## 10.7 Bias og algoritmisk transparens

Brugen af algoritmer giver også en anden etisk konflikt, nemlig en konflikt mellem effektivitet og retfærdighed. Det har i århundrede været helt almindeligt at bruge datadrevne sandsynlighedsberegninger til at støtte beslutninger, der går ud på at vurdere en form for risiko. Hvis man fx tegner en ulykkesforsikring, vil forsikringsselskabet bruge deres viden om, hvor hyppigt folk, der på relevante træk ligner dig, kommer til skade, til at fastsætte forsikringspræmien. Det er ret ukontroversielt. Det er straks mere kontroversielt, at man bl.a. i det amerikanske retssystem siden starten af den 20. århundrede har brugt lignende beregninger til at vurdere kriminelles risiko for at begå ny kriminalitet – og at den type beregninger er blevet brugt i vurderingen strafudmåling, prøveløsladelser m.v. (Carlson, A. (2017). The Need for Transparency in the Age of Predictive Sentencing Algorithms. Iowa Law Review, 103(1), 303–329.). I de sidste årtier har lettere adgang til store datamængder og fremkomsten af diverse machine learning teknikker gjort det lettere at udvikle den type systemer og diverse former for prædiktive algoritmer bruges i dag på en lang række områder fra reklamer til sortering af jobansøgninger.

Det er dog ikke helt uproblematisk at bruge prædiktive algoritmer. Som vi så tidligere, er der en række epistemiske problemer forbundet med machine learning. Der er typisk en meget direkte sammenhæng mellem epistemiske og etiske overvejelser; hvis de etiske overvejelser inddrager utilitaristiske komponenter er det i høj grad væsentligt at forstå, hvor godt en given algoritme i praksis virker. Ville den etiske vurdering af algoritmen til forudsigelse af frafald i gymnasiet fx falde anderledes ud, hvis systemet havde en træfsikkerhed på 67%? Der er imidlertid også et anden og vanskeligere etisk aspekt specielt hvis man bruger algoritmiske forudsigelser til at træffe afgørelser, der har betydning for enkeltindividers muligheder. Algoritmen til forudsigelse af frafald inddrog således direkte elevernes etnicitet som en faktor i beregningen, men hvad nu hvis algoritmen havde en bias så den systematisk gav elever med en etnicitet en højre risikovurdering end elever med en anden?

Spørgsmålet om diskrimination i algoritmer er navnlig blevet diskuteret i forbindelse med den såkaldte COMPAS-algoritme, der bruges i det amerikanske retssystem til vurdering af kriminelles tilbagefaldsrisiko. En gennemgang af systemet viste imidlertid en systematisk skævhed i de fejl, algoritmen lavede. Således havde farvede, der ikke senere begik ny kriminalitet, en langt højere chance for fejlagtigt at få en høj risikoscore end hvide, og hvide, der havde omvendt en højere chance for fejlagtigt at få en lav score end farvede (Angwin m.fl., 2016).

Hvis man imidlertid tager udgangspunkt i de to kategorier høj- og lav-risiko, og undersøtter ved at blive publiceret, hvis der bruges et mandligt forfatternavn end hvis der bruges et kvindeligt. Da det udelukkende er artiklens kvalitet og ikke forfatterens køn der skal afgøre, om en artikel er publicerbar eller ej, er der her tale om direkte (køns)diskrimination. I andre tilfælde er sagen imidlertid mindre klar. Fx lever mænd kortere end kvinder (en tommelfingerregel siger, at det er lige så farligt at være mand som at være overvægtig). Så hvis man tegner en livsforsikring, burde mænds præmier være højere end kvinders. Hvis præmien er ens for begge køn kommer de i gennemsnit fornuftige kvinder til at betale for de i gennemsnit ufornuftige mænds dumheder. Man kan sige, at køn i forhold til livsforsikring er en informationsbærende dimension, så hvis man afskaffer algoritmen fra at bruge din dimension, vil den naturligvis blive mindre effektiv. Og i dette tilfælde er det kvinderne, der vil komme til at betale for den manglende effektivitet. At man imidlertid inddrager køn i beregningen af pensionspræmie, vil en mand, der opfører sig fornuftigt og tager sit helbred

seriøst, komme til at betale en højre præmie, udelukkende fordi han har samme køn som en gruppe, der – gennemsnitligt – opfører sig ufornuftigt. Og hvad gør man så? Er det mest retfærdige at kønsdiskriminere i dette tilfælde (og lade de fornuftige mænd betale prisen) eller er det mest retfærdigt at undlade at kønsdiskriminere (og lade kvinderne betale prisen)?

Det bringer os tilbage til COMPAS-algoritmen. I den population, hvor algoritmen er blevet testet, har farvede fanger reelt en større risiko for at begå ny kriminalitet. Når algoritmen diskriminerer på baggrund af etnicitet er det derfor blot et udtryk for, at den har identificeret en informationsbærende dimension i datasættet. Og lige som ovenfor vil vi have valgt mellem at forhindre algoritmen i at bruge informationen, hvormed den vil blive mindre effektiv i den forstand at dens overordnede fejlrato vil stige, eller at tillade den at bruge informationen, hvorved den vil diskriminere enkeltindivider, således at farvede, der ikke vil begå ny kriminalitet har en meget større risiko for at få en høj kriminalitetsscore end ikke-farvede. Hvilket af de to scenarier forekommer dig mest retfærdigt, når du står bag uvidenhedens slør?

Der er dog også væsentlige forskelle på de to cases. Specielt kan man argumentere for, at den gennemsnitligt høje kriminalitetsrate blandt farvede til dels skyldes strukturelle forhold i det amerikanske samfund; pga. en generel diskrimination mod farvede har de sværere ved at få uddannelser og job end ikke-farvede, og derfor vil de hyppigere ende i kriminalitet. En algoritme som COMPAS, der har en større tilbøjelighed til at holde farvede i fængsel end ikke-farvede, vil bidrage til fastholde denne strukturelle ulighed. En datadreven algoritme afspejler virkeligheden, som den er, men hvis virkeligheden er racistisk, vil algoritmen også blive det, og hvis algoritmen bliver brugt uden forbehold, kan den bidrage til at fastholde historisk betingede diskriminerende strukturer. Der er dermed ikke noget, der tyder på, at mænds kortere levealder er et udtryk for strukturel diskrimination. Så med andre ord er det vigtigt, at se på den kontekst, en algoritme skal indgå i, når man overvejer, om den er retfærdig.

**Pandoras black box**

I en traditionel statistisk model vil man have nogenlunde styr på de variable, modellen inddrager. Det er dog værd at bemærke, at tingene selv med traditionelle algoritmer ikke altid er så klare; COMPAS-algoritmen inddrager således ikke direkte etnicitet som variabel. Den inddrager til gengæld flere variable, der er tæt korrelerede med etnicitet, og dermed indgår etnicitet indirekte i algoritmen. Det gør naturligvis tingene mere besværligt, at man på den måde kan komme til at bruge variable, man ikke direkte har indbygget i sin algoritme.

Det problem bliver imidlertid kraftigt forstærket, hvis man træner en algoritme med machine learning. Her har vi styr på de data, algoritmen er trænet med, og vi kan opstille forskellige mål for, hvor godt den virker, men resten er en black box; vi aner reelt ikke hvorfor den virker, og som vi berørte ovenfor, kan machine learning algoritmer med lethed diskriminere langs variable, der er af en helt anden type den de data, vi har trænet algoritmen med. Hvis vi vil sikre, at de algoritmer, vi udvikler, ikke diskriminerer på en etisk kritisabel måde, bliver vi nødt til at gøre dem transparente. Det er imidlertid ikke altid let, at gøre en blackbox transparent!