portal.org/smash/get/diva2:1088890/FULLTEXT01.pdf), oppslag 12.12.20.

- Guidotti et al. (2018) Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi og Fosca Giannotti, 21.6.2018. *A Survey of Methods for Explaining Black Boxes*. arXiv:1802.01933.
- Ikdahl (2016) Ikdahl, Ingunn, Hvilken vekt skal man legge på trygde-rettens praksis? I: Førde, Reidun, Morten Kjelland og Ulf Stridbeck (red.) 2016, *Cand.mag., cand.med., cand.jur., cand.alt. Festskrift til Aslak Syse 70 år*. Oslo: Gyldendal.
- Jaedick (4.3.20) Jaedick, Dana, *Er maskinlæring forenlig med personvern-prinsippene?* Foredrag på seminaret Maskinlæring og jus, Senter for rettsinformatikk (SERI), Universitetet i Oslo 4.3.2020.
- Kahneman et al. (2016) Kahneman, Daniel, Andrew M. Rosenfield, Linnea Gandhi og Tom Blaser, *Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making*. Harvard Business Review, oktober 2016, <https://hbr.org/2016/10/noise>, oppslag 26.9.20.
- Kaminski og Maglieri (2019) Kaminski, Margot E. og Gianclaudio Malgieri, 18.9.19, *Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations*. <http://dx.doi.org/10.2139/ssrn.3456224>.
- Kane (2018) Kane, Aina Aune, Navs begrunnelser av enkeltvedtak. Hvilke forhold kan påvirke utformingen av Navs enkeltvedtak, og hvilke konsekvenser kan det ha for brukernes rettssikkerhet? *Tidsskrift for velferdsforskning 04/2018 (Volum 21)*.
- Kaun (2021) Kaun, Anne, Suing the algorithm: the mundanization of automated decision-making in public services through litigation. *Information, Communication & Society*, doi: 10.1080/1369118X.2021.1924827.
- Larsson og Heintz (2020) Larsson, Stefan og Fredrik Heintz, Transparency in artificial intelligence. *Internet Policy Review*, 9(2). doi: 10.14763/2020.2.1469.
- Lipton (2017) Lipton, Zachary C., *The Mythos of Model Interpretability*. arXiv:1606.03490v3.

- Lovett (4.3.20) Lovett, Hilde, *Hva er maskinlæring?* Foredrag på seminaret Maskinlæring og jus, Senter for rettsinformatikk (SERI), Institutt for privatrett, Universitetet i Oslo 4.3.2020.
- Lyngstad (14.5.20) Lyngstad, Cathrine Pihl, *Forklarbar KI i praksis*. Foredrag på Teknas webseminar om forklarbar kunstig intelligens 14.5.20, <https://www.tekna.no/fag-og-nettverk/IKT/ikt-bloggen/forklarbar-kunstig-intelligens/>, oppslag 14.7.20.
- Løland (14.5.20) Løland, Anders, *Metodegjennomgang – forklaring-smetoder*. Foredrag på Teknas webseminar om forklarbar kunstig intelligens 14.5.20, <https://www.tekna.no/fag-og-nettverk/IKT/ikt-bloggen/forklarbar-kunstig-intelligens/>, oppslag 14.7.20.
- Martini (2020) Martini, Mario, Regulating Algorithms. How to Demystify the Alchemy of Code. I: Ebers, Martin og Susana Navas (red.): *Algorithms and Law*. Cambridge: Cambridge University Press, doi:10.1017/9781108347846.004.
- Mehrabi et al. (2019) Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman og Aram Galstyan, *A Survey on Bias and Fairness in Machine Learning*. arXiv:1908.09635.
- Munro (2019) Munro, Eileen, *Predictive analytics in child protection*. CHES Working Paper No. 2019-03, https://www.researchgate.net/publication/332528200_Predictive_analytics_in_child_protection, oppslag 29.7.20.
- NAV (26.2.20) Kurs i maskinlæring 26.2.20, som del av NAVs kompetanseprogram om kunstig intelligens i offentlig forvaltning.
- NTNU og Universitetet i Helsinki (2020) NTNU i samarbeid med Universitetet i Helsinki, 2020, *Elements of AI*. (Åpent nettkurs om kunstig intelligens), <https://www.elementsofai.com/no/ai-challenge>.
- Olsen et al. (2019) Olsen, Henrik Palmer, Jacob Livingston Slosser, Thomas Troels Hildebrandt og Cornelius Wiesener *What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration*. University of Copenhagen Faculty of Law Research Paper No. 2019-84, <http://dx.doi.org/10.2139/ssrn.3402974>.

- Oswald (2018) Oswald, Marion, 2018 *Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power*. <https://doi.org/10.1098/rsta.2017.0359>.
- Reisman et al. (2018) Reisman, Dillon, Jason Schultz, Kate Crawford og Meredith Whittaker, 2018 *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. <https://ainowinstitute.org/aiareport2018.pdf>, oppslag 18.9.20.
- Reutter og Broomfield (2019) Reutter, Lisa og Heather Broomfield, *Kunstig intelligens / data science: En kartlegging av status, utfordringer og behov i norsk offentlig sektor – første resultater*. <http://hdl.handle.net/11250/2634733>, oppslag 3.6.21.
- Rudin (2019) Rudin, Cynthia, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019 (1), 206-215, doi.1038/s42256-019-0048.
- Rudin (8.8.19) Rudin, Cynthia. *Do Simpler Models Exist and How Do We Find Them?* KDD Keynote Address, <https://www.youtube.com/watch?v=wL4X4lG20sM&t=2134s>, oppslag 19.7.20.
- Sandis (6.7.20) Sandis, Constantine, *AI Intelligibility and the Right to Explanation*. Foredrag i regi av University of Lorraine Philosophy Department <https://www.youtube.com/watch?v=Or1DfNxxNdw&t=5632s>, oppslag 27.8.20.
- Schartum (15.3.21) Schartum, Dag Wiese, *Vokser den digitale forvaltningen over hodet på oss?* Innlegg på eForvaltningskonferansen 15.3.21.
- Schartum (2015) Schartum, Dag Wiese, 2015, *Den elektroniske forvaltningen og loven*. Oslo: Kommuneforlaget.
- Schartum (2018) Schartum, Dag Wiese, 2018, *Digitalisering av offentlig forvaltning – Fra lovtekst til programkode*. Bergen: Fagbokforlaget.

- Schartum (2019) Schartum, Dag Wiese, Utilstrekkelig forslag om digital forvaltning i ny forvaltningslov. *Stat og styring* 2019 (3), 54-57.
- Schartum (4.3.20) Schartum, Dag Wiese, *Anvendelse av maskinlæring innen det juridiske domenet*. Foredrag på seminaret Maskinlæring og jus, Senter for rettsinformatikk (SERI), Universitetet i Oslo 4.3.2020.
- Schartum et al. (2017) Schartum, Dag Wiese, Arild Jansen og Tommy Tranvik, 2017, *Digital forvaltning – en innføring*. Bergen: Fagbokforlaget.
- Simplilearn (2018) Simplilearn, *Decision Tree Algorithm With Example*, <https://www.youtube.com/watch?v=RmajweUfKvM>, oppslag 16.7.20.
- Strümke (14.5.20) Strümke, Inga, *Forklaring, forståelse og forvirring*. Foredrag på Teknas webseminar om forklarbar kunstig intelligens 14.5.20, <https://www.tekna.no/fag-og-nettverk/IKT/ikt-bloggen/forklarbar-kunstig-intelligens/>, oppslag 14.7.20.
- Syse (2016) En tradisjonell forståelse av materielle rettigheter i velferdsretten. I: Ikdahl, Ingunn og Vibeke Blaker Strand, 2016, *Rettigheter i velferdsstaten. Begreper, trender, teorier*. Oslo: Gyldendal Juridisk.
- The Guardian (19.8.20) Today in Focus podkast 19.8.20, *The A-levels Fiasco* <https://www.theguardian.com/education/audio/2020/aug/19/the-a-levels-fiasco-podcast>, oppslag 18.9.20.
- The Royal Society (2019) The Royal Society, 28.11.2019. *Explainable AI: The Basics. Policy Briefing*. <https://royalsociety.org/topics-policy/projects/explainable-ai/>, oppslag 23.10.20.
- Veale og Brass (2019) Veale, Michael og Irina Brass, *Administration by Algorithm? Public Management Meets Public Sector Machine Learning*. I: Yeung, Karen og Martin Lodge (red.) *Algorithmic Regulation*. Oxford Scholarship Online: October 2019, doi: 10.1093/oso/9780198838494.001.0001.

- Wachter et al. (2017) Wachter, Sandra, Brent Mittelstadt og Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 2017, <http://dx.doi.org/10.2139/ssrn.2903469>.
- Wachter et al. (2018) Wachter, Sandra, Brent Mittelstadt og Chris Russell, Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 2018, arXiv:1711.00399.
- Wachter et al. (2020) Wachter, Sandra, Brent Mittelstadt og Chris Russell, *Why Fairness cannot be Automated – Bridging the Gap between EU Non-Discrimination Law and AI*, arXiv:2005.05906.
- Wachter og Mittelstadt (2019) Wachter, Sandra og Brent Mittelstadt, A right to reasonable inferences. Rethinking data protection in the age of Big Data and AI. *Columbia Business Law Review*, 2019 (2), <https://ssrn.com/abstract=3248829>.
- Wired (15.1.18) Wired, 15.1.18, *A Child Abuse Prediction Model Fails Poor Families*, <https://www.wired.com/story/excerpt-from-automating-inequality/>, oppslag 5.10.20.

7.2 Lover

- 1814 Lov 17. mai 1814 Kongeriket Norges Grunnlov (Grunnloven).
- 1967 Lov 10. februar 1967 om behandlingsmåten i forvaltningssaker (forvaltningsloven).
- 1992 Lov 4. desember 1992 nr. 126 om arkiv (arkivlova).
- 2003 Lov 19. mai 2003 nr. 16 om rett til innsyn i dokument i offentlig virksomhet (offentleglova).
- 2021 Lov 21. mai 2021 nr. 42 om språk (språklova) (I kraft 1.1.2022).

7.3 Internasjonale rettskilder

- EMK *Convention for the Protection of Human Rights and Fundamental Freedoms*, Roma 4.11.1950.
- Forordning 2016/679 Kommisjonsforordning (EU) nr. 2016/679 av 27.4.2016 om vern av fysiske personer i forbindelse med behandling av personopplysninger og om fri utveksling av slike opplysninger samt om oppheving av direktiv 95/94/EF (generell personvernforordning) [PVF, GDPR].

7.4 Forarbeider og offentlige dokumenter

- Administrative Review Council (2004) Administrative Review Council, Government of Australia 12.11.2004 *Automated Assistance in Administrative Decision Making*.
- Artikkel 29-arbeidsgruppen (2017) Article 29 Data Protection Working Party 3.10.17, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*.
- Artikkel 29-arbeidsgruppen (2017b) Article 29 Data Protection Working Party 4.4.17, *Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is “likely to result in a high risk” for the purposes of Regulation 2016/679*.
- Commonwealth Ombudsman (2017) Commonwealth Ombudsman 2017, *Lessons learnt about digital transformation and public administration: Centrelink’s online compliance intervention*, https://www.ombudsman.gov.au/_data/assets/pdf_file/0024/48813/AIAL-OCI-Speech-and-Paper.pdf, oppslag 4.10.20.
- Datatilsynet (2018) *Kunstig intelligens og personvern*. Rapport, 2018.
- Datatilsynet (2019) Datatilsynet, 20.6.19 *Datatilsynets innspill til Regjeringens KI-Strategi*, <https://www.datatilsynet.no/contentassets/b9cc1bec9a7440ca9329000a8cce1787/datatilsynet---innspill-ki-strategi.pdf>, oppslag 3.10.20.
- EU-kommisjonen (2020) *White Paper On Artificial Intelligence - A European approach to excellence and trust*, COM(2020)65, Brussel 19.2.2020.

EU-kommisjonen (2021)	2021/0106 (COD) <i>Proposal for the Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts.</i> Brussel, 21.4.2021.
EUs høynivåekspertgruppe på kunstig intelligens (2019)	European Commission Independent High-Level Expert Group on AI: <i>Ethics Guidelines for Trustworthy AI</i> , 8.4.2019.
FNs spesialrapportør på ekstrem fattigdom (2019)	UN Special Rapporteur on Extreme Poverty and Human Rights, <i>Report of the Special Rapporteur on extreme poverty and human rights</i> , A/74/493, 2019.
FNs spesialrapportør på ekstrem fattigdom (26.9.19)	UN Special Rapporteur on Extreme Poverty and Human Rights (26.9.19) <i>amicus curiae</i> , SyRI-saken.
Fornyings- og administrasjonsdepartementet (2009)	Fornyings- og administrasjonsdepartementet 16.10.09, <i>Statens kommunikasjonspolitikk</i> .
Justis- og politidepartementet (2009)	Justis- og politidepartementet 2009, <i>Rettleiar til offentleg-lova</i> .
KMD (2020)	Kommunal- og moderniseringsdepartementet, 14.1.2020. <i>Nasjonal strategi for kunstig intelligens</i> .
NAV (4.8.19)	NAV 4.8.19, <i>Innspill til strategi for kunstig intelligens</i> , https://www.regjeringen.no/contentassets/0e36c85fcd-143a5b626c53cf292cb3b/strategi-kunstig-intelligens_innspill-fra-nav.pdf , oppslag 4.10.20.
NOU 1999:27	«Ytringsfrihet bør finde Sted» – Forslag til ny Grunnlov § 100.
NOU 2003: 19	<i>Makt og demokrati</i> .
NOU 2003:30	<i>Ny offentlighetslov</i> .
NOU 2006: 7	<i>Det lokale folkestyret i endring? – Om deltaking og engasjement i lokalpolitikken</i> .

NOU 2019: 5	<i>Ny forvaltningslov. Lov om saksbehandlingen i offentlig forvaltning (forvaltningsloven).</i>
NOU 2019: 9	<i>Fra kalveskinn til datasjø. Ny lov om samfunnsdokumentasjon og arkiver.</i>
NUT 1958:3	<i>Innstilling fra Komiteen til å utrede spørsmålet om mer betryggende former for den offentlige forvaltning.</i>
OECD (2019)	<i>Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449.</i>
Ot.prp. nr. 38 (1964–65)	<i>Om lov om behandlingsmåten i forvaltningssaker (forvaltningsloven).</i>
Ot.prp. nr. 77 (1991–92)	<i>Om lov om arkiv.</i>
Prop. 108 L (2019–2020)	<i>Lov om språk (språklova).</i>
Riksrevisjonen (2020)	Riksrevisjonen i Finland, Tyskland, Nederland, Norge og UK, 24.11.20 <i>Auditing machine learning algorithms. A white paper for public auditors.</i> https://auditingalgorithms.net/index.html , oppslag 5.1.21.
St.meld. nr. 19 (2008–2009)	<i>Ei forvaltning for demokrati og fellesskap.</i>
Teknologirådet (2018)	Teknologirådet, 2018, <i>Kunstig intelligens – Muligheter, utfordringer og en plan for Norge.</i>
UDI høringssvar (05.09.19)	Utlendingsdirektoratet 5.9.19 <i>Høringssvar fra Utlendingsdirektoratet – Endringer i utlendingsforskriften om bruk av automatiserte avgjørelser, innhenting av opplysninger fra andre offentlige myndigheter m.m.</i>

7.5 Rettsavgjørelser

C-09-550982-HA ZA 18-388, ECLI:NL:RBDHA:2020:1878, Den Haag 5. februar 2020 (SyRI-saken).

Case of P.G. and J.H. v. the United Kingdom, no. 44787/98, 25. September 2001.

Case of S. and Marper v. the United Kingdom, nos. 30562/04 og 30566/04, 4. Desember 2008.

Case of Satakunnan Markkinapörssu Oy and Satamedia Oy v. Finland, no. 931/13, 27. Juni 2017.

Rt. 1981 s. 745 (Isene).

Rt. 2000 s. 1056.

Rt. 2014 s. 1105 (Overskuddsmateriale).

7.6 Oversikt over samtaler

- Systemjurist Torill Christine Edvardsen og sjefsinformasjonsarkitekt Daniel Engehagen, Seksjon for informasjonsforvaltning, NAV. Forslag til oppgavetema, diskusjon og dordling: 13.1, 14.2, 9.3, 29.6 og 2.10.20.
- Robindra Prabhu, data scientist ved NAV IT, utvikling og data, 1.7.20, 9.11.20 og 5.1.21.
- Lars Sutterud og Anders Topper, data scientister ved NAV IT, utvikling og data, 3.9.20.
- Anders Løland, assisterende forskningssjef, Norsk Regnesentral, 17.8.20.
- Stanley Greenstein, Assistant Professor in Law and Information Technology (IT), Universitetet i Stockholm, 9.10.20.
- Heather Broomfield, stipendiat ved Institutt for offentlig rett, UiO, 7.12.20.

7.7 Oversikt over konferanser, kurs og seminarer

- NAVs interne kompetanseprogram «Kunstig intelligens i offentlig forvaltning». Jeg deltok på samlinger 26.2, 25.6 og 8.10.20.
- Heldagsseminar på Institutt for privatrett 4.3.20: «Maskinlæring og jus».
- Foredrag for NAV 26.6.20: Martin Jullum, Norsk Regnesentral: «Proper prediction explanations with Shapr».
- Foredrag for NAV 3.7.02: Anders Løland, Norsk Regnesentral: «Forklarbar kunstig intelligens».

- Årskonferansen til Likestillings- og diskrimineringsombudet (LDO): «Kunstig intelligens, fare eller fremgang for likestillingen?», 29.10.20.
- XXXV Nordic Conference on Law and IT, «Law in the Era of AI», Universitetet i Stockholm 11-12.11.20.
- eForvaltningskonferansen 2021: «The Ghost in the Machine», Fagforbundet 15.3.21
- Åpent nettbasert kurs om kunstig intelligens i regi av NTNU og Universitetet i Helsinki: «Elements of AI».