Tidsskrift for velferdsforskning

Universitetsforlaget

VITENSKAPELIG PUBLIKASJON

Årgang 25, nr. 3-2022, s. 1–15
ISSN online: 2464-3076
DOI: https://doi.org/10.18261/tfv.25.3.4

Algoritmer til barnets beste?

Algorithms in child protection decision-making

Edmund Henden
Professor, Senter for profesjonsstudier, OsloMet
edhen@oslomet.no

Sammendrag
Det er grunn til å tro at bruk av maskinlæring som beslutningsstøtteverktøy for tildeling av velferdsytelser vil øke i fremtidens velferdsstat. Et velferdsområde hvor man i noen land har startet utprøving av denne teknologien, er i barnevernet. I denne artikkelen forklarer jeg hva som skiller maskinlæringsverktøy fra tradisjonelle aktuariske beslutningsstøtteverktøy, og redegjør for noe av bakgrunnen for at barnevernstjenesten i en del land har tatt maskinlæringsverktøy i bruk. Artikkelen viser deretter hvilke etiske utfordringer denne bruken kan skape for barnevernsarbeidere, og hvilke lærdommer som kan trekkes når det gjelder en eventuell fremtidig beslutning om å ta maskinlæringsteknologi i bruk i norsk barnevernssammenheng. Jeg argumenterer for at en grundig utredning av mulige konsekvenser når det gjelder yrkesetiske retningslinjer for sosialfaglig arbeid, bør være en sentral del av beslutningsgrunnlaget. Videre fremsetter jeg noen konkrete forslag til hva en slik utredning bør inneholde.

Nøkkelord
barnevern, risiko vurdering, beslutningsstøtte verktøy, aktuariske metoder, maskinlæring, yrkesetiske retningslinjer

Abstract
It is highly likely that the use of machine learning tools in welfare sector decision-making will only increase in the future. In some countries these tools have already been deployed in children's social care. This article explains what distinguishes machine learning tools from traditional actuarial decision methods and describes how they have been used in child protection decisions. The article further discusses the ethical challenges the use of such tools may create in children's social care systems, and what lessons can be drawn for any future decision to use them in Norwegian child welfare services. It is argued that a comprehensive inquiry into the possible consequences for ethical codes of conduct in social work should form part of the basis for such a decision. The article ends by suggesting some areas of potential ethical risk that an inquiry of this sort should particularly focus on.

Keywords
child protection, risk assessment, decision tools, actuarial methods, machine-learning, ethical codes of conduct

1. Innledning
Barnevernet skal «sikre at barn og unge som lever under forhold som kan skade deres helse og utvikling, får nødvendig hjelp, omsorg og beskyttelse» (Barnevernloven, 1992, § 1-1). Basert på undersøkelser av barnets omsorgsbetingelser kan barnevernstjenesten sette inn ulike hjelpetiltak i hjemmet eller fremme sak for Fylkesnemnda om plassering i fosterhjem

eller institusjon. Barnevernet er derfor en forvaltning med betydelig makt som typisk møter enkeltmennesker i sårbare situasjoner (Fjeld et al., 2020). Det stiller strenge krav til beslutningsprosessen som leder frem til tiltak i barnevernssaker. For å redusere risikoen for feil er det derfor enighet om at barnevernsfaglige beslutninger bør være basert på vitenskapelig kunnskap og teoretiske modeller som har bred tilslutning (Christiansen & Kojan, 2019). En ny type modeller som har begynt å gjøre inntog i velferdsprofesjonene, er maskinbaserte læringsmodeller (ML). Innen flere velferdsområder har slike modeller vist seg å være bedre til å gjøre sannsynlighetsberegninger enn menneskelig skjønn kombinert med datamodellering basert på tradisjonelle statistiske metoder. Et eksempel er i helsesektoren, hvor maskinlæring har blitt utprøvd som beslutningsstøtteverktøy når det gjelder både diagnostikk og prognostikk av en rekke alvorlige sykdommer (for en kunnskapsoppsummering, se Becker, 2019). Det er kanskje derfor ikke overraskende at man i noen land har startet utprøving av denne teknologien også i barnevernstjenesten, hvor risikovurderinger inngår som en sentral del av kjernevirksomheten. Beslutningskonteksten i barnevernet er imidlertid ikke bare empirisk kompleks. Den er også normativt kompleks. Det dreier seg ikke bare om avveininger mellom risiko og beskyttelsesfaktorer for å forutsi barns utvikling og trivsel, men også om avveininger mellom skade som kan forebygges ved iverksettelse av ulike tiltak, og skade som kan påføres gjennom iverksettelse av slike tiltak. Dette reiser et spørsmål om når og under hvilke omstendigheter det er etisk legitimt å bruke ML-modeller som beslutningsstøtteverktøy i sosialfaglig arbeid.

Internasjonalt har diskusjonen rundt bruk av slike modeller i barnevernstjenesten allerede foregått over lengre tid. I artikkelens første del skisseres litt av bakgrunnen for denne diskusjonen, og jeg forklarer hvordan ML-modeller har blitt tatt i bruk av barnevernet i noen land. I andre del redegjør jeg for noen felles etiske prinsipper for helse- og sosialfaglig arbeid. I tredje del forklarer jeg på hvilken måte bruk av ML-modeller kan komme i konflikt med disse prinsippene. I konklusjonen diskuterer jeg hvilke lærdommer som kan trekkes når det gjelder en eventuell vurdering av å ta maskinlæringsteknologi i bruk i norsk barnevernstjeneste. Jeg argumenterer for at en grundig utredning av mulige etiske konsekvenser bør inngå som en sentral del av beslutningsgrunnlaget. Videre fremsetter jeg noen konkrete forslag til hva en slik utredning bør inneholde.

2. Beslutningsstøtteverktøy i helse- og sosialfaglig arbeid

Ifølge barnevernloven (1992) krever begrunnelse av tiltak ikke bare en vurdering av barnets omsorgssituasjon i nåtid, men også av barnets utvikling og trivsel på sikt (NOU 2000: 12, kap. 8). Det innebærer å bruke informasjon om barnets nåværende omsorgssituasjon for å trekke slutninger om noe man ikke har informasjon om, nemlig barnets fremtidige omsorgssituasjon. Prediksjon av usikre utfall står derfor ofte sentralt i den barnevernsfaglige beslutningsprosessen. Prediksjon krever at barnevernsarbeideren kombinerer informasjon om barnets situasjon med kunnskap om vilkår for barns utvikling og trivsel (for eksempel risiko- og beskyttelsesfaktorer) til en sannsynlighetsvurdering. I helse- og sosialfaglig arbeid skilles det tradisjonelt mellom to hovedtyper av tilnærminger til slike vurderinger: *kliniske* og *aktuariske* (Dawes et al.,1989).

Kliniske tilnærminger innebærer at vurderingen styres av sosialarbeiderens kliniske skjønn og fortolkning av ulike aspekter ved barnets omsorgssituasjon (Baird et al., 1999; English & Pecora, 1994). Ved hjelp av egen fagkunnskap og erfaring fra lignende saker kombineres og vektes den tilgjengelige informasjonen om risiko- og beskyttelsesfaktorer basert på en subjektiv vurdering av relevans og prediksjonskraft. På dette grunnlaget trekkes det en

slutning om barnets utvikling og trivsel på sikt. Fordelene med kliniske tilnærminger er at de garanterer en individualisert behandling som kan fange opp sjeldne eller uvanlige tilfeller, for eksempel barn som opplever omsorgssvikt til tross for fravær av typiske risikofaktorer i familiesituasjonen (Baumann et al., 2005; Crea, 2010). Samtidig beskytter de mot brudd på individers rettssikkerhet og menneskerettigheter fordi de gir høy grad av fleksibilitet og mulighet til å oppdage individspesifikke faktorer i familiers situasjon og risikobilde. Men det er også noen velkjente ulemper med kliniske tilnærminger. Jevnt over har de dårlig test-retest-reliabilitet (identisk informasjon gir identisk vurdering uavhengig av personen som vurderer), og relativt lav prediktiv validitet, dvs. lav treffsikkerhet når det gjelder forutsigelse (Arad-Davidson & Benbenishty, 2008; D'Andrade et al., 2005; Lyons et al., 1996). Forskning viser for eksempel at forskjellige saksbehandlere som presenteres for samme sakstilfeller, ofte vurderer sakene svært ulikt (Jergeby & Soydan, 2002; Rossi et al., 1996). En del forskning tyder også på at kliniske tilnærminger ikke presterer særlig mye bedre enn ren gjetning, noe som innebærer at de ofte fører til ukorrekte vurderinger (Baird & Wagner, 2000; Van der Put et al., 2016). Den vanligste forklaringen på dette er at kliniske tilnærminger er basert på en skjønnsmessig kombinering og vekting av risiko og beskyttelsesfaktorer, og at skjønnsutøvelse er sårbart for påvirkninger av irrelevante faktorer. Eksempler på irrelevante faktorer forskning har vist at kan påvirke skjønnsutøvelse, inkluderer kulturen på det lokale kontoret, saksbehandlernes personlige verdier (Ægisdóttir et al., 2006; Dawes et al., 1989), men også ulike kognitive skjevheter («bias») i selve informasjonsprosesseringen (Plous, 1993; Kahneman & Klein, 2009). Det er for eksempel evidens for at barnevernsarbeidere har en tendens til en overdreven vektlegging av informasjon som er iøynefallende eller følelsesladet, at de kan ha vanskeligheter med å korrigere førsteinntrykk i lys av ny informasjon, og at de ofte benytter feilaktige heuristiske strategier når informasjonsmengden blir stor og uoversiktlig (Munro, 1999). Lav prediktiv validitet er problematisk fordi det kan innebære at barn ikke får den hjelpen de trenger, eller at feil tiltak iverksettes overfor sårbare familier. Dårlig test-retest-reliabilitet er problematisk fordi det undergraver prinsippet om likebehandling.

Den andre hovedtypen tilnærminger til vurderinger i helse- og sosialfaglig arbeid er aktuariske. Ved bruk av aktuariske tilnærminger overlates vektingen og kombineringen av risiko- og beskyttelsesfaktorer til en veldefinert regel, gjerne en matematisk formel. Det innebærer typisk at slutningen om barnets fremtidige trivsel og utvikling trekkes ved hjelp av regresjonsanalyse på statistiske data som avdekker hvilke variabler som er assosiert med det utfallet som skal predikeres, og som vekter disse variablene slik at de på en optimal måte tilpasses datasettet (Meehl, 1954). Selve beslutningsprosessen er fastlagt på forhånd, og vurderingen er utelukkende basert på historisk etablerte empiriske relasjoner mellom data og det utfallet man er interessert i, dvs. de variablene regresjonsanalyse har vist at er de sterkeste prediktorene for dette utfallet. For eksempel har empiriske studier vist at konflikter mellom foreldre, psykiske vansker hos en av dem, fattigdom i hjemmet, lav utdannelse osv. er faktorer som statistisk øker risikoen for omsorgssvikt ikke bare i nåtiden, men også i fremtiden (Kvello, 2015). Med en rent aktuarisk tilnærming vil den barnevernsfaglige beslutningen utelukkende være basert på denne typen statistiske data kombinert med informasjon om risikofaktorer i barnets situasjon.

Fordelene og ulempene med aktuariske tilnærminger er på mange måter de motsatte av fordelene og ulempene med kliniske. Sammenlignet med sistnevnte har aktuariske tilnærminger vist seg å ha bedre test-retest-reliabilitet og høyere treffsikkerhet når det gjelder å estimere risiko innenfor en rekke områder (Baird & Wagner, 2000; Dawes et al., 1989; Grove & Meehl, 1996). Den naturlige forklaringen på dette er at vektingen og kombineringen av risiko- og beskyttelsesfaktorer utføres basert på forhåndsdefinerte regler snarere enn basert

på helse- eller sosialarbeiderens subjektive skjønn (som forskningen viser er sårbart for irrelevante påvirkninger). Ved bruk av aktuariske tilnærminger vil derfor forskjellige helse- og sosialarbeidere i prinsippet vekte og kombinere den tilgjengelige informasjonen på samme måte. Fordi vurderingen baserer seg på statistiske data på gruppenivå, vil imidlertid en ulempe kunne være at de overser individuelle variasjoner som kan begrunne spesielle tiltaksvalg i mer uvanlige eller sjeldne tilfeller.

Kjernen i maskinlæring er statistisk modellering. Slik sett føyer diskusjonen av denne teknologien i velferdsyrker seg inn i den klassiske diskusjonen rundt bruk av kliniske vs. aktuariske metoder i profesjonell praksis. Mange av de tradisjonelle argumentene for å utvide bruk av aktuariske beslutningsverktøy (for eksempel bedre reliabilitet, høyere prediktiv validitet, økt effektivitet osv.) er derfor de samme som begrunner en utvidet bruk av maskinlæringsteknologi. Tilsvarende er mange av de tradisjonelle innvendingene mot en utvidet bruk av aktuariske beslutningsverktøy (for eksempel lav fleksibilitet og evne til å oppdage individspesifikke faktorer) de samme som begrunner en mer restriktiv tilnærming til bruk av denne teknologien. Mange vil antagelig være av den oppfatningen at med utvikling av maskinlæringsteknologi har den første typen argumenter blitt styrket, mens de tradisjonelle innvendingene har blitt svekket. Det skyldes at maskinlæringsverktøy på mange måter representerer en dramatisk forbedring i forhold til mer tradisjonelle aktuariske beslutningsverktøy. I neste avsnitt forklarer jeg nærmere på hvilken måte og hvordan maskinlæring har blitt tatt i bruk av barnevernet i noen land.

3. Bruk av maskinlæring i barnevernstjenesten

Det er både likheter og forskjeller mellom bruk av ML-modeller og tradisjonelle aktuariske metoder. I likhet med statistiske analyseverktøy som lineær regresjon modellerer maskinlæringsalgoritmer relasjoner mellom uavhengige og avhengige variabler i data. Den viktigste forskjellen er at mens det i klassisk regresjonsanalyse vil være en menneskelig modellerer som ved hjelp av statistiske analyseverktøy (som OLS) definerer hvilke regresjonsfunksjoner som best beskriver relasjonen mellom disse variablene, vil det i en ML-modell være algoritmen selv som finner frem til disse funksjonene (Breiman, 2001). Algoritmen trenes først på store datasett fra mange ulike kilder, hvor antallet variabler kan være enormt (gjerne i samme størrelsesorden som antall observasjoner). I løpet av treningsperioden identifiserer algoritmen selv hvilke relasjoner mellom variablene som optimaliserer den stokastiske modellens tilpasning til datasettet. Når treningsperioden er over, blir algoritmen testet på valideringsdata – data den ikke har «sett» tidligere. Læringssuksess måles av hvor god den er til å generalisere fra trenings- til valideringsdata, dvs. hvor god den er til å finne mønstre i valideringsdata og generere prediksjoner basert på disse mønstrene (Domingos, 2012). Evnen til denne typen læring gir ML-algoritmer en stor fordel. De blir i stand til å finne frem til mye mer komplekse relasjoner i det som kan være både heterogene og dynamiske datasett (inkludert ikke-lineære relasjoner) enn det som er mulig ved hjelp av vanlige regresjonsfunksjoner som er definert av mennesker (eller datamaskiner) på forhånd. Fraværet av sistnevnte begrensninger gir dem stor fleksibilitet og evne til å finne ukjente mønstre i data (Marshall & English, 2000).[1] Dette bidrar til at de får høyere prediktiv treffsikkerhet enn tra-

1. En fare er imidlertid at modellene kan bli så fleksible at de blir overtilpasset («over-fitted») treningsdataene. De finner rett og slett så mange mønstre i dette datasettet at når de får nye data, klarer de ikke å generalisere fra dem basert på mønstre som er dominerende eller relevante. Resultatet kan bli dårligere prediksjoner. Det finnes ulike tekniske metoder for å redusere faren for overtilpasning (Athey & Imbens, 2017).

disjonelle aktuariske beslutningsverktøy. I motsetning til disse verktøyene krever bruken av dem dessuten ingen spesiell utdanning eller trening ettersom de ikke er operatørdrevne, men datadrevne og automatiserte. I tillegg til at dette reduserer sjansen for ukorrekte vurderinger grunnet feilaktig menneskelig bruk (for eksempel feiltolkninger av risikoestimater), og dermed bedrer test-retest-reliabilitet, gjør det maskinlæringsverktøy både mer effektive og mindre ressurskrevende å bruke enn tradisjonelle aktuariske beslutningsverktøy (Cuccaro-Alamin et al., 2017).

Disse fordelene er en viktig del av bakgrunnen for at man i noen land har begynt å ta i bruk ML-modeller som beslutningsstøtte i barnevernstjenesten. En viktig oppgave for barnevernet er å forutsi omsorgssvikt med et mål om å forhindre at det skjer. Dette er nødvendig for at det skal kunne iverksettes hjelpetiltak så tidlig som mulig (for eksempel ekstra oppfølging og støtte i hjemmet), som kan forebygge fremtidig vanskjøtsel og overgrep. Forebyggende tiltak krever at man lykkes med å identifisere de mest sårbare barna før de utsettes for alvorlig skade. Det kan være en vanskelig og ressurskrevende oppgave. Avgjørelsen barnevernet fatter, vil dessuten kunne ha store konsekvenser. Feil vurderinger kan føre til at barn og familier ikke får den oppfølgingen og hjelpen de trenger. Men det kan også føre til at barn og familier som ikke trenger oppfølging og hjelp, utsettes for unødvendige inngrep fra barnevernets side. Begge deler kan påføre barn alvorlige skader. Kjernen i utfordringen slike avgjørelser skaper for barnevernet, kan beskrives som et prediksjonsproblem: Hvilke barn har høy risiko for å oppleve omsorgssvikt og derfor behov for tidlig intervensjon som kan beskytte dem mot dette? Evnen til å gi mest mulig korrekte svar på dette spørsmålet er av stor betydning når det gjelder å sette inn ressurser der det er størst behov for det. Det er her maskinlæring kommer inn i bildet. Av grunner som er beskrevet over, er ML-modeller overlegne alle andre kjente metoder når det gjelder nøyaktig prediksjon. De er dessuten kostnadseffektive og relativt enkle å ta i bruk. Maskinlæringsalgoritmer kan identifisere risikostratifiserte individer i en populasjon basert på sannsynligheten for at det enkelte individ vil oppleve et bestemt utfall (Cuccari-Alamin et al., 2017). Basert på søk gjennom store datasett finner algoritmene selv frem til hvilke variabler og relasjoner som inkluderes i modellen. Ut fra dette lager de en risikoprofil for individene i populasjonen.

I USA har maskinlæringsmodeller blitt tatt i bruk blant annet for å predikere risiko for omsorgssvikt, barnemishandling, barnedødsfall og mislykket familiegjenforening (Packard, 2016). Individenes risikoprofil har blitt benyttet av barnevernstjenesten for å identifisere de mest sårbare barna og familiene og dermed målstyre ressurser mot de med det antatt største behovet for ekstra oppfølging og hjelp (Daley et al., 2016). Maskinlæringsteknologi har dermed vært en viktig del av beslutningsgrunnlaget når det gjelder å bestemme hvilke barn og familier som skal kontaktes av barnevernstjenesten. Denne bruken har imidlertid blitt møtt med voksende kritikk blant barnevernsarbeidere og barnevernsforskere (Leslie et al., 2020). Et sentralt moment i kritikken har vært at den kommer i konflikt med etiske prinsipper for sosialfaglig arbeid. Før denne kritikken skal diskuteres nærmere, må det kort redegjøres for noen viktige etiske prinsipper i sosialfaglig arbeid.

4. Profesjonsetikk for helse- og sosialfaglig arbeid

Mye av den moderne inspirasjonen for etiske retningslinjer i helse- og sosialfaglig arbeid kommer fra menneskerettighetstradisjonen og tradisjonen fra bioetikk (Reamer, 1985). Historisk oppsto begge tradisjonene i midten av forrige århundre som en reaksjon på overgrep og vold begått mot sårbare grupper i vitenskapens navn (Kuhse & Singer, 1998). Mens menneskerettighetstradisjonen er forankret i en idé om at alle mennesker har lik moralske status

i kraft av sin iboende verdighet og autonomi, og derfor har krav på juridisk beskyttelse av de samme sosiale, politiske og moralske rettigheter, er tradisjonen fra bioetikk forankret i en idé om den moralske betydningen av omsorg og beskyttelse av sårbare grupper og individer mot institusjonelle overgrep og maktmisbruk. Dette er en viktig del av bakgrunnen for at moralske verdier som menneskelig verdighet, sosial rettferdighet og respekt for individets selvbestemmelse ansees å være grunnleggende i profesjonsetikk over store deler av verden (Leslie et al., 2020).

Når det gjelder det teoretiske grunnlaget for profesjonsetikk i helse- og sosialfaglig arbeid, har den viktigste innflytelsen utvilsomt vært Tom Beauchamp og James Childress *Principles of Biomedical Ethics* (1979). Sentralt i deres «prinsipalistiske» tilnærming til profesjonsetikk er de fire velkjente prinsippene om å respektere selvbestemmelse («respect for autonomy»), gjøre godt («beneficence»), ikke skade («non-maleficence») og å ivareta rettferdighet («justice»). Disse prinsippene, som Beauchamp og Childress mener er forankret i vår felles «common sense»-forståelse av moralitet, går igjen i etiske retningslinjedokumenter for helse- og sosialfaglig arbeid over hele verden (inkludert i Norge). «Prinsipalismen» til Beauchamp and Childress representerer en pluralistisk og non-absolutistisk tilnærming til profesjonsetikk. Den er pluralistisk fordi den bygger på et syn om at det eksisterer flere forskjellige grunnleggende etiske prinsipper som ikke lar seg utlede fra ett overordnet abstrakt prinsipp (som Mills nytteprinsipp eller Kants kategoriske imperativ). Ifølge Beauchamp og Childress er hverken innholdet i eller begrunnelsen for de fire prinsippene avhengig av én bestemt type etisk teori (konsekvensialisme, deontologi eller dygdsetikk). Tvert imot er de forenlige med forskjellige teorier og kan begrunnes basert på metoden for refleksiv likevekt (Rawls, 1971). Prinsipalisme er dessuten non-absolutistisk fordi den bygger på et syn om at ingen av de fire prinsippene er absolutte, men såkalte *prima facie*-prinsipper. Det betyr at de identifiserer etisk relevante egenskaper ved situasjoner, men at de i praksis alltid vil kunne tilsidesettes avhengig av hvilke andre egenskaper situasjonene har. For eksempel vil noen situasjoner ha egenskaper som skaper *konflikter* mellom ulike prinsipper (det er umulig å både respektere selvbestemmelse og gjøre godt). Da vil ett prinsipp kunne overstyre et annet (for eksempel vil hensynet til en pasients liv og helse i noen situasjoner overstyre kravet om å innhente pasientens samtykke før et medisinsk inngrep).

Innholdsmessig er de fire prinsippene i utgangspunktet altfor abstrakte og vage til å gi helse- og sosialfagsarbeidere noe særlig handlingsveiledning i konkrete situasjoner. Den praktiske anvendelsen av dem krever derfor det Beauchamp og Childress (2013) kaller «spesifisering». Med dette sikter de til prosessen med å gi dem deskriptivt innhold, dvs. konkretisere ulike betingelser for korrekt anvendelse («når, hvorfor, hvordan») og basert på dette formulere handlingsveiledende regler (en spesifisering av prinsippet om rettferdighet vil for eksempel kunne formuleres som en regel som forbyr fordeling av helseressurser basert på sosioøkonomisk status eller etnisitet). For denne artikkelens formål er imidlertid det viktigste at den praktiske anvendelsen av de fire prinsippene krever det Beauchamp og Childress kaller «balansering». Med dette sikter de til den komparative rangeringen og vektingen av dem i overveielsesprosesser hvor målet er å komme frem til en konklusjon om hva som bør gjøres i konkrete situasjoner. Hvordan disse prinsippene *bør* balanseres i slike situasjoner, vil avhenge av egenskaper ved situasjonen. Ifølge Beauchamp og Childress er ikke en slik balansering bare et spørsmål om magefølelse og intuisjon, men om å finne gode grunner som begrunner en vurdering av den relative vekten prinsippene bør tillegges i situasjonen. Et viktig spørsmål er hvilken metode eller regler som bør benyttes for å balansere etiske prinsipper og normer i konkrete situasjoner. I bioetikk kalles dette ofte «balanseringsproblemet». Beauchamp og Childress' svar er at denne metoden simpelthen er det vi legger i å utøve *god døm-*

mekraft. God dømmekraft krever en rekke karaktertrekk («virtues»): åpenhet, ryddighet, empati, sensitivitet, oppmerksomhet og selvinnsikt, for å nevne noen. Uten slike karaktertrekk (på norsk kalt «dygder») er det vanskelig å se hvordan man skulle være i stand til å skille det som er relevant og viktig i en situasjon, fra det som er mindre relevant og viktig. En grunnleggende antagelse i prinsipalismen er derfor at balansering forutsetter det som i den dygdsetiske tradisjonen kalles «praktisk klokskap» (eller *phronesis*). Praktisk klokskap handler ikke om å følge generelle regler – slike regler vil alltid før eller senere gi feil anbefalinger fordi de ikke tar hensyn til det som kan være spesielle eller unntaksvise omstendigheter i en situasjon – men om å utøve god dømmekraft eller godt skjønn *i den konkrete situasjonen* hvor beslutningen må fattes (McDowell, 1997; Audi, 2004; Oakley & Cocking, 2001).

Beauchamp og Childress' fire prinsipper (spesifisert på den ene eller andre måten) i kombinasjon med idéen om at korrekt anvendelse av dem krever praktisk klokskap og derfor en rekke profesjonelle dygder, kan sies å utgjøre kjernen i det etiske grunnlaget for helse- og sosialfaglig arbeid. Dette er godt reflektert i yrkesetisk grunnlagsdokument for sosialarbeidere i Norge, hvor respekt for den enkeltes selvbestemmelse, integritet og menneskeverd (autonomi), ikke-diskriminering, solidaritet og rettferdighet fremholdes som kjerneverdier, mens åpenhet, omsorg, og evne til «å se hele mennesket» fremholdes som profesjonelle dygder sosialarbeidere må ha for å ivareta disse verdiene i sin profesjonsutøvelse. I neste avsnitt forklarer jeg på hvilken måte bruk av maskinlæringsteknologi i barnevernet kan komme i konflikt med Beauchamp og Childress' fire prinsipper og dermed det etiske grunnlaget for sosialfaglig arbeid.

5. Etiske utfordringer ved bruk av maskinlæring i barnevernstjenesten

Et viktig prinsipp i sosialfaglig arbeid er prinsippet om å respektere brukeres selvbestemmelse (autonomi). Det omfatter respekt for deres integritet, verdighet og menneskerettigheter. Selvbestemmelsesprinsippet skaper velkjente utfordringer i barnevernsfaglige beslutningsprosesser. Det kan være vanskelig å finne rett balanse mellom hensynet til det enkelte barnets integritet og rettigheter på den ene siden og hensynet til foreldrenes på den andre (Christiansen & Kojan, 2019; Fjeld et al., 2020). Til tross for at hensynet til barnets beste skal være en overordnet norm i den barnevernsfaglige beslutningsprosessen, kan det i praksis ofte være uklart nøyaktig når det ene hensynet må vike for det andre. Denne utfordringen forsterkes av ulike ideologiske oppfatninger av statens rolle i familielivet og fraværet av generell politisk enighet når det gjelder hvor grensene går mellom barn og foreldres rettigheter (Benbenishty et al., 2016). Utfordringen knyttet til å balansere ulike etiske hensyn i denne normativt komplekse situasjonen uten å krenke familiers rett til selvbestemmelse stiller spesielle krav til den barnevernsfaglige beslutningsprosessen. Blant annet er det en utbredt oppfatning om at det krever *en helhetlig og relasjonsbasert* tilnærming til det enkelte barn og den enkelte familie (Gillingham, 2011; Goddard et al., 1999; Munro et al., 2013). En slik tilnærming innebærer at barna og familiene møtes med forståelse og anerkjennelse på et rent mellommenneskelig plan, og at oppmerksomhet rettes ikke bare mot *bestemte aspekter* ved deres situasjon, men også mot rammene rundt denne situasjonen, herunder relasjonen mellom barnet og foreldrene og eventuelle relasjoner barnet har til andre omsorgspersoner i nærmiljøet. Også den bredere sosiale konteksten (for eksempel etnisk, religiøs og kulturell bakgrunn), barnets behov for å bevare familiemiljøet samt barnet og foreldrenes forståelser av seg selv og egne roller bør tas i betraktning (Fjeld et al., 2020).

En viktig del av kritikken mot bruk av ML-modeller i barnevernsammenheng har vært

at det undergraver en slik helhetlig, relasjonsbasert tilnærming til barna og deres familier og dermed kommer i konflikt med selvbestemmelsesprinsippet. Dette henger sammen med at modellene prioriterer prediksjon av omsorgssvikt kun basert på fravær eller forekomst av (statistiske) risikofaktorer uten å ta hensyn hverken til den bredere sosiale konteksten, rammene rundt det enkelte barnets situasjon eller det som kan være spesielle omstendigheter i denne situasjonen (Glaberson, 2019; Vaithianathan et al., 2017). Modellene tar for eksempel ikke hensyn til mulige *positive aspekter* ved relasjonen mellom barnet og foreldrene eller mellom barnet og andre omsorgspersoner i nærmiljøet. Slike aspekter kan omfatte vanskelig målbare faktorer som foreldrenes hengivenhet og kjærlighet til barnet, emosjonell støtte fra andre pårørende eller positive utviklingsmuligheter i barnets situasjon (Keddell, 2019; Spratt & Callan, 2004). Det er grunn til å anta at *samspillet* mellom slike positive faktorer og mulige risikofaktorer vil være relevant for vurderingen av barnets totale omsorgssituasjon. En studie utført av Bosk (2018) illustrerer noe av problemet. Den viste at mange barnevernsarbeidere opplevde at statistiske modelleringsverktøy hadde en tendens til å overestimere risiko på måter som slo urimelig ut for noen grupper, for eksempel ved å «straffe» familier basert på demografiske faktorer som det å ha flere enn tre barn eller å ha barn født nært i tid. Oppsummerende har én viktig kritikk mot bruk av maskinlæringsverktøy i barnevernstjenesten derfor vært at det kan føre til en ensidig fokusering på *risiko* på bekostning av en helhetsvurdering som tar i betraktning også andre relevante omstendigheter i barnets situasjon. Resultatet kan bli unødvendige inngrep fra barnevernets side som vil kunne krenke barna og familienes rett til privatliv og selvbestemmelse (Leslie et al., 2020).

Å balansere ulike etiske hensyn uten å påføre barn og foreldre slike krenkelser krever ikke bare en helhetlig tilnærming til det enkelte barn og den enkelte familie. Det er bred enighet om at det krever at barna og foreldrene også gis anledning til deltakelse og innspill underveis i beslutningsprosessen, og at de får informasjon om bakgrunnen for den barnevernsfaglige beslutningen som gjør det forståelig for dem hvordan den er begrunnet (Fjeld et al., 2020; Fenton & Kelly, 2017; Healy & Darlington, 2009). Grunnen til dette er at respekt for brukernes selvbestemmelse forutsetter at de får slik informasjon, noe som skyldes at de må ha den for å kunne inkorporere egne preferanser og verdier i de valgene de må ta om hvordan livene deres skal bli (Rubel et al., 2021). Også på dette punkt skaper bruk av ML-modeller utfordringer. Ofte kan det være vanskelig (i noen tilfeller tilnærmet umulig) for mennesker å rasjonelt rekonstruere hva som er grunnlaget for risikoprediksjonene til en ML-modell. Dette er en konsekvens av modellenes iboende matematiske kompleksitet og det faktum at de selv finner frem til variablene og relasjonene som optimaliserer tilpasningen til datasettet (Burell, 2016). Det som ofte mangler, er kausal informasjon som forklarer *hvorfor* individene i populasjonen har en bestemt risikoprofil. Konsekvensen av dette såkalte «black box»-problemet er at en beslutning om å intervenere overfor en bestemt familie kun basert på risikoprediksjonen til en ML-modell kan være vanskelig å underbygge med noe annet enn henvisning til «at familien ansees for å være i høy risiko-kategorien» (Fenton & Kelly, 2017). Familier som blir fortalt at de har blitt identifisert som «høy risiko» av en ML-modell uten ytterligere forklaring, vil naturlig nok kunne oppleve dette både som fordømmende og stigmatiserende. Det er god grunn til å tro at det vil kunne svekke tillitsforholdet deres til barnevernstjenesten (Spratt & Callan, 2004). Mangel på forståelig informasjon om grunnlaget for barnevernets beslutning om intervensjon fratar dem dessuten en viktig form for kontroll over egne liv. Det er liten tvil om at det kan innebære en krenkelse av retten deres til selvbestemmelse.

I tillegg til prinsippet om å respektere brukernes selvbestemmelse står prinsippene om å gjøre godt («beneficence») og ikke skade («non-maleficence») sentralt i de yrkesetiske

retningslinjene for sosialfaglig arbeid i de fleste land (spesifisert på litt ulike måter). Den internasjonale forskningslitteraturen er full av eksempler på hvordan ukritisk bruk av ML-modeller kan komme i konflikt med disse prinsippene. Her er ikke problemet nødvendigvis at en slik bruk kan føre til unødvendig involvering av barnevernet, men det stikk motsatte: Det kan føre til *en mangel* på involvering, som resulterer i at sårbare barn ikke får den hjelp og omsorg de trenger. Det er flere grunner til dette.

Én grunn henger sammen med fravær av faglig enighet om en presis, operasjonaliserbar definisjon av hva «omsorgssvikt» *er*, hva som utgjør *risiko* for omsorgssvikt, og hva som bør være *terskelen* for en slik risiko (Taylor & White, 2001; Welbourne, 2002; Rose & Meezan, 1996). I praksis er dette forhold som typisk vurderes av barnevernsarbeidere i det enkelte tilfellet, basert på relasjonen han eller hun har til barnet, og hvordan han eller hun tolker de konkrete omstendighetene i barnets situasjon. Til tross for at det statistisk sett kan være lav sannsynligheten for omsorgssvikt i et slikt tilfelle («lav risiko»), kan det likevel være tegn i situasjonen som gir begrunnet mistanke om omsorgssvikt. Slike tegn vil kunne oversees ved ukritisk bruk av ML-modeller med det resultat at sårbare barn ikke fanges opp. En annen viktig grunn som nevnes i litteraturen, henger sammen med faren for seleksjonsskjevheter i datasettet ML-modellene trenes på. Barnevernsdata (som utgjør en viktig del av ML-modellenes treningsgrunnlag) er ikke nødvendigvis representative for populasjonen som helhet (Dingwall et al., 2014). Dersom modellene lages for å beregne risiko for omsorgssvikt i bestemte grupper, for eksempel grupper som er overrepresenterte i dette datasettet, vil de kunne underestimere risiko for omsorgssvikt for individer i underrepresenterte grupper (Leslie et al., 2020). I verste fall kan konsekvensen bli at mange sårbare barn i underrepresenterte grupper ikke fanges opp av barnevernet. En praksis som har denne typen konsekvenser, vil opplagt være i strid med prinsippene om å gjøre godt og ikke skade og dermed innebære et brudd på yrkesetiske retningslinjer for sosialfaglig arbeid.

Et fjerde viktig prinsipp i sosialfaglig arbeid er rettferdighetsprinsippet. Spørsmål om rettferdighet oppstår i omstendigheter hvor det er knapphet på goder, ressurser eller muligheter. Rettferdighetsprinsippet kan spesifiseres på litt ulike måter avhengig av kontekst og teori, men det som er felles, er at rettferdighet dreier seg om hvordan individer behandles ut fra hva de har rettmessig krav på i slike omstendigheter. Rettferdighet krever at tilfeller som er like i alle relevante henseender, behandles likt (Miller, 2021). I de internasjonale yrkesetiske retningslinjene for sosialfaglig arbeid (NASW Code of Ethics) spesifiseres dette dithen at alle brukere skal ha lik tilgang til informasjon, tjenester og ressurser, samt lik mulighet til medvirkning i den barnevernsfaglige beslutningsprosessen.

Mange faktorer kan bidra til at bruk av ML-modeller i barnevernstjenesten kan komme i konflikt med rettferdighetsprinsippet. En hovedutfordring er den allerede nevnte faren for seleksjonsskjevheter i ML-modellenes datagrunnlag. En viktig del av dette datagrunnlaget kommer fra barnevernsregistre. Data fra barnevernsregistre er imidlertid ikke basert på et representativt utvalg som gjenspeiler den faktiske forekomsten av omsorgssvikt i hele populasjonen, men på et utvalg som gjenspeiler *undersøkte* tilfeller av omsorgssvikt (Keddell 2019; McDonell et al., 2015). Det er mange grunner til at dette kan skape seleksjonsskjevheter i data. Én grunn henger sammen med at beslutninger om å iverksette undersøkelser kan være påvirket av faktorer som *ikke* nødvendigvis har noe med omsorgssvikt å gjøre. Forskning tyder for eksempel på at alt fra lokal ressurssituasjon og kultur på det enkelte kontoret til saksbehandleres personlige verdier kan påvirke beslutninger om å iverksette undersøkelser (Bywaters et al., 2018; Fallon et al., 2013). En annen grunn er at mange tilfeller av omsorgssvikt aldri oppdages eller rapporteres. I en studie på New Zealand av en kohort barn som ble født i 1998, fant man for eksempel at 10 % av barna var registrert i barnevernsdata

som tilfeller av omsorgssvikt, mens en representativ longitudinell studie senere viste at hele 27 % utover disse «sannsynligvis» hadde blitt utsatt for omsorgssvikt (Rouland & Vaithianathan, 2018). I tillegg til disse grunnene viser forskning at en disproporsjonal andel av barn som har hatt tidligere kontakt med barnevernstjenesten, er barn av familier med dårlig økonomi, lavt utdanningsnivå, med minst én forelder født i utlandet osv. (Bywaters et al., 2014; Pelton, 1989).

Til sammen kan faktorer som dette skape skjevheter i ML-modellenes datagrunnlag. Fordi det er dette datagrunnlaget modellene trenes på, kan konsekvensen bli at faktorer som for eksempel indikerer lav sosioøkonomisk status (lav SES) og minoritetsetnisk bakgrunn, og som samtidig kan være korrelerte med former for strukturell diskriminering, får høy prediksjonsverdi i forhold til utfallene modellene bruker for å måle risiko for omsorgssvikt (Leslie et al., 2020). Modellene vil på denne måten kunne forsterke og reprodusere ulike diskriminerende sosiale strukturer i risikoprediksjonene sine. For eksempel vil de kunne overidentifisere individer i lav-SES-grupper som «høy risiko», mens de overidentifiserer individer i høy-SES-grupper som «lav risiko», til tross for at fordelingen av omsorgssvikt mellom disse gruppene kan være mye jevnere i virkeligheten enn det som gjenspeiles i barnevernsdata. En viktig kritikk mot bruken av maskinlæringsverktøy i barnevernet har derfor vært at det kan føre til en barnevernspraksis som systematisk styrer tiltak mot bestemte marginaliserte sosiale grupper (Eubanks, 2017). En slik praksis vil opplagt være urettferdig og derfor innebære et brudd på yrkesetiske retningslinjer for sosialfaglig arbeid.

6. Konklusjon

Bakgrunnen for diskusjonen i denne artikkelen har vært internasjonale erfaringer med bruk av maskinlæringsverktøy i barnevernstjenesten. Barnevernsvirkeligheten i Norge er utvilsomt forskjellig fra den man har i mange av landene hvor denne teknologien har blitt tatt i bruk. Det er imidlertid liten grunn til å tro at de etiske utfordringene som diskuteres i den internasjonale forskningslitteraturen, ikke også vil kunne oppstå her. Fremtidig ressurssituasjon samt krav til økt effektivisering kan føre til at man også i Norge vil måtte ta stilling til om barnevernet er tjent med å ta i bruk maskinlæringsverktøy på den måten man har gjort i land som USA og Storbritannia. Dette vil bare kunne avgjøres basert på en bredt anlagt nytte–kost-analyse av en slik bruk i norsk sammenheng. Mange faktorer som ikke har blitt berørt i denne artikkelen, vil naturlig nok spille inn i en slik analyse, blant annet kvaliteten på eksisterende saksbehandlingsprosedyrer og beslutningsprosesser i norsk barnevern. Den internasjonale forskningslitteraturen gir ikke grunnlag for å trekke noen entydig konklusjon i den ene eller andre retningen når det gjelder dette spørsmålet. Det den derimot viser, er at en utredning av mulige konsekvenser i forhold til yrkesetiske retningslinjer for sosialfaglig arbeid bør være en sentral del av beslutningsgrunnlaget. Av dette følger det noen krav til hva en slik utredning bør inneholde. For å unngå etiske problemer knyttet til seleksjonsskjevheter i data, bør det for eksempel nøye vurderes om datasettet ML-modellen skal trenes på, er tilstrekkelig representativt, relevant og ferskt. I tråd med anbefalinger i den internasjonale forskningslitteraturen kan det bety at det bør inneholde data som indikerer bredere beskyttelsesfaktorer og andre positive aspekter ved familier som kan balanseres mot mer risikosentrerte data (Barocas & Selbst, 2016; Lehr & Ohm, 2017). Dette ansees nødvendig for å ivareta en helhetlig tilnærming til det enkelte barn og den enkelte familie. Det bør også vurderes om informasjonen som skal utgjøre grunnlaget for modellens risikoprediksjoner, kan kommuniseres på en måte som er forståelig for barnevernsarbeidere og brukere. Dette stiller antagelig krav til modellegenskaper. Mange mener for eksempel at maskinlæringsmo-

deller som skal brukes i barnevernet, bør være lineære og monotone[2] og ikke inkludere for mange variabler, slik at de statistiske slutningene samsvarer med rimelige menneskelige forventninger (Leslie et al., 2020). Videre er det en utbredt oppfatning om at formale rettferdighetskriterier bør spesifiseres på forhånd og innarbeides i modellene, dvs. definisjoner av utvalgsrammer samt kriterier som legger føringer på fordeling av type I- og type II-feil og utfall mellom ulike grupper i populasjonen (for diskusjon av formale rettferdighetskriterier, se Leslie, 2019). Dette innebærer fastsettelse av *risikoterskel*, noe som krever balansering av modellspesifisitet og modellsensitivitet. Mens en høy risiko-terskel vil øke modellens sensitivitet, men også andelen falske negativer, dvs. andelen barn som klassifiseres som «lav risiko» til tross for at de utsettes for omsorgssvikt (type II-feil), vil en lav risiko-terskel øke modellens spesifisitet, men også andelen falske positiver, dvs. andelen barn som klassifiseres som «høy risiko» til tross for at de ikke utsettes for omsorgssvikt (type 1-feil). Fordi det er barns liv og helse som står på spill, kan det være gode etiske grunner til å prioritere modellsensitivitet fremfor modellspesifisitet og dermed tolerere en høyere andel falske positiver (Baumann et al., 2005). Samtidig kan effekten av en stor andel falske positiver være diskriminerende fordi det kan føre til mange unødvendige inngrep fra barnevernets side overfor familier som klassifiseres som «høy risiko», og som (sannsynligvis) vil tilhøre lav-SES-grupper (McQuillan, 2015). Poenget er at de statistiske avgjørelsene som gjøres her, er dypt normative og kan ha konkrete konsekvenser for barns velferd. Det betyr at de bør underbygges av gode *etiske* argumenter (blant annet basert på en bredere rettighetsbasert forståelse av rettferdighet). Det sier seg selv at en utredning som oppfyller disse kravene, forutsetter godt samarbeid mellom barnevernsarbeidere, jurister, etikere og empiriske forskere med god områdekunnskap. Alle disse aktørene bør involveres tidlig i utviklingsprosessen av ML-modeller som tiltenkes brukt i norsk barnevernstjeneste.

Det er imidlertid liten grunn til å tro at dette alene vil være nok for å unngå alle etiske utfordringer. En hypotese kan være at roten til disse utfordringene er balanseringsproblemet. Balanseringsproblemet har interessante forbindelser til det klassiske «frame»-problemet i kunstig intelligens forskning. Litt forenklet dreier «frame»-problemet seg om hvordan praktisk (inkludert moralsk) kunnskap av det slaget mennesker baserer seg på for å fatte beslutninger, kan representeres i algoritmiske systemer. I én variant oppstår problemet fordi praktisk kunnskap er noe vi hele tiden reviderer og oppdaterer i lys av faktorer og mulige konsekvenser vi identifiserer som *relevante* i konkrete situasjoner, samtidig som vi ignorerer et stort antall andre faktorer og mulige konsekvenser (Shanahan, 2016). Når det gjelder oss mennesker, krever dette praktisk klokskap. I kunstig intelligens-forskning er problemet å få algoritmiske systemer til å revidere og oppdatere informasjon i lys av hva som er relevant i konkrete situasjoner, på samme måte som vi mennesker gjør det. Dette har imidlertid vist seg å være svært vanskelig. Per i dag eksisterer det ingen bred enighet om noen løsning på dette problemet eller om det finnes en løsning (Boden, 2016; Larson, 2021). Nå er ikke en løsning på «frame»-problemet en forutsetning for å kunne bruke maskinlæringsteknologi på en trygg og ansvarlig måte i barnevernssammenheng. Men det tjener likevel som en viktig påminnelse om at maskinlæringsteknologi antagelig aldri vil kunne erstatte barnevernsarbeideres etiske skjønn. Tvert imot er det grunn til å tro at ansvarlig bruk av denne teknologien stiller økte krav til deres etisk dømmekraft og ikke minst kunnskap om de ulike etiske utfordringene som kan oppstå i profesjonell praksis.

2. At modellene er «monotone», betyr at når verdien på prediktorvariabelen endrer seg i en bestemt retning, så endrer verdien på den avhengige variabelen seg enten i den samme eller motsatte retning. Dette gjør modellens prediksjoner mer intuitive.

Litteratur

Arad-Davidson, B. & Benbenishty, R. (2008). The role of workers' attitudes and parent and child wishes in child protection workers' assessment and recommendation regarding removal and reunification. *Children and Youth Services Review*, *30*(1), 107–121. https://doi.org/10.1016/j.childyouth.2007.07.003

Athey, S. & Imbens, G. (2017). The state of applied econometrics – causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32. https://doi.org/10.1257/jep.31.2.3

Audi, R. (2004). *The good in the right: A theory of intuition and intrinsic value*. Princeton, NJ: Princeton University Press.

Baird, C. & Wagner, D. (2000). The relative validity of actuarial and consensus based risk assessment systems. *Children and Youth Services Review*, *22*(11/12), 839–871. https://doi.org/10.1016/S0190-7409(00)00122-5

Baird, C., Wagner, D., Healy, T. & Johnson, K. (1999). Risk Assessment in Child Protective Services: Consensus and Actuarial Models of Reliability. *Child Welfare*, *78*(6), 723–748.

Barnevernloven (1992). *Lov om barneverntjenester* (LOV-1992-07-17-100). Lovdata. https://lovdata.no/dokument/NL/lov/1992-07-17-100

Barocas, S. & Selbst, A. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3), 671–732.

Baumann, D.J., Law, J.R., Sheets, J., Reid, G. & Graham, J.C. (2005). Evaluating the effectiveness of actuarial risk assessment models. *Children and Youth Services Review*, *27*(5), 465–490. https://doi.org/10.1016/j.childyouth.2004.09.004

Beauchamp, T.L. & Childress, J.F. (1979). *Principles of Biomedical Ethics*. (1. utg,). Oxford University Press. ————. (2013). *Principles of Biomedical Ethics*. (7. utg.). Oxford University Press.

Becker, A. (2019). Artificial intelligence in medicine: What is it doing for us today? *Health Policy and Technology*, *8*(2), 198–205. https://doi.org/10.1016/j.hlpt.2019.03.004

Benbenishty, R., Davidson-Arad, B., López, M., Devaney, J., Spratt, T., Koopmans, C., Knorth, E.J., Witteman, C.L.M., Del Valle, J.F. & Hayes, D. (2016). Decision making in child protection: An international comparative study on maltreatment substantiation, risk assessment and interventions recommendations, and the role of professionals' child welfare attitudes. *Child Abuse and Neglect*, 49, 63–75. https://doi.org/10.1016/j.chiabu.2015.03.015

Boden, M. (2016). *AI. Its Nature and Future*. Oxford: Oxford University Press.

Bosk, E.A. (2018). What counts? Quantification, worker judgment, and divergence in child welfare decision-making. *Human Service Organizations: Management, Leadership & Governance*, 42(2), 205–224. https://doi.org/10.1080/23303131.2017.1422068

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–215. DOI: 10.1214/ss/1009213726

Burell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Bywaters, P., Brady, G., Bunting, L., Brigid, D., Featherstone, B., Jones, C., Scourfield, J., Sparks, T. & Webb, C. (2018). Inequalities in English child protection practice under austerity: A universal challenge? *Child & Family Social Work*, *23*(1), 1365–2206. https://doi.org/10.1111/cfs.12383

Bywaters, P., Brady, G., Sparks, T. & Bos, E. (2014). Inequalities in child welfare
intervention rates: the intersection of deprivation and identity. *Child & Family Social Work*, *21*(4), 452–463. https://doi.org/10.1111/cfs.12161

Christiansen, Ø. & Kojan, B.H. (2019). *Beslutninger i barnevernet*. Universitetsforlaget.

Crea, T.M. (2010). Balanced decision making in child welfare: Structured processes informed by multiple perspectives. *Administration in Social Work*, *34*(2), 196–212. https://doi.org/10.1080/03643101003609529

Cuccaro-Alamin, S., Foust, R., Vaithianathan, R. & Putnam-Hornstein, E. (2017). Risk assessment and decision making in child protective services: Predictive risk modeling in context. *Children and Youth Services Review*, 79, 291–298. https://doi.org/10.1016/j.childyouth.2017.06.027

D'Andrade, A., Benton, A. & Austin, M.J. (2005) *Risk and Safety Assessment in Child Welfare: Instrument Comparisons*. Berkeley, CA: Bay Area Social Services Consortium.

Dawes, R. M., Faust, D. & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science,* 243, 1668–74. https://doi.org/10.1126/science.2648573

Dingwall, R., Eekelaar, J. & Murray, T. (2014). *The protection of children*. Quid Pro Books.

Domingos, P. (2012). A Few Useful Things to Know about Machine Learning. *Communications of the ACM, 55*(10), 78–87.

Daley, D., Bachmann, M., Bachmann, B.A., Pedigo, C., Bui, M.-T. & Coffman, J. (2016). Risk terrain modeling predicts child maltreatment. *Child Abuse & Neglect*, 62, 29–38. https://doi.org/10.1016/j.chiabu.2016.09.014

English, D.J. & Pecora, P.J. (1994). Risk Assessment as a Practice Method in Child Protective Services. *Child Welfare*, *73*(5), 451–473.

Eubanks, V. (2017). *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*. St. Martin's Press.

Fallon, B., Chabot, M., Fluke, J., Blackstock, C., MacLaurin, B. & Tonmyr, L. (2013). Placement decisions and disparities among Aboriginal children: Further analysis of the Canadian incidence study of reported child abuse and neglect part A: Comparisons of the 1998 and 2003 surveys. *Child Abuse & Neglect, 37*(1), 47–60. https://doi.org/10.1016/j.chiabu.2012.10.001

Fenton, J. & Kelly, T. (2017). 'Risk is King and Needs to take a Backseat!' Can social workers' experiences of moral injury strengthen practice? *Journal of Social Work Practice, 31*(4), 461–75. https://doi.org/10.1080/02650533.2017.1394827

Fjeld R., Sasaoka, K., Madland, S., Skivenes, M., Øvreeide, H. & Ordermann, H. (2020). *Det kan høres ut som en bagatell, men ikke for meg da. Gjennomgang av ti særlige konfliktfylte barnevernssaker i Bergen kommune*. Rapport til byrådet i Bergen. https://www.bergen.kommune.no/politikere-utvalg/api/fil/bk360/6513561/9-Barnevernrapport-fra-Fjeld-utvalget-2020

Gillingham, P. (2011). Decision-making tools and the development of expertise in child protection practitioners: Are we 'just breeding workers who are good at ticking boxes'? *Child & Family Social Work*, *16*(4), 412–21. https://doi.org/10.1111/j.1365-2206.2011.00756.x

Glaberson, S. (2019). Coding Over the Cracks: Predictive Analytics and Child Protection. *Fordham Urban Law Journal*, *46*(2), 307–363.

Goddard, C.R., Saunders, B.J, Stanley, J.R. & Tucci, J. (1999). Structured risk assessment procedures: Instruments of abuse? *Child Abuse Review*, *8*(4), 251–263. https://doi.org/10.1002/(SICI)1099-0852(199907/08)8:4%3C251::AID-CAR543%3E3.0.CO;2-M

Grove, W.M. & Meehl, P.E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Healy, K. & Darlington, Y. (2009). Service user participation in diverse child protection contexts: Principles for practice. *Child and Family Social Work*, *14*(4), 420–30. https://doi.org/10.1111/j.1365-2206.2009.00613.x

Jergeby, U. & Soydan, H. (2002). Assessment processes in social work practice when children are at risk: A comparative cross-national vignette study. *Journal of Social Work Research and Practice*, *3*(2), 127–144.

Kahneman, D. & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515–526. DOI: 10.1037/a0016755

Keddell, E. (2016). Substantiation, decision-making and risk prediction in child protection systems. *Policy Quarterly*, *12*(2), 46–59. http://dx.doi.org/10.26686/pq.v12i2.4587

Keddell, E. (2019). Algorithmic justice in child protection: Statistical fairness, social justice and the implications for Practice. *Social Sciences, 8*(281), 1–22. https://doi.org/10.3390/socsci8100281

Kuhse, H. & Singer, P. (1998). *A companion to bioethics.* (2. utg.) Wiley-Blackwell.

Kvello, Ø. (2015). *Barn i risiko.* Gyldendal akademisk.

Larson, E. (2021). *The myth of artificial Intelligence. Why computers can't think the way we do.* Belknap Press.

Lehr, D. & Ohm, P. (2017). Playing with the data: What legal scholars should learn about machine learning. *US Davis Law Review*, *51*(2), 653–717.

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *The Alan Turing Institute.* https://dx.doi.org/10.2139/ssrn.3403301

Leslie, D., Holmes, L., Hitrova, C. & Ott, E. (2020). *Ethics Review of Machine Learning in Children's Social Care.* https://whatworks-csc.org.uk/research-report/ethics-review-of-machine-learning-in-childrens-social-care/

Lyons, P., Doucek, H.J. & Wodarski, J.S. (1996). Risk assessment for child protective services: A review of the empirical literature on instrument performance. *Social Work Research*, *20*(3), 143–155. https://doi.org/10.1093/swr/20.3.143

Marshall, D.B. & English, D.J. (2000). Neural network modeling of risk assessment in child protective services. *Psychological Methods*, *5*(1), 102–124. https://doi.org/10.1037/1082-989X.5.1.102

McDowell, J. (1997). Virtue and Reason. I R. Crisp & M. Slote (Red.), *Virtue Ethics* (s. 141–162). Oxford University Press.

McDonell, J. R., Ben-Arieh, A. & Melton, G.B. (2015). Strong Communities for Children: Results of a Multi-Year Community-Based Initiative to Protect Children from Harm. *Child Abuse & Neglect*, *41*, 79–96. https://doi.org/10.1016/j.chiabu.2014.11.016

McQuillan, D. (2015). Algorithmic states of exception. *European Journal of Cultural Studies*, *18*(4/5), 564–76. https://doi.org/10.1177/1367549415577389

Meehl, P.E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis, MN: University of Minnesota Press.

Miller, D. (2021). Justice. *The Standford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/fall2021/entries/justice/

Munro, E., Taylor, J. & Bradbury-Jones, C. (2013). Understanding the Causal Pathways to Child Maltreatment: Implications for Health and Social Care Policy and Practice. *Child Abuse Review, 23*(1), 61–74. https://doi.org/10.1002/car.2266

Munro, E. (1999). Common errors in reasoning in child protection work. *Child Abuse & Neglect*, *23*(8), 745–758. https://doi.org/10.1016/S0145-2134(99)00053-8

NOU 2000: 12 (2000). *Barnevernet i Norge.* Barne- og familiedepartementet.

Oakley, J. & Cocking, D. (2001). *Virtue Ethics and Professional Roles.* Cambridge University Press.

Packard, T. (2016). Literature review: Predictive analytics in human services. Southern Area Consortium of Human Services. https://1library.net/document/z1302mpq-literature-review-predictive-analytics-in-human-services.html

Pelton, L.H. (1989). *For reasons of poverty: A critical analysis of the public child welfare system in the United States.* Praeger.

Plous, S. (1993). *The Psychology of Judgment and Decision-Making.* McGraw-Hill.

Rawls, J. (1971). *A theory of justice.* Belknap Press.

Reamer, F.G. (1985). The emergence of bioethics in social work. *Health & Social Work*, *10*(4), 271–281. https://doi.org/10.1093/hsw/10.4.271

Rose, S.J. & Meezan, W. (1996). Variations in perceptions of child neglect. *Child Welfare*, *75*(2), 139–160.

Rossi, P., Schuerman, J. & Budde, S. (1996). *Understanding child maltreatment decisions and those who make them*. Chapin Hall Center for Children at the University of Chicago.

Rouland, B. & Vaithianathan, R. (2018). Cumulative Prevalence of Maltreatment among New Zealand Children, 1998–2015. *American Journal of Public Health*, *108*(4), 511–13.

Rubel, A., Castro, C. & Pham, A. (2021). *Algorithms & Autonomy: The Ethics of Automated Decision Systems*. Cambridge University Press

Shanahan, M. (2016). The frame problem. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2016/entries/frame-problem/

Spratt, T. & Callan, J. (2004). Parents' Views on Social Work Interventions in Child Welfare Cases. *The British Journal of Social Work*, *34*(2), 199–224. https://doi.org/10.1093/bjsw/bch022

Taylor, C. & White, S. (2001). Knowledge, truth and reflexivity: The problem of judgement in social work. *Journal of Social Work*, *1*(1), 37–59. https://doi.org/10.1177/146801730100100104

Vaithianathan, R., Jiang, N., Maloney, T. & Putnam-Hornstein, E. (2017). Developing Predictive Risk Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation. Auckland: Centre for Social Data Analytics. https://www.alleghenycountyanalytics.us/wp-content/uploads/2017/04/Developing-Predictive-Risk-Models-package-with-cover-1-to-post-1.pdf

Van der Put, C.E., Bouwmeester-Landweer, M.B.R., Landsmeer-Beker, E.A., Wit, J.M., Dekker, F.W., Kousemaker, N.P.J. & Baartman, H.E.M. (2017). Screening for potential child maltreatment in parents of a newborn baby: The predictive validity of an Instrument for early identification of Parents at Risk for child Abuse and Neglect (IPARAN). *Child Abuse & Neglect*, 70, 160–68. https://doi.org/10.1016/j.chiabu.2017.05.016

Welbourne, P. (2002). Culture, children's rights and child protection. *Child Abuse Review*, *11*(6), 345–35. https://doi.org/10.1002/car.772

Ægisdóttir, S., White, M.J., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook, R.S., Nichols, N.C., Lampropoulos, G.K., Walker, B.S., Cohen, G. & Rush, J.D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341–382. https://doi.org/10.1177/0011000005285875