Dataetisk Råd 2024

København

Dataetisk Råd er en offentlig myndighed ved Nationalt Center for Etik, som rådgiver og skaber debat om etisk udvikling og anvendelse af digitale løsninger, data og kunstig intelligens

Projektet har i Dataetisk Råd været forankret i en referencegruppe bestående af:

Mathias Bukhave

Peter Damm

Mikkel Flyverbom

Camilla Gregersen

Rikke Frank Jørgensen

Eik Møller

Projektet er udarbejdet med sekretariatsbistand fra projektleder Frej Klem Thomsen, samt Thomas Nørskov Kjølbye og Asta Maria Jarner Bjerre

Indhold

1. Sammenfatning 6

2. Indledning: Velkommen til den nye arbejdsplads 9

2.1. Indsamling af medarbejderdata og automatiserede beslutningssystemer på arbejdspladsen 10

2.2. Hvilke dataetiske udfordringer rejser teknologien? 11

2.3. Rapporten i det dataetiske landskab 12

3. Digitale værktøjer til indsamling og anvendelse af medarbejderdata – teknologier og begreber 16

3.1. Internationale erfaringer med indsamling og anvendelse af medarbejderdata 16

3.2. Indsamling og anvendelse af medarbejderdata på danske arbejdspladser 19

3.3. Hvilke medarbejderdata kan indsamles på arbejdspladsen? 20

3.4. Hvordan anvendes automatiserede beslutningssystemer på arbejdspladsen? 23

4. Juridiske rammer 29

4.1. Centrale love og aftaler 29

4.2. Seks centrale juridiske rammer for indsamling og anvendelse af medarbejderdata 33

4.3. Fremtidig regulering af området 38

4.4. Dataetik og jura i denne rapport 40

5. Etiske udfordringer ved indsamling af medarbejderdata – overvågning og privatliv på arbejde 43

5.1. Hvad er privatliv på arbejdet? 43

5.2. To teorier om privatliv for personlig information 45

5.3. Hvad er retten til privatliv? 48

5.4. Hvorfor privatliv på arbejdspladsen? 52

5.5. Hvad er det særlige ved privatliv på arbejdspladsen? 58
5.6. Hvilken rolle spiller samtykke til indsamling af medarbejderdata? 59

6. Etiske udfordringer ved anvendelse af automatiserede beslutningssystemer på arbejdspladsen – fejlvurderinger, bias og adfærd 65

6.1. Beslutningssystemer begår fejl 66
6.2. Medarbejdertilfredshed, produktivitet og gaming-effekter 70
6.3. Algoritmisk bias i beslutningssystemer 72
6.4. Hvordan opstår algoritmisk bias? 81
6.5. Hvornår er algoritmisk bias uetisk? 89
6.6. Et menneske i kredsløbet 93

7. Tre illustrationer af dataetiske udfordringer ved indsamling og anvendelse af medarbejderdata på arbejdspladsen 100

7.1. Opgaveløsning 100
7.2. Rekruttering 102
7.3. Trivsel 104

Litteratur 107

1. Sammenfatning

Denne rapport handler om de dataetiske udfordringer som kan opstå, når en arbejdsplads bruger nye digitale værktøjer til at indsamle medarbejderdata, og støtte eller træffe beslutninger med automatiserede beslutningssystemer. Rapporten fokuserer især på hvordan indsamling af medarbejderdata kan udfordre medarbejderes ret til privatliv, og hvordan anvendelse af automatiserede beslutningssystemer kan føre til fejlagtige beslutninger og beslutninger præget af algoritmisk bias.

De dataetiske udfordringer, som nye digitale teknologier skaber, er på engang velkendte og underbelyste. Udfordringerne er anerkendte i mange dataetiske retningslinjer og anbefalinger, men de er komplekse, og bliver sjældent behandlet i detaljer på en måde, som gør det muligt at foretage præcise dataetiske vurderinger (se kapitel 2). **Denne rapport folder de dataetiske udfordringer ud, og viser hvilke dataetiske hensyn, der er på spil.**

Rapportens redegørelse for teknologien viser, at **digitale ledelsesværktøjer i disse år udbredes med stor hast, også på danske arbejdspladser** (se kapitel 3). Disse digitale værktøjer kan indsamle medarbejderdata, og anvende disse data i automatiserede beslutningssystemer. Automatiserede beslutningssystemer kan bruges til mange formål på arbejdspladsen, inklusiv at støtte eller udføre beslutninger med alvorlige konsekvenser om for eksempel ansættelse og afskedigelse.

Rapporten behandler også hvordan de juridiske rammer sætter grænser for arbejdspladsers lovlige indsamling og anvendelse af medarbejderdata (se kapitel 4). Analysen viser i den forbindelse, at dataetikken kan stille andre krav end lovgivningen – **indsamling og anvendelse af medarbejderdata kan være uetisk selvom den er lovlig.** Det er derfor afgørende at overveje, hvordan arbejdspladser bør anvende de digitale værktøjer i lyset af dataetiske hensyn.

En central etisk udfordring ved indsamling af medarbejderdata er, at denne indsamling reducerer medarbejderes privatliv. Rapporten peger på **fem dataetiske grunde til, at beskytte medarbejderes privatliv på arbejdet:** for at forhindre (1) ydmygelse, (2) observationsstress og (3) afskrækkelse, for at (4) beskytte værdifulde sociale relationer, og for at (5) begrænse medarbejderes sårbarhed overfor skadelige måder at bruge data på (se kapitel 5).

Privatliv kan være særligt vigtigt i arbejdssammenhæng, fordi arbejdspladser har mulighed for meget detaljeret og omfattende indsamling af medarbejderdata, og fordi der kan være snævre grænser for medarbejderes adfærd, og omkostningsfulde sanktioner (se kapitel 5). Medarbejderes samtykke til indsamling og anvendelse af medarbejderdata kan gøre en etisk forskel, hvis det er autonomt, informeret, og frit. Medarbejdere vil imidlertid ofte være underlagt en mulig eller konkret trussel om sanktioner, hvis de afviser samtykke, som betyder, at **medarbejdere ikke kan give et frit samtykke til indsamling og anvendelse af medarbejderdata.**

En anden central udfordring knytter sig til arbejdspladsers stigende anvendelse af automatiserede beslutningssystemer. **Automatiserede beslutningssystemer kan begå fejl og have negative sideeffekter** på medarbejderes motivation og præstation, blandt andet i form af såkaldt "gaming", hvor medarbejdere tilpasser deres adfærd til den måde arbejdspladsen indsamler og anvender data, på bekostning af arbejdspladsens egentlige mål (se kapitel 6).

Automatiserede beslutningssystemer kan også have algoritmisk bias, hvor medarbejdere systematisk behandles forskelligt (se kapitel 6). Et system kan have flere forskellige slags algoritmisk bias, og der findes mange mulige kilder til algoritmisk bias. **I mange tilfælde er det vanskeligt eller endda umuligt at undgå algoritmisk bias i et automatiseret beslutningssystem.** En dataetisk vurdering må tage stilling til både hvornår et system har for meget bias til, at det bør anvendes, og hvordan man skal vælge mellem forskellige bias. Rapporten peger på, at **algoritmisk bias kan være uetisk, når den bevarer eller skaber ulighed i personers muligheder for at leve gode liv eller når den gør skade.**

Afslutningsvis behandler rapporten automatiseret beslutningsstøtte, hvor et *"menneske-i-kredsløbet"* godkender eller udfører den endelige beslutning. Rapporten viser, at automatiseret beslutningsstøtte møder sine egne udfordringer, fordi beslutningerne kan være præget af støj, automatiseringsbias, algoritmisk aversionsbias, konfirmationsbias, og anker-effekter (se kapitel 6). Konsekvensen af disse effekter er, at **dataetiske fordele og ulemper ved automatiseret beslutningsstøtte som alternativ til rent menneskelige og fuldt automatiserede beslutninger må vurderes fra sag til sag.**

Opmærksomhed på disse dataetiske udfordringer er afgørende, når ledelse, medarbejdere og beslutningstagere skal overveje, hvordan man dataetisk indsamler og anvender medarbejderdata på arbejdspladsen. Ved at afklare udfordringerne håber Dataetisk Råd at skabe et stærkere fundament for de dataetiske samtaler, vi som samfund må tage, om hvordan vi bør bruge de nye teknologiers muligheder.

Dataetisk Råd

Johan Busse
Formand for Dataetisk Råd

17. Januar 2024

Centrale pointer:

Kapitel 3

- Digitale værktøjer til indsamling af medarbejderdata udbredes med stor hast på arbejdspladser. Mange værktøjer gør det muligt automatisk at støtte eller træffe ledelsesbeslutninger.

Kapitel 4

- Der er vide muligheder for lovligt at indsamle og anvende medarbejderdata, men lovlig indsamling og anvendelse er ikke nødvendigvis etisk.

Kapitel 5

- Der er dataetiske grunde til at beskytte medarbejderes privatliv, når kendskab til medarbejderdata kan være ydmygende, stressende, eller afskrækkende, skade værdifulde sociale relationer, eller gøre medarbejdere sårbare for skadelige handlinger.
- Ansættelsesforholdet vil ofte begrænse medarbejderes mulighed for at afgive frit samtykke til indsamling og anvendelse af medarbejderdata.

Kapitel 6

- Automatiserede beslutningssystemer kan begå fejl, have negative effekter på medarbejderes motivation og præstation, og føre til såkaldt "gaming" af systemet.
- Automatiserede beslutningssystemer vil ofte uundgåeligt have algoritmisk bias, som kan skabe ulighed i medarbejderes muligheder eller skade medarbejdere.
- Automatiseret beslutningsstøtte, hvor der er "et menneske i kredsløbet", rejser særegne udfordringer, i form af støj, automatiseringsbias, algoritmisk aversionsbias, konfirmationsbias, og anker-effekter.

2. Indledning: Velkommen til den nye arbejdsplads

IDA Forsikring er et forsikringsselskab ejet af fagforeningen IDA. Fagforeningen IDA organiserer i Danmark ca. 150.000 personer, som på forskellig vis har stærke tekniske kompetencer, blandt andet ingeniører og it-professionelle. I det lys er det ikke overraskende, at IDA Forsikring forsøger at være på forkant med den teknologiske udvikling.

Ikke desto mindre vakte det opsigt, da det i oktober 2022 kom frem, at IDA Forsikring havde indført en ny slags kunstig intelligens på arbejdspladsen.[1] Systemet lyttede med, når medarbejdere i selskabets kundeservice talte med kunderne. En kombination af stemmegenkendelse og en sprogmodel gjorde det muligt at analysere samtalerne, og en model for de egenskaber, som karakteriserer en god samtale, gjorde det muligt automatisk at evaluere medarbejdernes indsats med at hjælpe kunderne. Resultatet var en løbende måling af hver medarbejders præstation, som blev offentliggjort på en storskærm i det åbne kontorlandskab.[2]

Den teknologi, som IDA Forsikring introducerede, er et eksempel på et digitalt værktøj til indsamling af medarbejderdata og automatiseret beslutningsstøtte. Denne type digitale værktøjer bliver i disse år hastigt mere almindelige på danske arbejdspladser. Anvendelse af den nye teknologi kan skabe fordele, men også udfordringer. Det digitale værktøj, som IDA Forsikring benyttede, kan støtte medarbejdere i løbende at reflektere over kvaliteten af deres arbejde, og bruge disse indsigter til at forbedre deres præstation. Det kan også potentielt give en mere præcis og rimelig evaluering af hver medarbejders indsats end konventionelle metoder til at måle præstation.

Men det digitale værktøj var også afhængigt af, at IDA Forsikring indsamlede og anvendte store mængder data om kunder og medarbejdere. Og systemet kunne have svært ved at forstå sproglige nuancer, for eksempel ved forskelle i tonefald eller brug af ironi.

På den måde illustrerer IDA Forsikrings introduktion af teknologien både en igangværende udvikling på det danske arbejdsmarked, og de muligheder og udfordringer, som denne udvikling skaber. Digitale værktøjer, som indsamler data om medarbejdere, og systemer til at analysere disse data, og støtte eller automatisere beslutninger på arbejdspladsen, er et afgørende teknologisk nybrud, som kan komme til at påvirke mange danskeres hverdag fremover. Derfor er det vigtigt at tænke grundigt over, hvordan teknologien kan bruges, så den på engang realiserer de bedste potentialer og undgår de værste faldgruber.

---

1 Jens Bostrup, "Det ligner et almindeligt storrumskontor, men nyskabelsen bliver synlig, når man vender sig om" *Politiken* (København), October 18, 2022, https://politiken.dk/viden/art9018913/Det-ligner-et-almindeligt-storrumskontor-men-nyskabelsen-bliver-synlig-n%C3%A5r-man-vender-sig-om; Therese Moreau, "IDA overvåger medarbejdere med uigennemskuelig AI: "Vi ved, at systemet ikke forstår alt"" *ING/Datatech*, October 25, 2022, https://pro.ing.dk/datatech/artikel/ida-overvaager-medarbejdere-med-uigennemskuelig-ai-vi-ved-systemet-ikke-forstaar.

2 IDA Forsikring oplyser, at de efterfølgende har fjernet storskærmen.

Denne rapport fra Dataetisk Råd præsenterer en række af de centrale dataetiske udfordringer og perspektiver, som arbejdspladser, medarbejdere, interessenter og beslutningstagere, må forholde sig til, når de skal vurdere den nye teknologi.

2.1. Indsamling af medarbejderdata og automatiserede beslutningssystemer på arbejdspladsen

Rapporten begynder med at skitsere, hvordan moderne arbejdspladser kan indsamle og anvende medarbejderdata. Indsamling og anvendelse af disse data kan være attraktivt for en arbejdsplads, fordi dette kan hjælpe med at effektivisere eller forbedre beslutninger. De indsamlede data kan informere beslutningsprocesser, og automatiserede beslutningssystemer kan frigive menneskelige ressourcer, hjælpe med at prioritere indsatser, og i nogle tilfælde lave hurtigere, mere komplekse og mere præcise vurderinger end mennesker.

Indsamling og anvendelse af data i bred forstand er selvsagt ikke nyt – arbejdspladser har altid indsamlet og anvendt data til at kvalificere beslutninger. Det som er nyt, er dels anvendelsen af nye digitale teknologier, som kan indsamle medarbejderdata i et omfang, som er væsentligt større, end man tidligere har kunnet, og dels automatiserede beslutningssystemer, som kan anvende disse data på mere komplekse måder, end det tidligere har været muligt. Det kan være GPS, som registrerer medarbejderens fysiske placering i løbet af arbejdsdagen, software som registrerer hvad medarbejderen laver på sin arbejdscomputer og på hvilke tidspunkter, eller som i det indledende eksempel kunstig intelligens, som analyserer medarbejderens telefonsamtaler og registrerer længden og ordvalg. Når Dataetisk Råd i denne rapport behandler indsamling og anvendelse af data er det med fokus på disse nye teknologier, de muligheder de giver, og de udfordringer de rejser.

En arbejdsplads kan indsamle og anvende mange andre slags data end data fra medarbejdere. Det er eksempelvis i stigende grad almindeligt, at indsamle og anvende data om brugere eller kunder. Kunde- og brugerdata kan anvendes til for eksempel at tilpasse og målrette produkter. For visse virksomheder, herunder flere af de såkaldte tech-giganter, er indsamling og analyse af brugerdata i vid udstrækning kernen i virksomhedens forretningsmodel. Denne rapport fokuserer imidlertid på indsamling og anvendelse af data om arbejdspladsens egne medarbejdere. De informationer som udledes af medarbejderdata kan anvendes til for eksempel at fordele arbejdsopgaver, evaluere medarbejderes præstation, eller prioritere ansøgere ved rekruttering. Disse data udgør også grundlaget for, at beslutninger kan automatiseres helt eller delvist, det vil sige, at arbejdspladsen kan anvende software til at træffe eller anbefale beslutninger.

I den første del af denne rapport præsenterer vi de nye teknologier, hvordan de er blevet indført på det danske arbejdsmarked, samt de juridiske rammer for hvordan teknologierne må anvendes. I **kapitel tre** giver vi et overblik over teknologien, hvordan den er blevet udbredt, og hvilke muligheder den giver for at indsamle og anvende medarbejderdata. Vi introducerer og definerer også mange af de begreber, som vi anvender til at beskrive teknologien. I **kapitel fire** kigger vi på hvordan indsamling og anvendelse af medarbejderdata er reguleret, ved at sammenfatte seks centrale juridiske rammer, som findes i dansk og international

lovgivning. Vi diskuterer også forskellen på dataetik og jura, og hvilken rolle dataetik kan spille for en arbejdsplads, som overholder gældende lov.

2.2. Hvilke dataetiske udfordringer rejser teknologien?

Udbredelsen af de nye teknologier er i nogle sammenhænge blevet kritisk modtaget. Forskere, NGO'er og fagforeninger har givet udtryk for bekymringer, og peget på en række dataetiske udfordringer, som teknologien rejser. To centrale dataetiske udfordringer er, at indsamling af medarbejderdata reducerer medarbejderes privatliv, samt at automatiserede beslutningssystemer har risici for bias og fejl.

De fleste ved godt, at almindelige personer efterlader store mængder digitale fingeraftryk i hverdagen, fordi de har berøring med mange forskellige digitale systemer gennem computerens browser og telefonens mange apps. Mange er også i nogen udstrækning bekymrede for, hvordan disse data påvirker privatlivet.[3] Men bekymringen bliver nok endnu mere konkret, når det er arbejdspladsen, som indsamler data. Dels fordi en arbejdsplads har mulighed for at samle store mængder data om den enkelte, identificerbare medarbejder, og dels fordi arbejdspladsen kan anvende sådanne data på måder, der har umiddelbare konsekvenser for medarbejderen. Det kan imidlertid være svært, at sige præcist, hvordan indsamling af medarbejderdata udgør en dataetisk udfordring, hvis det er uklart hvordan man skal forstå medarbejderes ret til privatliv. Hvad vil det sige, at en medarbejder har og ikke har privatliv, og hvad er det etiske grundlag for retten til privatliv på en arbejdsplads?

Den samme kombination af kendte og nye udfordringer gør sig gældende, når arbejdspladser anvender medarbejderdata i automatiserede beslutningssystemer. Det seneste årti er det blevet stadigt mere almindeligt at møde sådanne systemer flere og flere steder, fra GPS'en som beregner og foreslår den hurtigste rute, over spamfilteret, som frasorterer irrelevant e-mail, og til digitale tjenester, som foreslår indhold, der kan være interessant for den enkelte bruger. På en arbejdsplads kan sådanne systemer støtte planlægning og organisering, inklusiv fordeling af opgaver og vagter, men det kan også evaluere medarbejdere og ansøgere i forbindelse med ansættelse, afskedigelse og lønforhandling. På arbejdspladsen får de automatiserede beslutningssystemers vurderinger potentielt større effekt, end i mange andre sammenhænge, og risikoen for, at systemet begår fejl eller har bias, bliver tilsvarende mere relevant. For at kunne vurdere denne dataetiske udfordring er man, ligesom for privatliv, nødt til at forstå, hvorfor systemer kan begå fejl og have bias, men også hvordan man etisk kan forklare, at sådanne fejl og bias er problematiske.

I anden halvdel af denne rapport går vi i dybden med de dataetiske udfordringer, for at afdække hvad de består i, og hvilke dataetiske hensyn, som er på spil. **Kapitel fem** undersøger de dataetiske privatlivsudfordringer ved indsamling af medarbejderdata, ved blandt andet at kigge på:

---

3 Se Dataetisk Råd og Analyse & Tal. *En hverdag af data* (2023), https://www.ogtal.dk/assets/files/En-hverdag-af-data_compressed.pdf.

- Forskellige former for privatliv og ret til privatliv på arbejdspladsen
- Etiske begrundelser for medarbejderes ret til privatliv, samt
- Muligheden for etisk relevant samtykke til deling af data.

Kapitel seks undersøger på tilsvarende vis de dataetiske udforinger med risici for fejl og bias i automatiserede beslutningssystemer, ved at kigge på:

- Automatiserede beslutningssystemers effekt på medarbejdere, inklusive præstation, medarbejdertrivsel og gaming-effekter,
- Forskellige typer algoritmisk bias, og kilder til algoritmisk bias, samt
- Styrker og svagheder ved automatiseret beslutningsstøtte.

På baggrund af de fire kapitler, som præsenterer teknologien, de juridiske rammer, og relevante dataetiske hensyn, diskuterer vi i **kapitel syv**, hvordan man kan forstå de dataetiske udfordringer i tre realistiske cases.

Ved at fokusere på disse dataetiske udfordringer, er det forhåbningen, at denne rapport kan hjælpe med at afklare nogle af de væsentligste dataetiske spørgsmål, som knytter sig til arbejdspladsers anvendelse af de nye digitale værktøjer, men det er vigtigt at anerkende, at der findes dataetiske udfordringer, som denne rapport ikke behandler.

Rapporten diskuterer eksempelvis ikke den dataetiske udfordring med, at automatiserede beslutningssystemer kan være uigennemskueligt komplekse, og at det derfor kan være vanskeligt for både ledelse og medarbejdere, at forstå hvordan systemet er nået frem til en vurdering. Den diskuterer heller ikke den dataetiske udfordring med, at ansvaret for beslutninger kan blive diffust, når arbejdspladser anvender automatiserede beslutningssystemer, fordi det eksempelvis kan være vanskeligt at afgøre, om det er udvikleren eller lederen, som skal holdes ansvarlig, når systemet begår fejl. Både uigennemskuelighed og diffus ansvarlighed er vigtige dataetiske udfordringer for automatiserede beslutningssystemer, som fortjener øget opmærksomhed, men det ville føre for vidt, at inddrage dem i denne rapport. Endelig behandler rapporten ikke dataetiske udfordringer knyttet til bredere sociale og kulturelle strømninger, for eksempel hvordan udbredelsen af de digitale værktøjer kan tænkes at påvirke normer og relationer på arbejdsmarkedet, og om sådanne ændringer i givet fald er etisk ønskværdige eller problematiske. Selvom disse spørgsmål uden tvivl er vigtige, kan de være svære at behandle, fordi de afhænger af empiriske forhold, som det kan være overordentlig vanskeligt at vurdere.

2.3. Rapporten i det dataetiske landskab

De dataetiske udfordringer, som ny teknologi kan rejse, har været genstand for intenst fokus de seneste år. Der findes da også i dag en lang række dataetiske principper og retningslinjer

udviklet af interesseorganisationer, myndigheder, forskere og tænketanke.[4] Hvor placerer denne rapport sig i dette dataetiske landskab?

På et meget overordnet niveau kan man skelne mellem en stor gruppe retningslinjer og anbefalinger, som formulerer generelle dataetiske udfordringer og hensyn, og en mindre gruppe, som forsøger at give konkret vejledning til arbejdspladser, der ønsker at arbejde dataetisk med indsamling og anvendelse af medarbejderdata.

Udfordringerne med, at indsamling af data reducerer personers privatliv, og at automatiserede beslutningssystemer kan være fejl- og biasbehæftede, er således velkendte, og optræder i flere af de mest prominente dataetiske retningslinjer og anbefalinger. Ofte behandler de mere generelle dataetiske retningslinjer og anbefalinger imidlertid dataetik på et meget overordnet niveau. Det kan gøre det vanskeligt at forstå udfordringerne, og vurdere hvordan de skal tackles.

I retningslinjer fra EU-kommissionens Ekspertgruppe på højt niveau om kunstig intelligens hedder det eksempelvis, at:

> *"AI-systemer skal garantere privatlivets fred og databeskyttelse i hele livscyklussen for et system. Dette omfatter de oplysninger, brugeren indledningsvis har afgivet, og oplysninger, der er genereret om brugeren i løbet af vedkommendes interaktion med systemet (f.eks. output, som AI-systemet har genereret for specifikke brugere, eller hvordan brugere reagerede på bestemte anbefalinger)."* [5]

Tilsvarende fremhæver UNESCOs anbefalinger om etik for kunstig intelligens, at:

> *"Privacy, a right essential to the protection of human dignity, human autonomy and human agency, must be respected, protected and promoted throughout the life cycle of AI systems. It is important that data for AI systems be collected, used, shared, archived and deleted in ways that are consistent with international law and in line with the values and principles set forth in this Recommendation, while respecting relevant national, regional and international legal frameworks."* [6]

---

4 Anna Jobin, Marcello Ienca, and Effy Vayena, "The global landscape of AI ethics guidelines." *Nature Machine Intelligence* 1, no. 9 (2019), https://doi.org/10.1038/s42256-019-0088-2, https://doi.org/10.1038/s42256-019-0088-2; Brent Mittelstadt, "Principles alone cannot guarantee ethical AI" *Nature Machine Intelligence* 1, no. 11 (2019), https://doi.org/10.1038/s42256-019-0114-4, https://doi.org/10.1038/s42256-019-0114-4; Arif Ali Khan et al., "Ethics of AI: A Systematic Literature Review of Principles and Challenges". Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering (2022).

5 Den uafhængige ekspertgruppe på højt niveau om kunstig intelligens, *Etiske retningslinjer for pålidelig kunstig intelligens*, Europa-Kommissionen (2018), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60420.

6 UNESCO, *Recommendation on the Ethics of Artificial Intelligence* (2021), https://unesdoc.unesco.org/ark:/48223/pf0000381137/PDF/381137eng.pdf.multi.

Endelig stiller Global Union i sine ti principper for arbejderes databeskyttelse og privatliv krav om at arbejdsgiveren skal:

> *"Show respect for human dignity, [sic] privacy and the protection of personal data should be safeguarded in the processing of personal data for employment purposes, notably to allow for the free development of the employee's personality as well as for possibilities of individual and social relationships in the work place."* [7]

De tre tekster peger altså alle på privatliv som et centralt dataetisk hensyn. Tilsvarende peger ekspertgruppens retningslinjer på den dataetiske udfordring, at anvendelsen af AI medfører en risiko for bias og diskrimination:

> *"Datasæt, der anvendes af AI-systemer (til både indlæring og drift), kan være præget af medtagelsen af utilsigtet historisk skævhed, ufuldstændighed og dårlige styringsmodeller. Videreførelsen af sådanne skævheder kan føre til utilsigtet (in)direkte forudindtagethed og diskrimination over for bestemte grupper eller personer, og det kan potentielt forværre fordomme og marginalisering. […] Identificerbar og diskriminerende skævhed bør fjernes i indsamlingsfasen, hvis det er muligt. Den måde, hvorpå AI-systemer udvikles (f.eks. programmering af algoritmer), kan også være påvirket af uretfærdig skævhed. Dette kan modvirkes ved at indføre kontrolprocesser med henblik på at analysere og tilpasse systemets formål, begrænsninger, krav og beslutninger på en klar og gennemsigtig måde."* [8]

UNESCOs anbefalinger fordrer, at:

> *"AI actors should make all reasonable efforts to minimize and avoid reinforcing or perpetuating discriminatory or biased applications and outcomes throughout the life cycle of the AI system to ensure fairness of such systems. Effective remedy should be available against discrimination and biased algorithmic determination."* [9]

Og Global Unions femte princip for etisk kunstig intelligens stiller krav om at sikre en ikke-biasbehæftet kunstig intelligens:

> *"In the design and maintenance of AI, it is vital that the system is controlled for negative or harmful human-bias, and that any bias—be it gender, race,*

---

7 UNI Global Union, *Top 10 Principles for Workers' Data Privacy and Protection* (Nyon, 2017), https://uniglobalunion.org/wp-content/uploads/uni_workers_data_protection-1.pdf.

8 Den uafhængige ekspertgruppe på højt niveau om kunstig intelligens, *Etiske retningslinjer for pålidelig kunstig intelligens*.

9 UNESCO, *Recommendation on the Ethics of Artificial Intelligence*.

sexual orientation, age, etc.—is identified and is not propagated by the system." [10]

Som eksemplerne viser, er de dataetiske udfordringer, som vi fokuserer på i denne rapport, bredt anerkendt. Men hvad udfordringerne består i, og hvordan man etisk må forholde sig til dem, præciseres ikke yderligere i de tre retningslinjer og anbefalinger, som er citeret her. Det bliver derfor ikke klart, hvordan man skal forstå centrale begreber, som "privatliv" og "algoritmisk bias", ligesom det etiske grundlag for hensyn til at beskytte privatliv og undgå bias kun skitseres, ved eksempelvis at henvise til et ikke-nærmere defineret begreb om "respekt for menneskers rettigheder". Når dataetiske retningslinjer og anbefalinger bevæger sig på dette abstrakte niveau, bliver det vanskeligt både at vurdere om de er plausible, og at anvende dem, til at tage stilling til specifikke dataetiske udfordringer, herunder at afgøre, hvordan dataetiske hensyn adskiller sig fra de krav, som lovgivningen stiller.

I denne rapport går vi i detaljer med både de dataetiske udfordringer, og de etiske hensyn, som man må være opmærksom på, når man skal forholde sig til dem. Rapporten forsøger ikke at give detaljerede anvisninger til, hvordan arbejdspladser skal indsamle og anvende medarbejderdata, men at præsentere en præcis forståelse af hvad der er på spil, som kan udgøre fundamentet for en reflekteret dataetisk analyse af udfordringerne i en konkret kontekst.

Konkrete anvisninger findes til gengæld i den anden gruppe af praksisnære anbefalinger. I Danmark har flere fagforeninger udviklet retningslinjer til både tillidsrepræsentanter og ledere, ligesom en bred gruppe af arbejdsmarkedets parter i samarbejde med Algoritmer, Data og Demokrati-projektet i slutningen af 2023 præsenterede anbefalinger til ansvarlig og værdiskabende anvendelse af medarbejderdata.[11]

Denne type anbefalinger giver umiddelbart anvendelig vejledning for medarbejdere og ledere om, hvordan man kan introducere digitale værktøjer til indsamling og anvendelse af medarbejderdata på danske arbejdspladser, men har ikke til formål, at give en detaljeret forståelse af de dataetiske udfordringer og hensyn. Dataetisk Råds rapport placerer sig således imellem de mere generelle præsentationer af dataetiske udfordringer og principper, som findes i visse retningslinjer, og den praksisnære vejledning af medarbejdere og ledere, som findes i andre. Vores forhåbning er, at rapporten derved kan supplere og berige den eksisterende debat.

---

10 UNI Global Union, *Top 10 Principles for Workers' Data Privacy and Protection*.

11 Se for eksempel IDA, *Guide til ledere om overvågning og monitorering på arbejdspladser* (2023); IDA, *Guide til tillidsvalgte om overvågning og monitorering på arbejdspladser* (2023), https://ida.dk/media/13402/overvaagning-paa-arbejdspladsen-2023.pdf; Algoritmer Data og Demokrati-projektet, *Ansvarlig og værdiskabende anvendelse af medarbejderdata - anbefalinger til den digitaliserede arbejdsplads* (2023), https://taenketanken.mm.dk/wp-content/uploads/2023/11/Anbefalingskatalog_Ansvarlig-og-vaerdiskabende-anvendelse-af-medarbejderdata.pdf.

3. Digitale værktøjer til indsamling og anvendelse af medarbejderdata – teknologier og begreber

Hvis man har været ansat på en moderne arbejdsplads, så har man med stor sandsynlighed erfaring med, at arbejdspladsen indsamler og anvender medarbejderdata. En arbejdsplads kan eksempelvis gennemføre den lovpligtige arbejdspladsvurdering (APV) ved at medarbejderne anonymt udfylder spørgeskemaer, hvis resultater ledelsen efterfølgende kan analysere, dele og diskutere med medarbejderne. Denne og beslægtede former for indsamling og anvendelse af medarbejderdata er blevet udført på danske arbejdspladser i årtier, og i mange tilfælde er de bredt accepterede.

Det seneste årti er nye og ofte mere kontroversielle måder at indsamle og anvende medarbejderdata vundet frem. Der findes i dag et stort marked for digitale værktøjer, som dels kan gøre det lettere at indsamle medarbejderdata, og dels kan indsamle flere, mere præcise medarbejderdata, end arbejdspladser tidligere har kunnet. Der findes også digitale værktøjer, som kan analysere medarbejderdata, sammenholde dem med andre relevante data fra arbejdspladsen, og trække på statistiske sammenhænge mellem data for at foretage vurderinger, som kan støtte eller udgøre en beslutning.

Disse værktøjer kan på den ene side effektivisere arbejdet ved at erstatte eller kvalificere menneskelig arbejdskraft, ensarte beslutninger og synliggøre beslutningsgrundlaget. På den anden side kan de rejse dataetiske udfordringer, eksempelvis ved at reducere medarbejderes privatliv, og ved at træffe eller anbefale biasbehæftede beslutninger.

Dette kapitel præsenterer hvordan disse værktøjer er blevet udbredt internationalt, hvordan de er i færd med at blive udbredt på det danske arbejdsmarked, hvilke data de kan indsamle, og hvilke beslutninger de kan hjælpe med at træffe. Undervejs definerer vi en række af de centrale teknologiske begreber, som optræder i rapporten.

3.1. Internationale erfaringer med indsamling og anvendelse af medarbejderdata

Nogle af de mest kendte og diskuterede eksempler på arbejdspladser, som i vid udstrækning indsamler og anvender medarbejderdata, er store internationale virksomheder i den såkaldte platformsøkonomi. Platformsvirksomheder er karakteriseret ved, at de stiller en digital platform til rådighed for platformens "medarbejdere", der formelt er eller arbejder på vilkår som på visse måder minder om selvstændige erhvervsdrivende. Platformsmedarbejderne bruger platformen til at opnå forbindelse med kunder, for eksempel ved at platformen modtager tilbud, som den så fordeler til platformsmedarbejderne, der løser opgaven for kunderne. Platformens indtægt kommer fra honorarer for stille platformen til rådighed for kunder og medarbejdere, eksempelvis i form af et gebyr lagt på prisen for den enkelte opgave.

Et kendt eksempel på en platformsvirksomhed er Uber, som opererede i Danmark fra 2014 til 2017. Uber er en digital platform for chauffører og kunder, der på mange måder leverer samme service som et konventionelt taxi-selskab, men adskiller sig ved, at chauffører formelt er selvstændige, og ved at Uber indsamler store mængder data fra både chauffører og kunder. Disse data anvendes blandt andet til at justere priser og fordele opgaver. Selskabet var både i Danmark og andre europæiske lande kontroversielt, blandt andet fordi etablerede taxachauffører mente, at selskabet udførte taxakørsel uden at leve op til de almindelige krav for taxavirksomhed. I 2017 afsagde EU-domstolen dom i en sag, og slog fast at Uber reelt var et transportselskab, som derfor skulle leve op til de almindelige krav for lovlig taxakørsel.[12] Uber lukkede i 2017 både i Danmark og i flere andre europæiske lande„ fordi selskabet vurderede, at det ikke var muligt at leve op til disse krav.[13]

En beslægtet debat har efterfølgende rettet sig mod den finske budservice Wolt. Wolt er, ligesom Uber, et platformsselskab, som stiller en digital platform til rådighed, hvor formelt selvstændige "kurerpartnere" kan finde og acceptere ordrer på levering af mad til kunder. Wolt opererer i 2023 efter egne oplysninger i 25 lande, fortrinsvis i Europa. Ligesom Uber indsamler virksomheden medarbejderdata blandt andet med henblik på at fordele arbejdsopgaver til de tilknyttede bude. Og ligesom ved Uber har bude i dansk sammenhæng argumenteret for, at de reelt er i et ansættelseslignende forhold, og at Wolt derfor er forpligtet til at overholde almindelige krav til arbejdsgivere på det danske arbejdsmarked.[14] I juli 2023 fastslog Arbejdsmarkedets Erhvervssikring, at Wolt var forpligtet til at tegne arbejdsskadeforsikring for deres bude, fordi disse reelt er underlagt instruktionsbeføjelse, når de arbejder på en ordre via platformen.[15] Denne afgørelse følger ovenpå Skatterådets afgørelse fra 2022, hvori rådet fastslår, at Wolt må anses for arbejdsgiver og platformens bude for lønmodtagere.[16]

Den store opmærksomhed, som en mindre gruppe internationale platformsvirksomheder har tiltrukket sig i den offentlige debat, kunne måske forlede en til at tro, at indsamling og anvendelse af medarbejderdata er et fænomen, som er knyttet til et mindretal af særlige arbejdspladser. Det er imidlertid ikke tilfældet. Indsamling og anvendelse af medarbejderdata er vidt udbredt på mange forskellige typer arbejdspladser. EU-agenturet Eurofound udgiver med nogle års mellemrum en analyse af tendenser på det europæiske arbejdsmarked. Den seneste undersøgelse fra 2019 viste, at mere end hver femte europæiske arbejdsplads brugte

12 Se Asociación Profesional Elite Taxi mod Uber Systems Spain SL, No. C-434/15 (EU-domstolen 2017).

13 Se Mathias Sommer, "OVERBLIK: Uber får sparket i flere europæiske lande" *DR*, March 28 2017, https://www.dr.dk/nyheder/penge/overblik-uber-faar-sparket-i-flere-europaeiske-lande. Uber opererer i 2023 efter egne oplysninger fortsat i omkring 70 lande verden over, inklusive Sverige, Norge, Finland, Holland, Tyskland, Frankrig, Spanien, Italien og Storbritannien.

14 Henrik Moltke og Marcel Mirzaei-Fard, "Wolt-budet Laura vil have en overenskomst: 'Det burde ikke være anderledes, fordi vi arbejder for en app'" *DR*, June 11 2020, https://www.dr.dk/nyheder/penge/wolt-budet-laura-vil-have-en-overenskomst-det-burde-ikke-vaere-anderledes-fordi-vi

15 Asger Havstein Eriksen, "Wolt skal betale erstatning til alle bude, der kommer til skade" *Fagbladet 3F*, July 5 2023, https://fagbladet3f.dk/artikel/wolt-skal-betale-erstatning-til-alle-bude-der-kommer-til-skade.

16 Peter Christian, "Her er Skatterådets Wolt-afgørelse i fuld længde" *Radar*, January 27 2022, https://radar.dk/artikel/her-er-skatteraadets-wolt-afgoerelse-i-fuld-laengde.

digitale værktøjer til at indsamle og anvende medarbejderdata for at måle medarbejdernes præstation.[17] Og siden 2019 er udviklingen gået stærkt.

I 2019 medførte Corona-pandemien langvarige nedlukninger i mange lande. For nogle arbejdspladser betød det, at medarbejdere blev opsagt eller fik orlov, men især for medarbejdere med kontorarbejde rykkede arbejdspladsen i mange tilfælde ind i hjemmet. Situationen efterlod mange ledere med en svær opgave: hvordan leder man medarbejdere, når al fysisk kontakt og sparring bliver erstattet af Zoom-møder og mailkommunikation? På mange arbejdspladser blev øget indsamling og anvendelse af medarbejderdata en del af løsningen. Det gælder især i USA, hvor andelen af virksomheder, der benytter digitale systemer til at monitorere medarbejderne steg fra en tredjedel til to tredjedele på blot to år.[18]

Denne hastige udbredelse illustreres også af en international undersøgelse i kølvandet på Coronapandemien, som viste at 70% af de adspurgte virksomheder havde gennemført eller planlagde tiltag, som skulle indsamle data om medarbejderes præstation efter overgangen til hybridarbejde.[19] Disse tiltag inkluderede indsamling af data fra e-mails (44%) og browsere (41%), samarbejdsprogrammer (43%), og sågar videoovervågning (29%). Men de inkluderede også værktøjer til såkaldt 'attention tracking', der registrerer, hvis en medarbejder klikker væk fra et igangværende videomøde (28%), samt "keylogger"-software (26%), der registrerer al input fra tastaturet.

En anden relevant undersøgelse måler tendenser i den globale interesse for værktøjer til indsamling og anvendelse af medarbejderdata, som denne kommer til udtryk ved aktiviteter på internettet, for eksempel i form af mængden af relevante søgninger. Undersøgelsen viser at der skete en skarp stigning i forbindelse med de første nedlukninger under Coronapandemien. Aktiviteten lå i marts 2020 73% højere end gennemsnittet for månederne i 2019.[20] Nok så bemærkelsesværdigt er aktiviteten fortsat med at vokse, også i årene efter pandemien.

Der sker altså rigeligt på området i udlandet. Men hvordan ser det ud, hvis man vender blikket mod indsamling og anvendelse af medarbejderdata på arbejdspladser i Danmark?

---

17 Eurofound, *European Company Survey 2019 - Workplace Practices Unlocking Employee Potential* (2019), https://www.eurofound.europa.eu/publications/flagship-report/2020/european-company-survey-2019-workplace-practices-unlocking-employee-potential. Målinger af kvaliteten af medarbejderes arbejde kaldes i både faglitteraturen, den offentlige debat, og de digitale værktøjer, som her er på spil, flere forskellige ting. Udover måling af præstation, taler man også om måling af "performance" og måling af "produktivitet". Begreberne anvendes ofte synonymt, og der forekommer ikke at være konsensus om, hvordan de eventuelt varierer i betydning. For enkelhedens skyld benytter vi konsistent "måling af medarbejderes præstation" som samlet betegnelse, for denne type data og vurderinger.

18 Christopher Mims, "More Bosses Are Spying on Quiet Quitters. It Could Backfire" *The Wall Street Journal*, September 17 2022, https://www.wsj.com/articles/more-bosses-are-spying-on-quiet-quitters-it-could-backfire-11663387216.

19 VMware, *The New Remote Work Era: Trends in the Distributed Workforce* (2023), https://www.vmware.com/content/microsites/learn/en/655785_REG.html

20 Simon Migliano, *Employee Monitoring Software Demand Trends 2020-23*, Top10VPN (2023), https://www.top10vpn.com/research/covid-employee-surveillance/. Bemærk at analysen fortolker denne aktivitet som udtryk for efterspørgsel på disse teknologier. Den fortolkning har vi her valgt ikke at adoptere. Selvom det virker rimeligt at antage, at der findes en sammenhæng mellem den pågældende internetaktivitet på den ene side og efterspørgsel på (og anvendelse af) de digitale værktøjer på den anden side, så forekommer det også rimeligt at antage, at dette forhold kan være komplekst, for eksempel fordi andre faktorer end ren efterspørgsel kan påvirke den målte internetaktivitet.

3.2. Indsamling og anvendelse af medarbejderdata på danske arbejdspladser

Anvendelsen af ny teknologi formes af den sociale og kulturelle kontekst hvori den optræder. Man kan derfor rimeligvis spørge, hvordan teknologier til indsamling og anvendelse af medarbejderdata udbredes på et dansk arbejdsmarked, der i international sammenhæng er karakteriseret ved blandt andet høj tillid mellem ledelse og medarbejdere, høj faglig organisering, stor fleksibilitet, og et stærkt socialt sikkerhedsnet?

Zoomer man ind på den danske kontekst, så viser flere undersøgelser, at digitale værktøjer til indsamling og anvendelse af medarbejderdata også er blevet udbredt til danske arbejdspladser, ligesom undersøgelserne indikerer, at de kan rejse nogle af de samme udfordringer.

I slutningen af 2022 udarbejdede Dataetisk Råd, IDA og HK i samarbejde med Algoritmer, Data og Demokrati-projektet en undersøgelse med omtrent 1100 medarbejdere på danske arbejdspladser, som tilkendegav deres holdninger til og oplevelser med dataindsamling på arbejdspladsen.[21] Undersøgelsen viste, at godt to ud af tre medarbejdere oplever, at der indsamles digitale data om dem på arbejdspladsen, mens 15% er i tvivl, og kun 22% ikke har denne oplevelse.

Udbredelsen kan imidlertid variere for forskellige typer arbejdspladser. I Eurofounds internationale undersøgelse er digitalisering mest udbredt i finanssektoren, men langt mindre udbredt i eksempelvis bygge- og industrisektorerne.[22]

I dansk sammenhæng har DM akademikerforeningen foretaget en undersøgelse med svar fra 5.386 medarbejdere, som er organiseret af DM.[23] I denne gruppe, som må forventes fortrinsvis at have akademisk arbejde, oplever fire ud af fem, at der indsamles digitale medarbejderdata på arbejdspladsen.

Den undersøgelse, som Dataetisk Råd medvirkede til, viste også, at blandt de medarbejdere som oplevede, at arbejdspladsen indsamlede data, mente kun en ud af fire, at de havde talt med deres nærmeste leder om formålet med dataindsamlingen, og hver femte oplevede dataindsamlingen som en form for overvågning.[21] Endelig viste undersøgelsen, at de mest almindelige formål med dataindsamling ifølge medarbejdere var måling af trivsel og tilfredshed (38%), registrering af møde- og gåtider (25%), samt tidsforbrug på forskellige arbejdsopgaver (21%), mens kun mindre grupper oplevede, at arbejdspladsen indsamlede data om eksempelvis medarbejderes præstation (12%) og lokation (6%).

---

21 Grit Munk et al., *Danskernes holdninger til og oplevelser med indsamling af digitale medarbejderdata på arbejdspladsen*, Mandag Morgen, Algoritmer, Data og Demokratiprojektet, IDA – Ingeniørforeningen, HK Danmark, og Dataetisk Råd (2023), https://algoritmer.org/medarbejderdata/.

22 Eurofound, *European Company Survey 2019 - Workplace Practices Unlocking Employee Potential.*

23 Sofie Caroline Falkenberg Holm and Sofie Dalum, *Fire ud af fem får indsamlet medarbejderdata*, DM Akademikerforeningen (2023), https://dm.dk/media/1can022l/fire-ud-af-fem-faar-indsamlet-medarbejderdata.pdf.

I 2023 vendte endnu en undersøgelse fokus mod 600 lederes erfaringer med at bruge digitale værktøjer til at indsamle medarbejderdata på danske arbejdspladser.[24] I denne undersøgelse svarer fire ud af fem ledere, at de anvender medarbejderdata, som er indsamlet med digitale værktøjer, og to ud af tre, at de har talt med deres medarbejdere om, at der indsamles medarbejderdata. Den mest udbredte type medarbejderdata som indsamles er også i denne undersøgelse data om medarbejdertrivsel (56%), men i modsætning til medarbejderundersøgelsen er dette tæt efterfulgt af måling af medarbejderes præstation (46%).

Undersøgelserne viser at digitale værktøjer til indsamling og anvendelse af medarbejderdata allerede er vidt udbredte på danske arbejdspladser. De illustrerer også nogle af de udfordringer dette kan give, blandt andet ved at flere ledere angiver at benytte værktøjer (79%), end medarbejdere oplever at de benyttes (63%), ved at ledere i langt højere grad angiver at indsamle data om medarbejderes præstation (46% mod 12%), og ved at et stort flertal af ledere mener at have talt om dataindsamling med deres medarbejdere (68%), mens et mindretal af medarbejderne oplever, at have talt med en leder om dataindsamlingen.[25]

3.3. Hvilke medarbejderdata kan indsamles på arbejdspladsen?

Indsamling af medarbejderdata på arbejdspladsen er blevet udbredt med stor fart både internationalt og på danske arbejdspladser. Men hvilke medarbejderdata kan arbejdspladser indsamle og hvordan?[26]

---

24 Casper Waldemar Hald, Julie Karnøe Tranholm-Mikkelsen, and Katrine Lindtner Andersen, *Digital dataindsamling på arbejdspladsen - En undersøgelse af lederes holdninger til og oplevelser med indsamling af digitale medarbejderdata på arbejdspladsen*, Mandag Morgen, Algoritmer, Data og Demokrati-projektet, Dansk Erhverv, DI, Djøf, DM, FH, Finansforbundet, Forsikringsforbundet, og IDA (2023), https://taenketanken.mm.dk/wp-content/uploads/2023/09/Minirapport_Digital-dataindsamling-Lederanalyse08.pdf.

25 Tallene for samtaler om dataindsamling er ikke fuldt sammenlignelige, fordi medarbejdere er blevet spurgt om, hvorvidt en leder har talt med dem om *formålet* med at der indsamles medarbejderdata, mens ledere er blevet spurgt om, hvorvidt de har talt med medarbejdere om, at der *anvendes* digitale værktøjer til at indsamling af medarbejderdata. Det forekommer imidlertid tvivlsomt, at hele forskellen på de to tal kan forklares ved, at ledere og medarbejdere har talt om indsamling af data, men ikke om formål med indsamling af data. Forskellen i antallet af ledere som angiver at benytte værktøjer til at indsamle medarbejderdata, og antallet af medarbejdere, som oplever at der indsamles medarbejderdata, kan også i teorien skyldes tendenser i hvilke arbejdspladser, som indsamler data. Hvis det eksempelvis er mere almindeligt at indsamle medarbejderdata, på arbejdspladser, hvor ledere har små grupper af medarbejdere, og mindre almindeligt, på arbejdspladser, hvor ledere har store grupper af medarbejdere, så vil dette af sig selv føre til, at andelen af ledere som rapporterer at bruge værktøjer, er højere end andelen af medarbejdere, som rapporterer at de bruges. Igen må det dog anses for tvivlsomt, om hele forskellen kan forklares på denne måde.

26 Præsentationen af forskellige muligheder for at indsamle data i dette afsnit trækker på Alexandra Mateescu and Aiha Nguyen, *Workplace Monitoring & Surveillance*, Data & Society (2019), https://datasociety.net/wp-content/uploads/2019/02/DS_Workplace_Monitoring_Surveillance_Explainer.pdf; Trade Union Congress, *Technology Managing People* (London, 29 November 2020), https://www.tuc.org.uk/sites/default/files/2020-11/Technology_Managing_People_Report_2020_AW_Optimised.pdf; Valerio De Stefano, "'Negotiating the algorithm': Automation, artificial intelligence and labour Protection" *Comparative Labor Law & Policy Journal* 41, no. 1 (2019); AI Now Institute, *Algorithmic Management: Restraining Workplace Surveillance* (11 April 2023), https://ainowinstitute.org/publication/algorithmic-management; Sara Baiocco et al., *The Algorithmic management of work and its implications in different contexts*, International Labour Organisation & European Commission (2022), https://www.ilo.org/wcmsp5/groups/public/---ed_emp/documents/publication/wcms_849220.pdf.

Lokationsdata

En første type data, som arbejdspladsen kan indsamle, er **lokationsdata** om medarbejderens fysiske placering. En simpel variant af sådanne data er de data, som en arbejdsplads kan lagre fra digitale adgangssystemer. Hvis eksempelvis medarbejdere skal låse sig ind på arbejdspladsen med en unik kode eller et adgangskort, så kan adgangssystemet lagre disse data, som indikerer hvilke medarbejdere, der var fysisk til stede på arbejdspladsen, på hvilke tidspunkter. En arbejdsplads kan også indsamle mere detaljerede lokationsdata, ved eksempelvis at trække data fra en GPS som medarbejderen bærer eller har umiddelbart i nærheden, for eksempel på medarbejderens arbejdsmobiltelefon, eller fra arbejdspladsens køretøjer. Hvis arbejdspladsen indsamler mange lokationsdata, kan de give et meget detaljeret billede af ikke kun medarbejderens fysiske placering, men også af medarbejderens aktivitet i løbet af arbejdsdagen.

Billeddata

En anden type data er **billeddata** om medarbejderens udseende og adfærd. På nogle arbejdspladser kan sådanne data trækkes fra overvågningskameraer, som er placeret på arbejdspladsen. I andre tilfælde kan arbejdspladsen trække billeddata fra medarbejderens mobiltelefon eller arbejdscomputer. Billeddata kan efterfølgende behandles og analyseres på flere forskellige måder, for at udlede information om medarbejderen. Billeder kan eksempelvis analyseres for at kontrollere medarbejderes identitet, for løbende at vurdere aktivitet, for eksempel udførelse af arbejdsopgaver eller opmærksomhed under møder, og for at evaluere medarbejderens følelsesmæssige tilstand.

Biometriske data

En tredje type data, er **biometriske data** om medarbejderens fysiske tilstand. I enkelte brancher har man konventionelt indsamlet sådanne data i forbindelse med helbredstjek.* Det seneste årti er det imidlertid blevet stadig mere almindeligt, at arbejdspladser benytter digitale værktøjer placeret i eksempelvis et smart-watch til løbende at indsamle biometriske data. Det kan eksempelvis være kropstemperatur, puls, antal skridt på en dag, og søvnrytme. Hvis disse data indsamles af eller deles med arbejdspladsen, kan den bruge de indsamlede data til at evaluere eksempelvis medarbejderes stress-niveau, balance mellem arbejde og fritid, vaner af relevans for sundhed, og overordnet fysisk helbred.

* Et almindeligt eksempel er piloter, som skal gennemgå et helbredstjek med henblik på helbredsgodkendelse for at få autorisation til kommerciel flyvning.

Systemdata

En fjerde type data er **systemdata** om medarbejderens aktivitet i arbejdspladsens systemer. Eksempelvis bruger mange arbejdspladser digitale systemer til at registrere og organisere arbejdsopgaver, og vil typisk i den forbindelse gemme data om hvilke medarbejdere, der er knyttet til opgaven, hvornår opgaven startes, hvornår forskellige skridt tages, og hvornår opgaven løses. Et eksempel på denne type dataindsamling kunne være et lager, hvor medarbejdere registrerer hvilke varer de henholdsvis placerer og henter på lageret. Arbejdspladsen kan også logge information om, hvilke af arbejdspladsens systemer som anvendes af hvilke medarbejdere på forskellige tidspunkter, og hvilken information de tilgår. Sådanne data er oplagt relevante, når medarbejdere har adgang til potentielt følsomme data gennem arbejdspladsens systemer. Medarbejdere kan også have digitale arbejdskalendere, e-mailkonti, og arbejdstelefoner, som gemmer digitale data om medarbejderens aktiviteter. Arbejdspladsen vil eksempelvis kunne samle data om hvilke aftaler og møder medarbejderen har, hvem medarbejderen mødes med og hvad dagsordenen for mødet er. Tilsvarende kan arbejdspladsen gemme metadata om hvor mange mails og beskeder medarbejderen skriver, hvem medarbejderen kommunikerer med, og hvor lange e-mails og beskeder er, og om hvilke telefonopkald medarbejderen henholdsvis modtager og foretager. Men arbejdspladsen kan også tilgå indholdet i sådan kommunikation, ved at analysere tekst i e-mails og beskeder, og ved at optage telefonsamtaler. I job, hvor medarbejdere bruger meget af tiden på at arbejde i sådanne systemer, får arbejdspladsen mulighed for at indsamle store mængder af meget detaljerede data om medarbejderens aktiviteter.

Aktivitetsdata

I tillæg til alle disse data kan arbejdspladsen anvende specialiserede digitale værktøjer til at trække udvalgte data, typisk endnu mere detaljerede **aktivitetsdata**, fra for eksempel medarbejderens arbejdscomputer eller telefon. Arbejdspladsen kan eksempelvis indsamle metadata om hvilke hjemmesider medarbejderen besøger via sin browser og hvilke filer medarbejderen downloader. Men digitale værktøjer kan også lave langt mere detaljeret registrering af eksempelvis tastetryk og musebevægelser, herunder antallet af pauser, eller screenshots af medarbejderens skærm på udvalgte eller tilfældige tidspunkter i løbet af arbejdsdagen.

3.4. Hvordan anvendes automatiserede beslutningssystemer på arbejdspladsen?

Arbejdspladser kan bruge indsamlede medarbejderdata på mange måder. Oplagt kan ledelsen på arbejdspladsen kigge på de indsamlede data, og bruge dem, når de finder det relevant. I stigende grad benytter arbejdspladser imidlertid også automatiserede beslutningssystemer til at behandle medarbejderdata. Det skyldes en kombination af, at det kan være en uoverkommelig opgave for mennesker at gennemgå de store mængder data, og at disse systemer ofte kan behandle data på måder, som mennesker kan have svært ved, for eksempel ved at trække på komplekse statistiske sammenhænge mellem mange forskellige typer data.

I dette afsnit kigger vi på, hvad automatiserede beslutningssystemer er, hvordan de udvikles, og hvilke typer beslutninger de kan bruges til at støtte eller træffe.

3.4.1. Hvad er et automatiseret beslutningssystem?

Et automatiseret beslutningssystem er software, som ved at kombinere data med en statistisk model for sammenhænge i disse data, foretager en vurdering, der kan støtte eller udgøre en beslutning.[27] De digitale værktøjer, som vi kigger på i denne rapport, er automatiserede beslutningssystemer som behandler medarbejderdata, eventuelt i kombination med andre data, og foretager en vurdering, som kan støtte eller udgøre en beslutning på arbejdspladsen.

Automatiseret beslutningssystem:

Software som bruger data og en statistisk model til at foretage en vurdering, som kan støtte eller udgøre en beslutning.

Hvad betyder det, at systemet foretager en vurdering? Hvordan bruger systemet medarbejderdata? Og hvad vil det sige, at systemet har en statistisk model? To enkle, fiktive eksempler kan illustrere, hvordan et automatiseret beslutningssystem på en arbejdsplads kan virke.

Den første arbejdsplads har udviklet et automatiseret beslutningssystem til at hjælpe med at beskytte arbejdspladsen mod sikkerhedsbrister. Konkret forsøger systemet at identificere risici for, at medarbejdere lækker følsomme data.

---

27 Automatiserede beslutningssystemer beskrives ofte som en variant af kunstig intelligens. Kunstig intelligens kan imidlertid forstås på forskellige måder, og det afhænger af definitionen om automatiserede beslutningssystemer er eller ikke er en form for kunstig intelligens. Selmer Bringsjord and Naveen Sundar Govindarajulu, "Artificial Intelligence" in *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta (2018). https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=artificial-intelligence. Samtidig er kunstig intelligens meget andet end automatiserede beslutningssystemer. I den offentlige debat diskuteres automatiserede beslutningssystemer også ofte under betegnelsen "algoritme". En algoritme er et sæt af instrukser for at udføre en serie af logiske operationer. Robin K. Hill, "What an Algorithm Is" *Journal of Philosophy & Technology* 29, no. 1 (2016), https://doi.org/10.1007/s13347-014-0184-5, https://doi.org/10.1007/s13347-014-0184-5. Automatiserede beslutningssystemer er derfor en form for algoritme. Men ligesom kunstig intelligens findes algoritmer mange andre steder, end i automatiserede beslutningssystemer. For at fokusere på det relevante fænomen, benytter vi derfor i denne rapport betegnelsen automatiseret beslutningssystem, snarere end kunstig intelligens eller algoritme.

Systemet samler data fra mange af arbejdspladsens andre systemer om hver enkelt medarbejders aktivitet. Hver medarbejder har en unik profil, som opdateres dagligt, der indeholder den seneste måneds data. De relevante data blev identificeret, da systemet blev trænet med maskinlæring. Denne træning viste, at nogle typer data er vigtige at holde øje med, mens andre ikke er vigtige.

Systemet indeholder en statistisk model, som viser hvordan forskellige slags medarbejdere typisk arbejder i arbejdspladsens systemer. Modellen kan for eksempel vise hvor ofte en medarbejder med en bestemt stilling gennemsnitligt tilgår følsomme data, eller hvor meget datatrafik en medarbejder har ud af arbejdspladsen. Modellen giver afvigelser fra normal aktivitet en score. Små afvigelser giver en lav score, mens store afvigelser giver en høj score, men nogle data vejer tungere end andre. Selv små afvigelser fra den almindelige aktivitet i tilgang til følsomme data kan give en høj score. Hver medarbejder får en samlet score ved at lægge scoren for de enkelte datapunkter sammen.

Modellen har også en tærskel, hvor den samlede score bliver så høj, at der statistisk set er en relevant risiko for en sikkerhedsbrist. Hvis scoren for en medarbejder overskrider denne tærskel, så slår systemet alarm, ved at sende en notifikation til medarbejderen, medarbejderens nærmeste leder og arbejdspladsens sikkerhedsspecialister. Ledelse og sikkerhedsspecialister kigger i fællesskab på sagen, og kan i de fleste tilfælde hurtigt afblæse advarslen som en falsk alarm. I enkelte tilfælde træffer de beslutning om, at der er grund til at undersøge, om der er sket en sikkerhedsbrist.

Det automatiserede beslutningssystem vurderer altså, om medarbejderes aktiviteter giver grund til at tro, at der er risiko for en sikkerhedsbrist. Systemet bruger medarbejderes individuelle data, ved at sammenligne dem med data om, hvordan almindelig aktivitet ser ud. Systemets model er en måde at beregne, hvordan usædvanlig aktivitet statistisk hænger sammen med risiko for en sikkerhedsbrist.

Den anden arbejdsplads har udviklet et automatiseret beslutningssystem til prioritering af ansøgere. Konkret forsøger systemet at vurdere, hvor god hver enkelt ansøger vil være, hvis vedkommende bliver ansat.

Systemet scanner ansøgeres CV og ansøgning, og bruger kunstig intelligens med en sprogmodel til at fortolke disse tekster, og opsummere hver ansøgers relevante data, for eksempel uddannelse, anciennitet, relevant erfaring, og match mellem ansøgerens kompetencer, og de kompetencer, som stillingen kræver.

Systemet har en model, som viser hvordan forskellige data statistisk hænger sammen med, hvor godt en ansøger gennemsnitligt præsterer, hvis vedkommende bliver ansat. Denne model er trænet ved at kigge på, hvordan tidligere ansøgere har klaret sig, når de blev ansat. Nogle data spiller en stor rolle – ansøgere med bestemte kompetencer, har historisk klaret sig virkelig godt – mens andre data spiller en mindre rolle – uddannelse er relevant, fordi ansøgere med nogle uddannelser klarer sig bedre end andre ansøgere, men uddannelse har vist sig, at være langt mindre vigtigt, end man troede.

Systemets model har en matematisk funktion, som beregner hvordan de mange forskellige data tilsammen stiller ansøgere. Resultatet udtrykkes som en score fra 1-10, der angiver hvilken gruppe ansøgeren tilhører, fra de 10% svageste ansøgere (1), over de 10-20% svageste

ansøgere (2), og helt op til de 10% stærkeste ansøgere (10). Ledelsen anvender denne vurdering, når de beslutter hvilke ansøgere, som skal kaldes til samtale.

Det automatiserede beslutningssystem vurderer altså hver ansøgers kvalifikationer. Systemet bruger ansøgeres individuelle data fra CV og ansøgning i en model, baseret på historiske data om, hvad der karakteriserer succesfulde medarbejdere. Systemets model lader det vurdere, hvordan hver ansøgere er placeret i forhold til andre ansøgere.

3.4.2. Beslutningsstøtte og fuld automatisering

Et automatiseret beslutningssystem foretager en vurdering, baseret på et input af data og en model for statistiske sammenhænge i disse data. Det karakteristiske ved automatiserede beslutningssystemer er, at systemet er designet til at foretage en vurdering, som er relevant for en beslutning. Hvis systemet eksempelvis peger på mulige sikkerhedsbrister, så kan arbejdspladsen undersøge og lukke dem. Hvis det prioriterer ansøgere, så kan arbejdspladsen invitere de bedst kvalificerede til samtale.

En arbejdsplads kan anvende vurderingen fra et automatiseret beslutningssystem på to forskellige måder. Den ene mulighed er, at vurderingen informerer en menneskelig beslutning. I dette tilfælde har systemets vurdering ingen selvstændig effekt på det forhold, som vurderes, eksempelvis lukning af sikkerhedsbrister, eller prioritering af ansøgere. Vurderingen har alene effekt på beslutninger ved, at en menneskelig beslutningstager kan inddrage den i de overvejelser, som fører til en beslutning. Denne anvendelse af automatiserede beslutningssystemer kaldes ofte for automatiseret beslutningsstøtte (se også afsnit 6.6 om et-menneske-i-kredsløbet).

Automatiseret beslutningsstøtte:

Det automatiserede beslutningssystems vurdering indgår i beslutningsgrundlaget for en menneskelig beslutning på arbejdspladsen.

Den anden mulighed er, at arbejdspladsen lader vurderingen udgøre en beslutning. I dette tilfælde påvirker systemets vurdering direkte hvordan arbejdspladsen handler i det forhold, som systemet vurderer: Identificerede sikkerhedsbrister lukkes og de højest prioriterede ansøgere inviteres til samtale. Menneskelige beslutningstagere kan i dette tilfælde have mulighed for at observere og ændre beslutningen, men menneskelig inddragelse er ikke nødvendig for, at systemets vurdering har effekt som beslutning. Denne anvendelse af automatiserede beslutningssystemer kaldes ofte for fuldt automatiserede beslutninger.

Fuldt automatiserede beslutninger:

Det automatiserede beslutningssystems vurdering udgør eller udføres direkte som en beslutning på arbejdspladsen.

Det kan være vanskeligt at trække en skarp grænse mellem beslutningsstøtte og fuldt automatiserede beslutninger af mindst to grunde. For det første er det ofte uklart, hvad der udgør

den relevante beslutning. Er det eksempelvis en relevant beslutning, at undersøge mulige sikkerhedsbrist, eller er det først i relevant forstand en beslutning, hvis man konkluderer, at en hændelse udgør en sikkerhedsbrist? For det andet kan det være svært at sige, hvornår en menneskelig beslutningstager er tilstrækkeligt involveret til, at der er tale om beslutningsstøtte snarere end en fuldt automatiseret beslutning. Er det eksempelvis tilstrækkeligt, at en menneskelig beslutningstager formelt skal godkende en beslutning, som i udgangspunktet er baseret på systemets vurdering? Eller forudsætter det, at mennesker aktivt kontrollerer systemets vurdering, og tager selvstændigt stilling? Hvor grundig skal denne kontrol i givet fald være? Som det nok fremgår, kan der være gråzoner, hvor det er uklart, om vi bedst kan kategorisere anvendelsen af et automatiseret beslutningssystem som beslutningsstøtte eller fuld automatisering. Ikke desto mindre er de to begreber nyttige idealtyper for forskellige måder, at anvende automatiserede beslutningssystemer på arbejdspladsen.

3.4.3. Udvikling af automatiserede beslutningssystemer

Et automatiseret beslutningssystem på arbejdspladsen er software, som anvender relevante data, herunder medarbejderdata, til at foretage en vurdering, som kan udgøre eller informere en beslutning. Hvordan udvikler man sådanne systemer?

Der findes i dag to fundamentalt forskellige måder, at udvikle et automatiseret beslutningssystem. Indtil for knap ti år siden var det almindeligt at udvikle systemer ved, at mennesker i detaljer designede systemet. Dette design kunne afspejle både ekspertvurderinger og analyser af data. En udvikler, som eksempelvis ønskede at lave et system til prioritering af ansøgere i forbindelse med ansættelse, kunne konsultere arbejdspladsens HR-afdeling, for at høre hvilke erfaringer de havde. HR-medarbejderne kunne pege på de træk, som de vurderede var karakteristiske for ansøgere, som ved ansættelse var blevet de mest succesfulde medarbejdere. Udvikleren kunne også analysere data fra arbejdspladsen, for at finde statistiske sammenhænge mellem udvalgte træk ved tidligere ansøgere på den ene side, og success-kriterier for medarbejdere på den anden side, for eksempel evalueringer af medarbejderes præstationer. En sådan analyse kunne eksempelvis nå frem til, at ansøgere med visse uddannelsesprofiler, eller et bestemt forudgående karriereforløb, typisk viste sig at være de mest succesfulde medarbejdere. Udvikleren kunne på den baggrund designe et system som prioriterede ansøgere, ved at score ansøgere som indfriede disse kriterier højt, og ansøgere som ikke eller i mindre grad indfriede kriterierne lavt.

Langt de fleste automatiserede beslutningssystemer udvikles i dag med en anden metode, såkaldt "maskinlæring". Når en udvikler anvender maskinlæring, så designer en menneskelig udvikler ikke systemet i detaljer. I stedet indstiller udvikleren en læringsalgoritme, og fodrer denne læringsalgoritme med historiske data. Disse data indeholder eksempler, på den type vurdering, som systemet skal foretage – jo flere, jo bedre. Derpå "træner" læringsalgoritmen systemet. Konkret justerer

Maskinlæring:

En metode til at udvikle et automatiseret beslutningssystem, hvor en læringsalgoritme analyserer historiske data, og træner systemets statistiske model til at repræsentere sammenhænge i disse data.

læringsalgoritmen den matematiske model, som er kernen i et beslutningssystem, indtil modellen er optimeret. At modellen er optimeret betyder, at den repræsenterer de statistiske sammenhænge i træningsdata på den måde, som samlet set fører til de bedste vurderinger. Hvad der tæller som de samlet set bedste vurderinger bestemmes af læringsalgoritmens succeskriterier, ofte i form af en såkaldt "tabsfunktion".

Systemer udviklet med maskinlæring[28] kan fungere på samme måde som et system udviklet af mennesker, men er ofte betragteligt mere komplekse og præcise. Selvom det ved maskinlæring er en læringsalgoritme, som træner systemet, så er det også værd at holde sig for øje, at menneskelige udviklere træffer afgørende valg i processen. Udvikleren vælger hvilken læringsalgoritme som skal anvendes, hvilke data læringsalgoritmen skal træne på, og hvad der skal tælle som succeskriterier for det færdige system. Mange af de dataetiske udfordringer, som vi diskuterer i denne rapport, kan optræde uanset om systemet udvikles på den ene måde eller den anden måde. For overskuelighedens skyld fokuserer vi imidlertid på systemer udviklet med maskinlæring.

3.4.4. Hvilke typer beslutninger kan automatiseres på arbejdspladsen?

Hvilke beslutninger kan en arbejdsplads anvende automatiserede beslutningssystemer til at støtte eller udføre?

Udvikling af automatiserede beslutningssystemer er i disse år i en rivende udvikling, hvor det ofte kan føles som om, at det kun er fantasien, som sætter grænser for, hvilke beslutninger systemerne kan støtte eller træffe. I praksis er udviklingen af automatiserede beslutningssystemer typisk begrænset af, hvilke data udvikleren har til rådighed – man kan ikke udvikle et pålideligt automatiseret beslutningssystem, hvis man ikke har adgang til store mængder relevante data af høj kvalitet. Det betyder, at man kun kan udvikle automatiserede beslutningssystemer, for de beslutninger, hvor man har gode data om beslutninger, og de forhold som påvirker resultatet.

Hvor meget data, der er til rådighed om beslutningerne på en arbejdsplads, kan forventes at variere betydeligt fra branche til branche og fra beslutning til beslutning. Det betyder, at det er vanskeligt på et helt generelt niveau at sige, hvilke beslutninger, som kan støttes eller træffes af automatiserede beslutningssystemer på arbejdspladser. Omvendt er det i princippet muligt at skabe datasæt med de relevante data for en næsten hvilken som helst beslutning, og derfor principielt muligt, at udvikle automatiserede beslutningssystemer for nærmest alle typer beslutninger.

Når man kigger på de digitale værktøjer, som findes på markedet, giver de da også mulighed for at anvende automatiserede beslutningssystemer til en bred vifte af forskellige typer beslutninger. Den amerikanske NGO Coworker har udarbejdet en database med mere end

28 Maskinlæring anvendes også til at træne andre typer systemer, men vi fokuserer i denne sammenhæng på maskinlæring til udvikling af automatiserede beslutningssystemer.

550 digitale værktøjer, til at indsamle og anvende medarbejderdata.[29] De mange forskellige digitale værktøjer kan indsamle vidt forskellige data, og kan have automatiserede beslutningssystemer til vidt forskellige typer beslutninger. Værktøjer kan registrere arbejdstid (for eksempel Upwork), tastetryk (for eksempel Interguard), skærmbilleder (for eksempel Forcepoint), aktivitet i andre softwaresystemer, som Office-pakken eller browseren (for eksempel Celonis), metadata om møder, e-mails og telefoni (for eksempel Microsoft Viva), og lokationsdata (for eksempel Comfy).[30] Systemerne kan tilsvarende foretage vurderinger af medarbejderes præstation (f.eks. Upwork), arbejdsprocesser (for eksempel Celonis), sikkerhedsrisici (for eksempel Interguard), trivsel på aggregeret niveau (Microsoft Viva), motivation (for eksempel Forcepoint), og trivsel på individuelt niveau (for eksempel Comfy).[31]

Fordi de digitale værktøjer indsamler forskellige typer og mængder data, og anvender forskellige automatiserede beslutningssystemer til forskellige typer beslutninger, kan de møde meget forskellige dataetiske udfordringer. Ét værktøj kan vise sig at rejse meget alvorlige etiske udfordringer, mens et andet kan være etisk uproblematisk. De dataetiske udfordringer må vurderes for det konkrete værktøj, som en arbejdsplads anvender.

---

29 Wilneida Negrón, *Little Tech is coming for workers*, CoWorker (2021), https://home.coworker.org/wp-content/uploads/2021/11/Little-Tech-Is-Coming-for-Workers.pdf.

30 Se Baiocco et al., *The Algorithmic management of work and its implications in different contexts*; "Celonis" https://www.celonis.com/; InterGuard, "Employee Productivity Tracking Software" https://www.interguardsoftware.com/employee-productivity-tracking/; Stine Lomborg, "Everyday AI at Work - Self-tracking and automated communication for smart work" in *Everyday Automation*, ed. Sarah Pink et al. (Routledge, 2022); "Forcepoint" https://www.forcepoint.com/; "Comfy App" https://comfyapp.com/.

31 Der findes, så vidt vi har kunnet konstatere, ikke undersøgelser af de mange forskellige værktøjers udbredelse på det danske arbejdsmarked. De tidligere citerede medarbejder- og lederundersøgelser giver et vist indblik i hvilke data som indsamles og til hvilke formål, men der findes ikke et overblik over hvilke specifikke værktøjer som optræder på danske arbejdspladser.

4. Juridiske rammer

Vi har i det foregående kapitel præsenteret nogle af de forskellige muligheder, som moderne teknologi giver for indsamling af medarbejderdata og anvendelse af automatiserede beslutningssystemer på arbejdspladsen. Både indsamling og anvendelse af medarbejderdata er underlagt juridiske rammer. På det danske arbejdsmarked er de især reguleret af Den Europæiske Menneskerettighedskonvention (EMRK), Databeskyttelsesforordningen (GDPR), Databeskyttelsesloven, Forskelsbehandlingsloven, Arbejdsmiljøloven, og Aftalen om kontrolforanstaltninger mellem arbejdsmarkedets parter.[32] I dette afsnit skitserer vi udvalgte forpligtelser og rettigheder i disse centrale reguleringer, og opsummerer de væsentligste juridiske rammer, som de tilsammen definerer. Efter at have skitseret rammerne kigger vi kort på hvordan lovgivningen kan udvikle sig i de kommende år. Det er vigtigt at understrege, at formålet ikke er, at give en detaljeret og udtømmende redegørelse for, hvordan indsamling og anvendelse af medarbejderdata er reguleret, men derimod at præsentere de juridiske rammer i bred forstand som baggrund for den dataetiske analyse.[33] Afslutningsvis diskuterer vi forskellen på dataetik og jura, og den rolle, som de juridiske rammer spiller for dataetisk analyse. Udgangspunktet for Dataetisk Råds arbejde i denne rapport er at arbejdspladser, som indsamler og anvender medarbejderdata, overholder gældende lov. Arbejdspladser, som ønsker at bruge de nye digitale værktøjer, er nødt til at kende de grænser, som de juridiske rammer sætter for indsamling og anvendelse af medarbejderdata. Spørgsmålet er hvordan dataetiske overvejelser kan bidrage i forlængelse af de juridiske rammer. Vi peger i den forbindelse på, at dataetik kan være relevant i forbindelse med udvikling af regulering og aftaler, og i situationer, hvor en arbejdsplads lovligt kan handle på måder, som strider mod dataetiske hensyn.

4.1. Centrale love og aftaler

4.1.1. Den Europæiske Menneskerettighedskonvention

Den Europæiske Menneskerettighedskonvention (EMRK) blev vedtaget af Europarådet i 1950 med det formål, at beskytte borgernes grundlæggende rettigheder.[34] Den centrale rettighed i forbindelse med indsamling af medarbejderdata og anvendelse af automatiserede beslutningssystemer er retten til privatliv. EMRK fastslår således:

---

32 I tillæg til de ovennævnte retskilder kan også blandt andet Straffeloven, Offentlighedsloven, Forvaltningsloven og Helbredsoplysningsloven på forskellig vis have betydning. Vi fokuserer i denne sammenhæng på et mindre udsnit af centrale reguleringer.

33 Opgaven med at producere detaljerede og udtømmende redegørelser for de juridiske rammer varetages i Danmark af blandt Datatilsynet. Se vejledninger om "Databeskyttelse i ansættelsesforhold" (2023), https://www.datatilsynet.dk/Media/0/8/Vejledning%20om%20databeskyttelse%20i%20forbindelse%20med%20ans%c3%a6ttelsesforhold.pdf, samt "Kontrol af medarbejdere" (2023), https://www.datatilsynet.dk/Media/638348919997326341/Kontrol%20af%20medarbejdere.pdf.

34 https://www.retsinformation.dk/eli/lta/1996/423

- at enhver har ret til respekt for sit privatliv og familieliv, sit hjem og sin korrespondance (Artikel 8, stk.1).

Konventionen giver visse muligheder for indgreb i de konventionsbestemte rettigheder, men kun når sådanne indgreb er lovbestemte og nødvendige. Indgreb i privatlivet kan eksempelvis legitimeres med henvisning til, at de er nødvendige for at forhindre kriminalitet, beskytte borgeres sundhed eller sikre andre af konventionens rettigheder.

Artikel 8's betydning for indsamling af medarbejderdata er blandt andet behandlet i sagen Copland v. UK fra 2007.[35] Skolelæreren Lynette Copland blev først suspenderet og siden afskediget, efter at have brugt skolens computer og internet til private formål. Arbejdspladsen havde installeret et system, der logførte ansattes internetbrug, herunder hjemmesidebesøg og elektronisk kommunikation. Systemet blev implementeret af skolen med det formål at sikre, at de ansatte brugte skolens ressourcer og tid i overensstemmelse med skolens regler og politikker. Copland hævdede i søgsmålet, at den dataindsamling, som afslørede hendes brug af skolens computer til private formål, krænkede hendes ret til privatliv i henhold til Artikel 8. Den Europæiske Menneskerettighedsdomstol (EMD) afgjorde sagen til fordel for Copland. Domstolen vurderede at dataindsamlingen var ulovlig, fordi i) hverken arbejdspladsen eller staten officielt havde autoriseret denne type dataindsamling, og ii) det forhold at dataindsamling hverken var lovliggjort eller var blevet oplyst skabte en rimelig forventning om privatliv, selvom kommunikationen foregik via arbejdspladsens telefon og computer. Afgørelsen understregede således, at lovlig indsamling af medarbejderdata forudsætter, at der findes klare og eksplicitte regler for denne dataindsamling, samt at medarbejdere informeres om, hvordan og hvornår deres data kan blive indsamlet.

4.1.2. Den Europæiske Databeskyttelsesforordning og Databeskyttelsesloven

Den vigtigste regulering af indsamling og anvendelse af persondata findes i EU's Databeskyttelsesforordning (GDPR), som blev vedtaget i 2016.[36] GDPR gælder direkte og umiddelbart i EU's medlemsstater, men gennemføres og suppleres i Danmark af Databeskyttelsesloven.[37] Forordningen og loven fastsætter sammen en række klare grænser for, hvordan en arbejdsplads kan "behandle" medarbejderdata, herunder indsamle data og anvende data i et automatiseret beslutningssystem. Disse inkluderer:

- At arbejdspladsen kun må behandle data, når det sker på et af de grundlag, som forordningen definerer, herunder i) når medarbejderen samtykker til databehandlingen, ii) når det er nødvendigt for at indfri arbejdspladsens lovmæssige eller kontraktlige

35 Copland v. the United Kingdom, No. 62617/00 (Den Europæiske Menneskerettighedsdomstol 2007).

36 https://gdpr-info.eu/

37 https://www.retsinformation.dk/eli/lta/2018/502

forpligtelser, eller iii) når det er nødvendigt for arbejdspladsens legitime interesser (GDPR art. 6, stk. 1).[38]

- At arbejdspladsen kun må indsamle data, til udtrykkeligt angivne og legitime formål, samt at indsamlede data ikke må behandles til andre formål, der er uforenelige med det formål, som de oprindeligt blev indsamlet til ("formålsbegrænsning", GDPR art. 5, stk. 1, litra b, se også art. 6, stk. 4).
- At arbejdspladsen kun må behandle data i det omfang, som er nødvendigt for at indfri formålet med databehandlingen, herunder at arbejdspladsen ikke må indsamle mere data end nødvendigt, eller gemme data, når det ikke længere skal anvendes ("dataminimering" og "opbevaringsbegrænsning", GDPR art. 5, stk. 1, litra c og e).
- At arbejdspladsens legitime interesse i at behandle data kan være et ugyldigt grundlag for databehandling, når databehandling strider mod medarbejdernes interesser og rettigheder, og hensynet til disse går forud for hensynet til arbejdspladsens interesse (GDPR art. 6, stk. 1, litra f).
- At selv med et ellers gyldigt grundlag må arbejdspladsen ikke, med visse undtagelser, behandle særligt følsomme persondata, som data om medarbejderes etnicitet, politiske og religiøse overbevisninger, tilknytning til faglige organisationer, biometriske data, helbredsdata og data om seksuel orientering (GDPR art. 9, stk. 1). Den mest relevante undtagelse er muligheden for, at medarbejdere kan give udtrykkeligt samtykke til behandling af sådanne data (GDPR art. 9, stk. 2, litra a).
- At arbejdspladsen skal oplyse medarbejdere om dataindsamlingen, når den indsamler medarbejderdata, herunder hvilke(t) formål data indsamles til (GDPR art. 13, stk. 1, se også betragtning 60).
- At arbejdspladsen ikke må anvende automatiserede beslutningssystemer til at træffe fuldt automatiserede beslutninger som i væsentlig grad påvirker medarbejdere, med mindre dette specifikt lovliggøres på nationalt niveau, eller medarbejderen samtykker (GDPR art. 22, se også betragtning 71).
- At arbejdspladsen skal oplyse medarbejdere om fuldt automatiserede beslutningssystemer, som anvendes i forbindelse med beslutninger der i væsentlig grad påvirker medarbejdere, herunder meningsfuld information om systemets logik, formålet med og konsekvenserne af sådanne systemer (GDPR art. 13, stk. 2, litra f, se også betragtning 60).
- At fuldt automatiserede beslutninger på arbejdspladsen, som i væsentlig grad påvirker medarbejdere, kun i visse tilfælde må anvende særligt følsomme persondata om medarbejderes etnicitet, politiske og religiøse overbevisninger, tilknytning til faglige organisationer, biometriske data, helbredsdata og data om seksuel orientering (GDPR art. 22, stk. 4, se også betragtning 71).

38 Det er i den forbindelse værd at bemærke, at offentlige myndigheder er underlagt særlige krav, idet de ikke kan påberåbe sig legitim interesse som behandlingsgrundlag, jf. GDPR art. 6, samt at der er begrundet tvivl om, i hvilken udstrækning en arbejdsplads kan indhente et frit – og dermed gyldigt – samtykke fra medarbejdere (se afsnit 4.2.3 nedenfor).

4.1.3. Forskelsbehandlingsloven

Forskelsbehandlingsloven regulerer det danske arbejdsmarked, ved at forbyde visse måder, hvorpå en arbejdsplads kan forskelsbehandle medarbejdere og ansøgere.[39] Bestemmelserne har betydning for indsamling af medarbejderdata, men også for anvendelsen af automatiserede beslutningssystemer, som kan have algoritmisk bias, der påvirker relevante grupper. Lovens bestemmelser inkluderer blandt andet:

- At en arbejdsgiver ikke må forskelsbehandle lønmodtagere eller ansøgere til ledige stillinger ved ansættelse, afskedigelse, forflyttelse, forfremmelse eller med hensyn til løn- og arbejdsvilkår (§2, stk.1). Forskelsbehandling kan i den forbindelse være direkte, når en person på grund af race, hudfarve, religion eller tro, politisk anskuelse, seksuel orientering, alder, handicap eller national, social eller etnisk oprindelse behandles ringere end en anden bliver, er blevet eller ville blive behandlet i en tilsvarende situation (§1, stk.2). Forskelsbehandling kan også være indirekte, når en bestemmelse, et kriterium eller en praksis, der tilsyneladende er neutral, reelt vil stille personer af en bestemt race, hudfarve, religion eller tro, politisk anskuelse, seksuel orientering eller national, social eller etnisk oprindelse eller med en bestemt alder eller med handicap ringere end andre personer, medmindre den pågældende bestemmelse, betingelse eller praksis er objektivt begrundet i et sagligt formål og midlerne til at opfylde det er hensigtsmæssige og nødvendige (§1, stk.3).
- At en arbejdsgiver i forbindelse med eller under ansættelsen af en lønmodtager ikke må anmode om, indhente eller modtage og gøre brug af oplysninger om dennes race, hudfarve, religion eller tro, politiske anskuelse, seksuelle orientering eller nationale, sociale eller etniske oprindelse (§4).

4.1.4. Arbejdsmiljøloven

Arbejdsmiljøloven er af central betydning i reguleringen af det danske arbejdsmarked.[40] Loven er ikke i udgangspunktet rettet mod indsamling af medarbejderdata eller anvendelse af automatiserede beslutningssystemer, men indeholder ikke desto mindre mere generelle bestemmelser, som i visse tilfælde kan være relevante. De vigtigste bestemmelser er:

- At arbejdsgivere er forpligtet til at foretage systematiske vurderinger af arbejdsmiljøet på arbejdspladsen. Sådanne vurderinger kan også omfatte risici forbundet med indsamling og anvendelse af medarbejderdata. Arbejdsgiveren skal i den forbindelse identificere og vurdere potentielle negative virkninger på arbejdsmiljøet, og iværksætte passende foranstaltninger for at minimere disse (§15a).
- At arbejdsgiveren er forpligtet til at sikre et sundt arbejdsmiljø, og at dette jf. lovens formålsbestemmelse (§1) omfatter det psykiske arbejdsmiljø. Sikringen af det psykiske

39 https://www.retsinformation.dk/eli/lta/2017/1001. Loven implementerer det europæiske direktiv om ligebehandling med hensyn til beskæftigelse og erhverv: https://eur-lex.europa.eu/legal-content/DA/TXT/HTML/?uri=CELEX:32006L0054

40 https://www.retsinformation.dk/eli/lta/2021/2062

arbejdsmiljø kan være særlig relevant, fordi indsamling og anvendelse af medarbejderdata kan tænkes at påvirke dette.

4.1.5. Aftalen om kontrolforanstaltninger mellem arbejdsmarkedets parter

I tillæg til de juridiske rammer, som er defineret i international og national lovgivning, har arbejdsmarkedets parter i Danmark indgået en juridisk bindende aftale om kontrolforanstaltninger, der på en række måder definerer grænser for arbejdspladsens mulighed for, at indsamle og anvende medarbejderdata.[41] LO og DA's aftale om kontrolforanstaltninger fastslår blandt andet:

- At kontrolforanstaltninger kun må indføres, hvis de er sagligt begrundet og har et fornuftigt formål (punkt 1).
- At kontrolforanstaltninger ikke må være krænkende for medarbejderne eller forvolde tab eller nævneværdig ulempe, samt at de skal indrettes således, at der er et rimeligt forhold mellem formål og midler (punkt 1).
- At medarbejdere, med visse undtagelser, skal underrettes om nye kontrolforanstaltninger senest 6 uger inden de iværksættes. Undtagelserne omfatter situationer, hvor formålet med kontrolforanstaltningen forpurres ved underretning, og situationer hvor der findes såkaldt "tvingende driftsmæssige grunde". Ved sådanne undtagelser skal medarbejdere underrettes hurtigst muligt (punkt 2).
- At den enkelte medarbejder ikke kan samtykke til indsamling af data – hverken i forbindelse med ansættelsen eller på et senere tidspunkt (punkt 3).
- Samt, at kontrolforanstaltninger ved hjemmearbejdspladser ikke må "krænke privatlivets fred" (punkt 4).

4.2. Seks centrale juridiske rammer for indsamling og anvendelse af medarbejderdata

Konventioner, forordninger, love og aftaler definerer tilsammen de juridiske rammer for indsamling af medarbejderdata og anvendelse af automatiserede beslutningssystemer. Nogle af de væsentligste bestemmelser kan sammenfattes under seks juridiske rammer:

- Arbejdspladsens legitime interesse som grundlag for behandling af medarbejderdata
- Nødvendighed og proportionalitet som betingelser for behandling af medarbejderdata

41 https://fho.dk/wp-content/uploads/lo/2017/03/aftaleomkontrolforanstaltninger.pdf

- Medarbejderes samtykke som grundlag for behandling af medarbejderdata
- Forbud mod indsamling og anvendelse af særligt følsomme medarbejderdata
- Forbud mod anvendelse af fuldt automatiserede beslutningssystemer på arbejdspladsen
- Arbejdspladsens pligt til at oplyse medarbejdere om indsamling og anvendelse af medarbejderdata

I dette afsnit skitserer vi disse seks rammer, som regulerer arbejdspladsers muligheder for at indsamle medarbejderdata, og anvende automatiserede beslutningssystemer, inden vi efterfølgende ser på hvordan reguleringen kan udvikle sig i de kommende år.

4.2.1. Arbejdspladsens legitime interesse som grundlag for behandling af medarbejderdata

En central bestemmelse i lovgivningen er, at behandling af medarbejderdata skal foregå på et lovligt grundlag. Ét sådant grundlag er arbejdspladsens "legitime interesse" i at behandle data. Et oplagt spørgsmål er derfor, hvad det betyder, at en arbejdsplads har en legitim interesse i at behandle medarbejderdata?

En meget almindelig interesse som italesættes af de virksomheder, som udvikler produkter til indsamling af medarbejderdata, er forbedring af arbejdspladsens sikkerhed. Denne sikring kan handle om at opdage og lukke uintenderede sikkerhedsbrister, som en medarbejder, der kommer til at efterlade en dør ulåst, eller installerer software på en arbejdscomputer, der gør arbejdspladsens systemer sårbare. Men den kan også dreje sig om at opdage og forhindre tyveri, svindel, eller deling af fortrolig information. Udover interessen i sikkerhed kan arbejdspladser eksempelvis have en legitim interesse i at indsamle data for at forbedre ledelsens beslutningsgrundlag, ved at måle medarbejderes trivsel, produktivitet, tidsforbrug på forskellige opgaver, eller brug af arbejdspladsens ressourcer.

Præcist hvad der kan og ikke kan tælle som en legitim interesse er ikke juridisk veldefineret, men begrebet må fortolkes relativt bredt. Indtil 2018 blev de europæiske databeskyttelsesregler fortolket af et rådgivende organ nedsat af EU, den såkaldte "Artikel 29 gruppe".[42] Gruppen har i et positionspapir lagt op til at begrebet "legitim interesse" må fortolkes således, at en arbejdsplads har en legitim interesse når denne interesse er i) aktuel og reel, ii) lovlig, og iii) kan præciseres i en grad, så den kan afvejes med modstridende interesser.[43] For denne brede fortolkning er det således først og fremmest kravene om, at databehandling skal være nødvendig og proportionel, som sætter grænser for arbejdspladsers mulighed for at behandle medarbejderdata.

42 Artikel 29 gruppen blev i 2018 erstattet af Det europæiske databeskyttelsesråd (EDPB).

43 Article 29 Data Protection Working Party, *Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC* (2014), https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf.

4.2.2. Nødvendighed og proportionalitet som betingelser for behandling af medarbejderdata

Udover at være baseret på et lovligt grundlag, stiller lovgivningen også krav om, at arbejdspladsens behandling af medarbejderdata skal være nødvendig og proportionel.

At arbejdspladsens indsamling af medarbejderdata er nødvendig betyder, at de pågældende data kan anvendes til et specifikt og legitimt formål, typisk at forfølge en af arbejdspladsens legitime interesser, og at dette formål ikke på rimelig vis kan indfries på anden måde. Kravet medfører, at arbejdspladser ikke må indsamle flere eller andre medarbejderdata end de, som er nødvendige til formålet.

Når grundlaget for behandling af medarbejderdata er arbejdspladsens legitime interesse, skal databehandlingen også være proportionel, i den konkrete betydning, at arbejdspladsens interesse i databehandling skal veje tungere, end de modstridende interesser, som medarbejderen måtte have. Det kan, som artikel 29 gruppen har påpeget, være vanskeligt at sige, præcist hvordan domstole i praksis skal foretage denne afvejning.[44] På et overordnet niveau er det imidlertid klart, at kravet om interesseafvejning betyder, at de grunde, som taler for at behandle data, skal være tilstrækkeligt tungtvejende, til at opveje de grunde, som taler imod behandling af data. Arbejdspladsens forskellige interesser kan i denne henseende have forskellig vægt – visse interesser er vægtigere end andre – ligesom behandling af data kan tjene arbejdspladsens interesser i forskellig grad. I nogle tilfælde vil behandling af data gøre en stor forskel for en stærk interesse. I sådanne tilfælde vil grundene til fordel for databehandling veje relativt tungt. I andre tilfælde vil behandling af data gøre en lille forskel eller tjene en mindre vigtig interesse. I sådanne tilfælde vil grundene til fordel for databehandling veje mindre tungt. Tilsvarende kan databehandling påvirke forskellige interesser og rettigheder hos medarbejderne, og påvirke dem i forskellig grad. I tilfælde, hvor databehandling i høj grad vil have negativ virkning på medarbejderes stærke interesser eller grundlæggende rettigheder, vil der være meget tungtvejende grunde, som taler mod databehandlingen. I andre tilfælde, vil databehandling kun have lille indvirkning på medarbejdernes interesser og rettigheder, eller påvirke mindre vægtige interesser. I sådanne tilfælde vil der være mindre tungtvejende grunde, som taler imod databehandlingen.

4.2.3. Medarbejderes samtykke som grundlag for behandling af medarbejderdata

Et alternativt grundlag for databehandling er medarbejderens samtykke til indsamling og anvendelse af data. Når grundlaget er samtykke, skal arbejdspladsens databehandling fortsat leve op til de generelle krav, for eksempel om, at behandlingen er nødvendig for opfyldelse af et legitimt formål. Men arbejdspladsen behøver ikke at leve op til kravet om, at arbejdspladsens interesse skal veje tungere end medarbejderens interesser og grundlæggende rettigheder. Samtykke udgør også en mulig undtagelse fra det generelle forbud mod anvendelse

44 Article 29 Data Protection Working Party, *Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC*.Se især afsnit III.3.3 og III.3.4.

af fuldt automatiserede beslutningssystemer som har væsentlig effekt på medarbejdere (se afsnit 4.2.5 nedenfor).

Samtykke fordrer til gengæld, at medarbejderen er informeret. Det betyder, at medarbejderen forstår præcis hvilken databehandling der er tale om, og hvilke konsekvenser den kan have, således at vedkommende kan samtykke til specifikt denne databehandling. Et samtykke fordrer også, at medarbejderen frit kan tage stilling – et samtykke er ugyldigt, hvis en person samtykker under pres, for eksempel under trusler om, at det vil have ansættelsesmæssige konsekvenser, at nægte samtykke.[45]

Kravet om, at et samtykke skal være afgivet frit, har fået den tidligere artikel 29 gruppe til, at udtrykke betænkelighed ved medarbejderes muligheder for, at afgive et juridisk gyldigt samtykke til indsamling af medarbejderdata på arbejdspladsen (se også afsnit 5.6 om moralsk relevant samtykke). Et ansættelsesforhold indebærer som regel en ulige magtrelation mellem arbejdsgiver og den enkelte medarbejder, som kan gøre det vanskeligt for medarbejderen at samtykke frit. Selv når nægtelse af samtykke ikke vil have negative konsekvenser for medarbejderen, så kan alene bekymringen for eller mistanken om, at dette kunne være tilfældet, underminere medarbejderens mulighed for frit at samtykke til databehandling.[46]

4.2.4. Forbud mod indsamling og anvendelse af særligt følsomme medarbejderdata

Selv når en arbejdsplads indfrier de øvrige krav, for at indsamle og anvende medarbejderdata, er der visse typer medarbejderdata, som arbejdspladsen kun i særlige situationer må indsamle og anvende. Det drejer sig især om de særligt følsomme data, som er omdrejningspunktet for forbud mod diskrimination, blandt andet data om etnicitet, religiøse og politiske overbevisninger, seksuel identitet, og tilhørsforhold til faglige organisationer.

Disse typer data må kun indsamles, når særlige betingelser er opfyldt, ligesom fuldt automatiserede beslutninger der i væsentlig grad påvirker medarbejdere, kun i visse situationer må anvende sådanne data.

---

45 GDPR, betragtning 32 præciserer, at: "Consent should be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject's agreement to the processing of personal data relating to him or her."

46 Article 29 Data Protection Working Party, *Opinion 2/2017 on data processing at work* (2017), https://ec.europa.eu/newsroom/article29/redirection/document/45631.: "Employees are almost never in a position to freely give, refuse or revoke consent, given the dependency that results from the employer/employee relationship. Given the imbalance of power, employees can only give free consent in exceptional circumstances, when no consequences at all are connected to acceptance or rejection of an offer." Se også GDPR, betragtning 43: "in order to ensure that consent is freely given, consent should not provide a valid legal ground for the processing of personal data in a specific case where there is a clear imbalance between the data subject and the controller…" GDPR artikel 7, stk. 4 slår også fast, at "When assessing whether consent is freely given, utmost account shall be taken of whether, *inter alia*, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract." Se også betragtning 43. I dansk kontekst har Datatilsynet i en afgørelse om anvendelse af automatiseret beslutningsstøtte på danske jobcentre fra 2022 slået fast, at samtykke ikke kunne danne grundlag for databehandling, fordi det pågældende samtykke ikke kunne anses for frivilligt. Se Datatilsynet, *Udtalelse fra Datatilsynet: Kommuners hjemmel til AI-profileringsværktøjet Asta* (2022), https://www.datatilsynet.dk/afgoerelser/afgoerelser/2022/maj/udtalelse-vedroerende-kommuners-hjemmel.

4.2.5. Forbud mod anvendelse af fuldt automatiserede beslutningssystemer på arbejdspladsen

Lovgivningen giver relativt brede muligheder for, at anvende automatiserede beslutningssystemer på arbejdspladsen. Den fastsætter dog et forbud mod afgørelser, som træffes *alene* af et automatiseret beslutningssystem, når disse beslutninger i væsentlig grad påvirker medarbejderen. Med undtagelse af situationer, hvor en lov specifikt tillader anvendelse af et sådant system i en konkret sammenhæng, og situationer hvor medarbejderen har givet samtykke til anvendelsen af systemet, er sådan anvendelse af automatiserede beslutningssystemer altså ulovlig.[47]

Forbuddet rejser spørgsmålet om, hvordan bestemmelsen om, at afgørelsen skal træffes af et automatiseret beslutningssystem alene, skal fortolkes. Hvis formuleringen fortolkes snævert, så vil bestemmelsen kun sjældent finde anvendelse, idet selv en helt symbolsk involvering af en menneskelig beslutningstager kan være tilstrækkeligt til at gøre anvendelsen af det automatiserede beslutningssystem lovlig.[48] Artikel 29 gruppen har derfor lagt op til at fortolke loven således, at bestemmelsen finder anvendelse i et bredere sæt af situationer.[49] I denne fortolkning kræver bestemmelsen, at en menneskelig beslutningstager tager stilling til beslutningen, herunder vurderer de data, som systemet har anvendt, og har mulighed for at omgøre beslutningen.

4.2.6. Arbejdspladsens pligt til at oplyse medarbejdere om indsamling og anvendelse af medarbejderdata

Et sidste krav som er værd at fremhæve er, at arbejdspladsen altid skal oplyse medarbejdere om, at den indsamler medarbejderdata, og hvilke formål de anvendes til. Medarbejdere har i den forbindelse en række rettigheder til blandt andet indsigt, klageadgang, og til at få rettet fejlbehæftede data.

Udover oplysning om indsamling af medarbejderdata vil arbejdspladsen i nogle situationer også være forpligtet til at oplyse medarbejdere om anvendelsen af et automatiseret beslutningssystem. Det er i udgangspunktet tilfældet, når arbejdspladsen anvender et fuldt automatiseret beslutningssystem til at træffe afgørelser, som i væsentlig grad påvirker medarbejdere (jf. afsnit 4.2.5 ovenfor). Et væsentlig spørgsmål i den forbindelse angår *hvilke* oplysninger, som arbejdspladsen i denne situation skal give medarbejdere. Også hvad dette spørgsmål angår,

---

47 Bemærk i den forbindelse, at de betænkeligheder, som knytter sig til medarbejderes mulighed for frit at samtykke til databehandling på arbejdspladsen, også gør sig gældende her, jf. afsnit 4.2.3.

48 Se Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation" *International Data Privacy Law* 7, no. 2 (2017) https://papers.ssrn.com/abstract=2903469.

49 Article 29 Data Protection Working Party, *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* (2017), https://ec.europa.eu/newsroom/article29/redirection/document/49826: "To qualify as human intervention, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision. As part of the analysis, they should consider all the available input and output data."

er det muligt, at anlægge henholdsvis mere lempelige og mere krævende fortolkninger, af arbejdspladsens forpligtelser. For en lempelig fortolkning er arbejdspladsen alene forpligtet til, at formidle generel information om systemets formål, generelle funktioner og forventede resultater. Information af denne type vil typisk være let at opnå, let at kommunikere, og vil normalt ikke risikere at afsløre fortrolige detaljer om systemets funktionalitet. Til gengæld vil den kun i meget begrænset omfang gøre medarbejdere i stand til at forstå, vurdere, og om nødvendigt udfordre systemets beslutninger. Blandt andet derfor argumenterer visse jurister for en mere krævende fortolkning, hvor arbejdspladsen er forpligtet til at forklare hvordan systemet teknisk fungerer på en måde, så de berørte medarbejdere kan forstå det.[50] En sådan forpligtelse vil typisk være langt vanskeligere for arbejdspladsen at løfte.

4.3. Fremtidig regulering af området

Der sker i disse år en hastig udvikling i reguleringen af digitale teknologier, blandt andet med EU's kommende forordning om kunstig intelligens. Samtidig har forskere, interesseorganisationer og tænketanke stillet en række forslag om yderligere regulering for at begrænse indsamling af medarbejderdata og anvendelse af automatiseret beslutningsstøtte på arbejdspladsen.[51] Sådanne forslag inkluderer:

- Et forbud mod indsamling af visse typer medarbejderdata, samt et forbud mod indsamling af medarbejderdata til visse formål.
- Et snævert nødvendighedskriterium for anvendelse af automatiserede beslutningssystemer på arbejdspladsen, også når systemet anvendes til beslutningsstøtte.
- En vidtrækkende forpligtelse til oplysning af den enkelte medarbejder, inklusive information om automatiserede beslutningssystemer, som anvendes til beslutningsstøtte.
- En forpligtelse til at informere arbejdspladsens medarbejderrepræsentanter om indsamling og anvendelse af medarbejderdata for arbejdspladsens medarbejdere.
- Et forbud mod alle former for anvendelse af automatiserede beslutningssystemer, også beslutningsstøtte, til visse typer beslutninger, især beslutninger der har vidtrækkende konsekvenser for medarbejderen.

Samlet ville gennemførelsen af sådanne forslag begrænse arbejdspladsers mulighed for at indsamle og anvende medarbejderdata betragteligt. Forslagene illustrerer at der både politisk og mellem arbejdsmarkedets parter foregår en intens debat om, hvordan teknologierne skal have mulighed for, at forme arbejdsmarkedet fremover.

---

50 Se Andrew D Selbst and Julia Powles, "Meaningful information and the right to explanation" *International Data Privacy Law* 7, no. 4 (2017), https://doi.org/10.1093/idpl/ipx022, https://doi.org/10.1093/idpl/ipx022.

51 Jeremias Adams-Prassl, "Regulating algorithms at work: Lessons for a 'European approach to artificial intelligence'" *European Labour Law Journal* 13, no. 1 (2022); Trade Union Congress, *Technology Managing People*; De Stefano, "'Negotiating the algorithm': Automation, artificial intelligence and labour protection."; AI Now Institute, *Algorithmic Management: Restraining Workplace Surveillance*.

4.3.1. EU's kommende forordning om kunstig intelligens

Et centralt initiativ i den igangværende udvikling af regulering er den kommende EU-forordning om kunstig intelligens.[52] Siden 2021 har man i EU forhandlet om en forordning, som skal regulere udvikling og anvendelse af kunstig intelligens i medlemslandene. Forhandlingen af denne forordning er på tidspunktet for denne rapport (vinteren 2023) i en afsluttende fase, men ikke afsluttet. Når forordningen træder i kraft kan den få afgørende betydning for arbejdspladsers muligheder for at indsamle medarbejderdata og anvende automatiserede beslutningssystemer, idet især udvikling og anvendelse af automatiserede beslutningssystemer kan blive underlagt nye krav. Selvom forordningen ikke er vedtaget, er det værd her at fremhæve nogle af de relevante bestemmelser, som med stor sandsynlighed vil blive indført med forordningen.[53]

Den overordnede tilgang i forordningen er, at skelne mellem forskellige former for og anvendelser af kunstig intelligens, afhængigt af hvilke risici de udsætter mennesker for.[54] Forordningen arbejder med fire niveauer, fra "uacceptable risici", over "høje risici", "begrænsede risici", og til "minimale risici". Forordningen lægger op til at underlægge især de første kategorier forbud og øget kontrol, mens begrænsede og minimale risici kun i mindre omfang underlægges strammere regulering.

Forordningens eksempler på uacceptable risici drejer sig om kunstig intelligens som i) anvendes til at manipulere borgeres adfærd på en måde som skader borgeren, eksempelvis gennem subliminale teknikker eller ved at udnytte kognitiv sårbarhed hos visse grupper, som børn, ældre eller personer med handicap, ii) anvendes af offentlige myndigheder til at overvåge og give borgere en "social score", og iii) anvendes til realtids biometrisk identifikation (med enkelte undtagelser) for eksempel i form af ansigtsgenkendelse. Denne kategori af risici forbydes, således at kunstig intelligens ikke må udvikles eller anvendes til disse formål eller med disse funktioner. De måder, arbejdspladser typisk indsamler og anvender medarbejderdata, vil således næppe klassificeres som tilhørende kategorien af uacceptable risici.

Automatiserede beslutningssystemer på arbejdspladsen falder til gengæld eksplicit ind under systemer, som klassificeres som højrisiko. For udviklere af højrisiko systemer stiller forordningen krav om datakvalitet, teknisk dokumentation, transparens, menneskelig kontrol,

---

52 Der forhandles også i EU om det såkaldte af "Direktiv om at forbedre arbejdsvilkårene ved platformsarbejde", https://data.consilium.europa.eu/doc/document/ST-14450-2021-INIT/en/pdf. Når direktivet vedtages kan det have stor betydning for indsamling og anvendelse medarbejderdata for den gruppe medarbejdere, som arbejder i den såkaldte platformsøkonomi. Her fokuserer vi imidlertid på AI forordningen, som vil have betydning for medarbejdere på tværs af sektorer

53 Europaparlamentet vedtog i Juni 2023 et forslag til AI forordningen med en række ændringer, som i efteråret 2023 udgør udgangspunktet for forhandlinger i EU: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf. For kritisk analyse af forordningen, se eksempelvis Connor Dunlop, *An EU AI Act that works for people and society*, Ada Lovelace Institute (2023), https://www.adalovelaceinstitute.org/policy-briefing/eu-ai-act-trilogues/; Claudio Novelli et al., "Taking AI Risks Seriously: A New Assessment Model for the AI Act" *AI & Society* 38, no. 3 (2023), https://doi.org/10.1007/s00146-023-01723-z; Michael Veale and Frederik Zuiderveen Borgesius, "Demystifying the Draft Artificial Intelligence Act" *Computer Law Review International* 4 (2021)

54 Forordningen er, som nævnt, ikke vedtaget. Når vi i dette afsnit henviser til "forordningen" refererer vi således til de offentligt tilgængelige udkast, som i vinteren 2023 forhandles. Det er vigtigt at holde sig for øje, at forordningen, når den vedtages, kan se anderledes ud.

robusthed, præcision og cybersikkerhed. I forhold til de juridiske rammer, som vi ovenfor har skitseret, pålægges udviklere konkret:

- at udføre kvalitetskontrol og risikoanalyser for systemet.
- at sikre tilstrækkelig kvalitet i de data, som systemet trænes på. Udviklere får i denne forbindelse undtagelsesvis tilladelse til at anvende særligt følsomme persondata, for eksempel data om personers etnicitet, med det eksklusive formål at teste systemers tendens til algoritmisk bias og indirekte diskrimination, herunder bias i de data som systemet trænes på.
- at sikre, at systemet har et vist niveau af gennemsigtighed for brugeren, dog kun ved at brugere skal kunne fortolke og anvende systemets resultater med hjælp fra dokumentation og vejledninger leveret af udvikleren
- at sikre, at mennesker kan kontrollere systemet, ved at give brugere af systemet adgang til informationer om systemet, og muligheder for at ændre en beslutning.

4.4. Dataetik og jura i denne rapport

Som vi ovenfor har set, så sætter både dansk og international lovgivning juridiske rammer for indsamling og anvendelse af medarbejderdata, ved eksempelvis at begrænse indsamling af medarbejderdata, som ikke i tilstrækkelig grad tjener arbejdspladsens legitime interesse, relativt til den omkostning indsamlingen af data har for medarbejdernes interesser og rettigheder. Lovgivningen forbyder også i udgangspunktet brug af fuldt automatiserede beslutningssystemer til beslutninger som væsentligt påvirker medarbejdere, og stiller krav til, at medarbejdere skal oplyses om indsamling og anvendelse af data.

De juridiske rammer kan virke både som et oplagt udgangspunkt og en udfordring for dataetiske overvejelser. Mange vil eksempelvis nok have en umiddelbar følelse af, at det er ikke kun ulovligt, men også moralsk problematisk, hvis en arbejdsplads overtræder de juridiske grænser for indsamling og anvendelse af medarbejderdata.[55] Når der findes intuitivt tiltalende juridiske rammer, så kan det være fristende at slutte direkte fra disse rammer til hvad der er etisk rigtigt og forkert: Hvis en indsamling af medarbejderdata er lovlig, så er den etisk; hvis anvendelsen af et automatiseret beslutningssystem er ulovlig, så er den uetisk. Hvis man laver denne slutning kan det være nærliggende også at tænke, at der ikke er behov for yderligere dataetisk vurdering. Hvis lovgivningen definerer både de juridiske og de etiske grænser, så kan en selvstændig dataetisk vurdering fremstå som spildt arbejde.

I dette afsnit skitserer vi forholdet mellem dataetik og jura, og peger på hvordan dataetisk analyse kan være relevant både for arbejdspladsers indsamling og anvendelse af medarbejderdata indenfor lovens rammer, og for overvejelser om ændring af eksisterende regulering.

55 "Etik" og "moral" anvendes ind imellem med forskellige betydninger, for eksempel således at etik rummer universelle fordringer mens moral rummer en kulturs sædvaner. I denne rapport anvendes de to udtryk imidlertid synonymt, på linje med den dominerende praksis i den moderne forskningslitteratur.

4.4.1. Dataetik er mere end lovlighed

Hvis handlinger er etisk korrekte, når de er lovlige, og etisk forkerte, når de er ulovlige, så behøver man ikke bekymre sig om dataetik – man kan nøjes med at fokusere på de juridiske rammer for indsamling og anvendelse af medarbejderdata. Det er også i nogle sammenhænge almindeligt, at eksempelvis en arbejdsplads, som udsættes for kritik, besvarer kritikken ved at henvise til, at man blot har fulgt gældende regler. Implicit i denne type svar ligger en forudsætning om, at man ikke handler etisk forkert, når man følger loven. Et første og afgørende spørgsmål er derfor, om der findes en sådan enkel sammenhæng mellem dataetik og jura? Kan en handling være uetisk, selvom den er lovlig, eller etisk, selvom den er ulovlig?

Når man besvarer disse spørgsmål er det vigtigt, at holde sig for øje, at dataetik og jura kan overlappe, uden at de nødvendigvis hænger sammen. Selv hvis lovlige handlinger typisk er etiske, og ulovlige handlinger typisk er uetiske, så kan dette være et resultat af, at vi som samfund har været dygtige til at regulere, snarere end af at etik og jura uundgåeligt følges ad. Og når man tænker efter, så er der meget der taler for, at handlinger kan være moralske forkerte, selvom de er lovlige, og etiske, selvom de er ulovlige.[56] Det kan let illustreres ved at pege på to typer eksempler.

Den første type eksempler angår handlinger, som har været lovlige, men siden er blevet kriminaliserede. I Danmark blev "revselsesretten", det vil sige forældres mulighed for lovligt at udøve fysisk vold som afstraffelse af egne børn, først definitivt afskaffet i 1997.[57] Det er naturligvis muligt at mene, at det først var med lovændringen i 1997, at det blev moralsk forkert for forældre, at slå deres børn. Men det forekommer også muligt at mene, og som et måske mere plausibelt synspunkt, at det var forkert for forældre at slå deres børn, også inden det i 1997 blev ulovligt. I så fald var der tale om en lovlig, men moralsk forkert handling.

Den anden type eksempler drejer sig om handlinger, som har været ulovlige, men som siden er blevet afkriminaliseret. Det kunne eksempelvis være sex mellem voksne, samtykkende personer før ægteskabet eller når personerne har samme køn. Begge typer seksuelle handlinger har været ulovlige i Danmark, men de færreste vil i dag mene, at sådanne handlinger er moralsk problematiske.[58] Hvis handlinger nødvendigvis er moralske forkerte, når de er ulovlige, må man hævde, at disse handlinger var moralsk forkerte indtil deres lovliggørelse, og først med lovliggørelsen blev etiske. Ligesom med eksemplerne ovenfor, så forekommer et alternativt, og mere plausibelt synspunkt at være, at disse handlinger aldrig har været moralsk forkerte, og at lovliggørelsen derfor ikke har ændret på deres moralske status. Der er snarere tale om, at lovene var uretfærdige, fordi disse handlinger ikke er eller var moralsk forkerte, og ikke bør kriminaliseres.

---

56 En klassisk analyse af, hvordan jura og etik stiller forskellige krav til personers handlinger, findes i Kenneth Einar Himma, "Positivism, Naturalism, and the Obligation to Obey Law" *The Southern Journal of Philosophy* 36, no. 2 (1998), https://doi.org/https://doi.org/10.1111/j.2041-6962.1998.tb01749.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-6962.1998.tb01749.x

57 "Lov om ændring af lov om forældremyndighed og samvær (Afskaffelse af revselsesretten)", Lov nr 416 af 10/06/1997, https://www.retsinformation.dk/eli/lta/1997/416.

58 Sex mellem voksne, samtykkende personer af samme køn blev lovliggjort i Danmark i 1930. https://historielab.dk/til-undervisningen/kildebank/koen-og-seksualitet/3-paa-kanten/den-grimme-lov/?kilde= Seksuelle forhold før ægteskabet, i datidens lovgivning kendt som "lejermål", blev endeligt afkriminaliseret i 1866. https://danmarkshistorien.dk/vis/materiale/seksuallovgivning-foer-1849

De to typer eksempler viser, at etik og jura er principielt forskellige. Spørgsmålet om, hvorvidt indsamling og anvendelse af medarbejderdata på en arbejdsplads er dataetisk eller ej, kan derfor ikke besvares alene ved at afklare om indsamling og anvendelse er lovlig. Dataetikken må vurderes selvstændigt.

4.4.2. Dataetik inden for lovens rammer

Selvom dataetik og jura må vurderes selvstændigt, tager denne rapport udgangspunkt i, at arbejdspladser overholder gældende lov. På den måde er de juridiske rammer med til at definere hvilke formål, en dataetisk analyse kan tjene.

Som det fremgår af dette kapitels redegørelse, efterlader de juridiske rammer et bredt råderum for, at arbejdspladser på lovlig vis kan indsamle og anvende medarbejderdata. Men fordi jura og dataetik kan rejse forskellige krav, er det ikke givet, at enhver lovlig indsamling og anvendelse af medarbejderdata er etisk. For en arbejdsplads, som ønsker at handle dataetisk, er det afgørende, at kunne skelne mellem etiske og uetiske måder at indsamle og anvende medarbejderdata. I disse situationer kan dataetisk analyse hjælpe med at evaluere, hvilke af de lovlige måder at indsamle og anvende medarbejderdata, som også er etiske. Dataetikken lægger sig, i dette perspektiv, i forlængelse af lovgivningen.

Dataetisk analyse kan også være relevant for beslutningstagere og interessenter – både arbejdspladser, fagforeninger, arbejdsgiverorganisationer, interesseorganisationer og politikere – som overvejer, om der er grund til at ændre eksisterende regulering. Et almindeligt argument for at ændre eksisterende lovgivning er netop, at lovgivningen forbyder handlinger, som ikke er etisk forkerte, eller at lovgivningen tillader handlinger, som er etisk forkerte. Dataetikken udgør i dette perspektiv en central del af beslutningsgrundlaget, når man skal vurdere, om der er behov for ny eller ændret regulering.

De dataetiske hensyn, som vi behandler i denne rapport er relevante for begge typer overvejelser. De kan informere dataetiske vurderinger af, hvordan en arbejdsplads bør handle *indenfor det råderum*, som loven giver. De kan også informere vurderinger af, om visse former for indsamling og anvendelse af medarbejderdata bør *reguleres anderledes*, enten ved at handlinger som aktuelt er tilladt forbydes, eller ved at aktuelt forbudte handlinger tillades (se afsnit 4.3 for eksempler på forslag til ændret regulering).

I de to næste kapitler præsenterer og diskuterer vi centrale dataetiske hensyn som knytter sig til medarbejderes privatliv, når arbejdspladser indsamler medarbejderdata, og til risici for fejl og bias, når arbejdspladser anvender automatiserede beslutningssystemer.

5. Etiske udfordringer ved indsamling af medarbejderdata – overvågning og privatliv på arbejde

Et afgørende træk ved moderne teknologier som indsamler medarbejderdata er, at disse teknologier kan reducere medarbejderes privatliv. Privatliv er en egenskab, som mange umiddelbart vil mene er moralsk vigtig. Det er også let at pege på former for dataindsamling, som de fleste mener er moralsk forkert, fordi de krænker privatlivet. Et enkelt eksempel kunne være en arbejdsplads, som indsamler data fra et kamera, der er installeret i medarbejdernes omklædningsrum. Omvendt findes der også former for dataindsamling, som kan forekomme ikke blot moralsk acceptable men moralsk prisværdige, for eksempel en børnehave, som indhenter en såkaldt børneattest ved ansættelse af nye medarbejdere. Et vigtigt spørgsmål er således, hvorfor vi oplever nogle former for dataindsamling som etisk problematiske krænkelser af privatlivet, og andre som acceptable eller endda ønskværdige? Og hvad skal vi mene om alle de mange former for dataindsamling, hvor det ikke er klart, hvad arbejdsgivere bør eller ikke bør gøre, af respekt for medarbejderes privatliv?

For at besvare sådanne spørgsmål, er det nødvendigt at overveje hvilken moralsk ret til privatliv medarbejdere har på arbejdspladsen. En sådan overvejelse kræver blandt andet at man tager stilling til, hvad det vil sige at have privatliv på en arbejdsplads. Men det kræver også, at man overvejer hvad det vil sige at have en ret til privatliv, og hvad der kan gøre det moralsk problematisk, at reducere medarbejderes privatliv.

For at besvare sådanne spørgsmål, er det nødvendigt at overveje hvilken moralsk ret til privatliv medarbejdere har på arbejdspladsen. En sådan overvejelse kræver blandt andet at man tager stilling til, hvad det vil sige at have privatliv på en arbejdsplads. Men det kræver også, at man overvejer hvad det vil sige at have en ret til privatliv, og hvad der kan gøre det moralsk problematisk, at reducere medarbejderes privatliv.

5.1. Hvad er privatliv på arbejdet?

Når man taler om "privatliv" eller det, at noget er "privat", så kan det betyde flere forskellige, beslægtede ting.[59]

En første betydning drejer sig om, at visse beslutninger har en privat karakter. Derved menes typisk, at det er beslutninger som den enkelte har ret til at træffe, uden at andre blander sig.

59 I meget af forskningslitteraturen tales om det engelske "privacy", som er nært beslægtet med, men også subtilt anderledes end det danske "privatliv". Hvor det engelske "privacy" kan forstås som de betingelser eksempelvis information skal indfri for at være privat, så betegner det danske privatliv i højere grad et domæne af for eksempel informationer som har (og måske også bør have) den egenskab at være private. Ikke desto mindre anvender vi her privatliv som det på dansk mest egnede udtryk.

Eksempelvis ville mange nok opleve det som grænseoverskridende, hvis en anden person uopfordret forsøgte at diktere hvilken sport man skulle dyrke i sin fritid, eller hvem man skulle være kærester med. Når det gælder private beslutninger, så tillader vi ofte allerhøjst velmente råd fra andre, og det er en almindelig erfaring, at selv sådanne råd skal fremsættes med stor omhu, for ikke at virke påtrængende.

Det er imidlertid ikke sådanne private beslutninger som er i fokus i spørgsmålet om privatliv ved indsamling af medarbejderdata på arbejdspladsen.[60] I denne sammenhæng er det snarere private, personlige informationer, som er på spil.[61] Information er personlig, når den er information *om* en person. Det er imidlertid værd at bemærke, at meget personlig information er information om ting, som på forskellig vis er knyttet til personer. Det er således i den relevante forstand en personlig information om en person, at farven på vedkommendes bil er gul.

Det er også værd at skelne mellem en værdiladet og en beskrivende betydning af privatliv. Når vi snakker om "privatliv" eller det, at noget er "privat" på dansk, så er begreberne ofte værdiladede på den måde, at vi også mener, at der er en grund til at respektere privatlivet, eller sikre, at det pågældende forbliver privat. Det kender de fleste nok fra den type situation, hvor en person afviser at besvare et spørgsmål, for eksempel om finansielle eller familiemæssige forhold, ved at sige at "det er privat". Det kan imidlertid være en fordel at skille spørgsmålene om, hvornår noget reducerer en persons privatliv, og hvornår det er moralsk forkert at reducere en persons privatliv, ad.[62] Det første er et rent deskriptivt spørgsmål, som handler om for eksempel hvorvidt andre personer har kendskab til en personlig information, mens det andet er et etisk spørgsmål. Når begrebet privatliv anvendes på den værdiladede måde kan det være svært at skelne det rent deskriptive spørgsmål fra det etiske spørgsmål. Derfor bruger vi i denne rapport privatliv i den lidt specielle, rent deskriptive betydning, og taler om de etiske spørgsmål under betegnelsen, at der kan være en "moralsk ret til privatliv". Når vi i denne rapport skriver, at visse teknologier "reducerer privatliv", så betyder det altså ikke at disse teknologier nødvendigvis krænker en ret til privatliv, eller på anden vis er moralsk problematiske.

---

60 Indsamling af medarbejderdata kan måske i nogle tilfælde påvirke medarbejderes mulighed for at træffe private beslutninger, men denne problemstilling er i givet fald en konsekvens af den forudgående indsamling af privat, personlig information, som vi her fokuserer på.

61 På engelsk tales i forskningslitteraturen om "informational privacy". Forskellen mellem private beslutninger og privat information diskuteres blandt andet af Herman T. Tavani, "Philosophical theories of privacy: Implications for an adequate online privacy policy" *Metaphilosophy* 38, no. 1 (2007); Alan Rubel, "The Particularized Judgment Account of Privacy" *Res Publica* 17, no. 3 (2011).

62 Dette er en central og ofte understreget pointe. Se Steven Davis, "Is there a right to privacy?" *Pacific Philosophical Quarterly* 90, no. 4 (2009); Ruth Gavison, "Privacy and the Limits of Law" *The Yale Law Journal* 89, no. 3 (1980), https://doi.org/10.2307/795891, http://www.jstor.org/stable/795891; H. J. McCloskey, "Privacy and the Right to Privacy" *Philosophy* 55, no. 211 (1980); Madison Powers, "A cognitive access definition of privacy" *Law and Philosophy* 15, no. 4 (1996); William A. Parent, "Privacy, morality, and the law" *Philosophy and Public Affairs* 12, no. 4 (1983).

5.2. To teorier om privatliv for personlig information

Privatliv kan imidlertid betyde flere forskellige ting, også når det gælder personlige informationer. På et generelt niveau kan man skelne mellem to udbredte måder at forstå, hvad der skal til for at en persons personlige informationer er private: kontrolteorien og kendskabsteorien.[63]

Fælles for de to teorier er, som den amerikanske jurist Alan Rubel har påpeget, at privatliv ikke er et enten-eller, men et spørgsmål om grader.[64] En personlig information kan være mere eller mindre privat, og en person kan have mere eller mindre privatliv. Rubel peger også på, at graden af privatliv synes af afhænge af relationer til andre personer. En information er meget privat, hvis få mennesker kender til eller har adgang til den, og mindre privat, hvis mange personer kender til eller har adgang til den. Og en medarbejder har mere privatliv, jo mere private personens informationer er, og mindre privatliv, jo mindre private de er.

I de næste to afsnit introducerer vi først kontrolteorien og så kendskabsteorien, og præsenterer nogle af de styrker og svagheder, som er blevet fremhævet ved hver af teorierne.

5.2.1. Kontrolteorien om privatliv

Den første udbredte teori om privatliv er, at en medarbejders personlige information er privat i den udstrækning den pågældende person kan kontrollere hvem, som har adgang til denne information.[65]

Det betyder først og fremmest, at information er mindre privat, jo mindre kontrol medarbejderen har over andre personers adgang til informationen. En medarbejder kan have forskellige grader af kontrol relativt til forskellige andre personer. Medarbejderen kan have en høj grad af kontrol over information i forhold til nogle personer, og en meget lav grad eller ingen kontrol i forhold til andre personer. For privatliv som hele betyder det, at medarbejderen har

63 Det er værd at understrege, at der fortsat er en intens diskussion i faglitteraturen om, hvordan privatliv skal forstås, herunder om fordele og ulemper ved at skelne mellem kontrol og adgangsteorier. Se Haleh Asgarinia, "Convergence of the source control and actual access accounts of privacy" *AI and Ethics* (2023), https://doi.org/10.1007/s43681-023-00270-z, https://doi.org/10.1007/s43681-023-00270-z; Leonhard Menges, "A Defense of Privacy as Control" *The Journal of Ethics* 25, no. 3 (2021), https://doi.org/10.1007/s10892-020-09351-1, https://doi.org/10.1007/s10892-020-09351-1; Jakob Mainz, "An Indirect Argument for the Access Theory of Privacy" *Res Publica* 27, no. 3 (2021), https://doi.org/10.1007/s11158-021-09521-4, https://doi.org/10.1007/s11158-021-09521-4; Björn Lundgren, "A Dilemma for Privacy as Control" *The Journal of Ethics* 24, no. 2 (2020), https://doi.org/10.1007/s10892-019-09316-z, https://doi.org/10.1007/s10892-019-09316-z; Lauritz Munch and Jakob Mainz, "To Believe, or Not to Believe – That is Not the (Only) Question: The Hybrid View of Privacy" *The Journal of Ethics* (2023), https://doi.org/10.1007/s10892-023-09419-8, https://doi.org/10.1007/s10892-023-09419-8; Helen Nissenbaum, "Privacy as contextual integrity" *Washington Law Review* 79, no. 1 (2004); Beate Roessler, *The Value of Privacy* (Polity Press, 2005).

64 Alan Rubel, "Claims to Privacy and the Distributed Value View" Article, *San Diego Law Review* 44, no. 4 (2007), http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=31197946&site=ehost-live.

65 A.F. Westin, *Privacy and Freedom* (New York: Ig Publishing, 1967); Andrei Marmor, "What Is the Right to Privacy?" *Philosophy and Public Affairs* 43, no. 1 (2015); James Rachels, "Why privacy is important" *Philosophy & Public Affairs* 4, no. 4 (1975); Adam D. Moore, "Privacy: Its Meaning and Value" *American Philosophical Quarterly* 40, no. 3 (2003), http://www.jstor.org/stable/20010117.

mere privatliv, jo mere kontrol vedkommende samlet set har, over hvem som har adgang til personlige informationer.

Indsamling af medarbejderdata på arbejdspladsen kan indlysende reducere medarbejderes privatliv i denne forstand, ved at reducere medarbejderes kontrol over andre personers adgang til deres personlige informationer.

Kontrolteorien om privatliv:

En persons personlige information er mere privat relativt til en anden person, jo mere kontrol vedkommende har over den anden persons adgang til informationen. En personlig information er samlet set mere privat, jo mere kontrol personen har relativt til andre personer. En person har mere privatliv, jo mere private vedkommendes personlige informationer er.

Kontrolteorien er blevet kritiseret for, at den i nogle tilfælde synes at stille de forkerte krav.[66] Et eksempel som viser udfordringen kunne være følgende: En medarbejder sender en e-mail til en kollega, og vedhæfter ved en fejl et dokument fuld af følsomme personlige oplysninger. Medarbejderen opdager fejlen, og skynder sig at ringe til kollegaen og forklare situationen. Kollegaen sletter e-mailen uden at åbne dokumentet. Ifølge kontrolteorien er de personlige informationer i dokumentet ikke private i den periode, hvor e-mailen lå uåbnet i kollegaens indbakke. Men for mange vil det nok virke mere oplagt at sige, at informationerne også i den periode var private, fordi der faktisk ikke var nogen, som fik adgang til dem.[67]

Samlet vil nogle kritikere sige, at kontrolteorien er god til at pege på, at det kan være vigtigt at sikre, at personer har kontrol over adgang til deres personlige informationer, fordi sådan kontrol er en central måde at beskytte privatliv på, men også at det er denne rolle, som kontrol spiller, snarere end rollen som den afgørende betingelse for, at information er privat.

5.2.2. Kendskabsteorien om privatliv

Den anden centrale teori om privatliv er, at personlig information er privat i den udstrækning andre personer ikke har kendskab til den pågældende information.[68] Ligesom med kontrolteorien kan en personlig information her forstås som mere eller mindre privat afhængigt af hvor mange andre personer, som har kendskab til den pågældende information. Og ligesom

66 Davis, "Is there a right to privacy?."; Mainz, "An Indirect Argument for the Access Theory of Privacy."

67 Et beslægtet eksempel anvendes til at fremføre denne kritik mod kontrolteorien af den britiske filosof Kevin Macnish. Se Kevin Macnish, "Government Surveillance and Why Defining Privacy Matters in a Post-Snowden World" *Journal of Applied Philosophy* 35, no. 2 (2018).

68 Teorien behandles i den engelsksprogede forskningslitteratur under betegnelsen "the access-account of privacy", men jævnfør teoriens ide om, at adgang betyder, at man enten modtager eller kan genkalde sig information oversætter vi den her som "kendskabsteorien", fremfor den direkte oversættelse "adgangsteorien".

for kontrolteorien kan en medarbejder siges at have mere eller mindre privatliv, afhængigt af hvor private personens personlige informationer er.[69]

Det at have kendskab til en information skal i denne sammenhæng forstås bredt, for eksempel således, at man direkte observerer informationen, at man får den formidlet, eller at man tidligere har modtaget informationen, og nu kan huske den. Eksempelvis har en medarbejder, som lige nu ser en kollega udføre en arbejdsopgave, kendskab til den information, at kollegaen arbejder på opgaven. Tilsvarende har en medarbejder kendskab til information om hvilken opgave kollegaen arbejdede på i går, hvis vedkommende så kollegaen arbejde på opgaven, og i dag kan huske det. Omvendt er det næppe nok, at en person på et tidspunkt har haft kendskab til informationen. En medarbejder som har set en kollega udføre en arbejdsopgave, men glemt det igen, har ikke kendskab til information om, at kollegaen udførte opgaven. Den pågældende information er blevet privat igen, da medarbejderen glemte den. Det er heller ikke nok, at en person har adgang til information i den betydning, hvor vedkommende *kunne* få kendskab til informationen. En information om hvilken opgave en medarbejder arbejder på ophører ifølge kendskabsteorien ikke med at være privat, alene fordi en kollega kunne gå hen og se efter, men først hvis vedkommende rent faktisk går hen og ser efter.

Kendskabsteorien om privatliv:

En persons personlige information er privat relativt til en anden person, hvis vedkommende ikke har kendskab til informationen. En personlig information er samlet set mere privat, jo færre personer som har kendskab til informationen. En person har mere privatliv, jo mere private vedkommendes personlige informationer er.

Indsamling af medarbejderdata på arbejdspladsen kan indlysende reducere medarbejderes privatliv i denne forstand, ved at gøre andre personer bekendt med medarbejdernes personlige informationer.

Kritikere af kendskabsteorien har fremført, at den synes at give det forkerte svar i situationer, hvor en person frivilligt deler sine egne personlige informationer. Hvis for eksempel en medarbejder af egen drift fortæller en kollega eller leder om sine ideer eller bekymringer, så fører dette ifølge kendskabsteorien til, at medarbejderen har mindre privatliv, fordi andre personer derved får adgang til information om disse ideer og bekymringer. Men, hævder nogle kritikere, i sådanne tilfælde er det misvisende at påstå, at en persons privatliv er blevet reduceret.[70]

Et svar på indvendingen kan være, at den synes at trække på et værdiladet begreb om privatliv. Man kan måske forklare den intuitive forskel på frivillig og ufrivillig deling af information ved at pege på, at der kan være en moralsk forskel på de to situationer (se også afsnit 5.6 nedenfor om samtykke til indsamling af medarbejderdata). Det er også værd at påpege, at også kontrolteorien synes at møde udfordringen, fordi en medarbejder næppe kan siges

69 Parent, "Privacy, morality, and the law."; Powers, "A cognitive access definition of privacy."; Mainz, "An Indirect Argument for the Access Theory of Privacy."

70 Julie Inness, *Privacy, intimacy, and isolation* (Oxford: Oxford University Press, 1992).

stadig at have kontrol over eksempelvis en kollegas adgang til en personlig information, når medarbejderen (frivilligt) har delt informationen med kollegaen.[71]

5.2.3. Kontrol eller kendskab?

I forskningslitteraturen behandles de to teorier om privatliv typisk som konkurrerende ideer, der hver for sig forsøger at give "det rigtige" svar på, hvad privatliv er. Men det er værd at holde sig for øje, at der synes at kunne være dataetiske udfordringer både ved tab af kontrol over adgang til medarbejderdata og ved udbredelsen af kendskab til medarbejderdata.

Eksempelvis kan en medarbejder have mulighed for at slukke for et stykke software, som registrerer medarbejderens aktivitet på en computer, og i den forstand have kontrol over ledelsens adgang til disse informationer. Men selvom medarbejderen har sådan kontrol, så synes dataindsamlingen at kunne rejse dataetiske udfordringer knyttet til privatliv, i de tilfælde hvor medarbejderen ikke slukker for softwaren, og ledelsen derfor får adgang til informationerne. På lignende vis synes det at kunne rejse dataetiske udfordringer knyttet til privatliv, hvis en medarbejder arbejder på en computer, hvor software obligatorisk registrerer aktivitet, selv hvis ledelsen aldrig kigger på disse data, og derfor ikke får kendskab til medarbejderens personlige information

Selvom begge teorier synes at pege på relevante perspektiver på privatliv, så vil det ofte være en fordel at lægge sig fast på en bestemt betydning. Derved undgår man risikoen for forvirring om hvad det betyder, når man siger eller skriver "privatliv". I denne rapport diskuterer vi i udgangspunktet privatliv i den betydning, som kendskabsteorien definerer. Når vi i de følgende afsnit henviser til "privatliv", handler det altså om, i hvilken grad andre personer har kendskab til en medarbejders personlige information. Vi inddrager imidlertid løbende indsigter om de særegne dataetiske udfordringer, som det kan rejse, når medarbejdere savner kontrol over adgang til deres personlige informationer.

5.3. Hvad er retten til privatliv?

I de første afsnit har vi diskuteret hvordan man kan forstå det, at personlig information er privat, og at en medarbejder har privatliv. I dette og de følgende afsnit vender vi nu blikket imod spørgsmål om hvilken etisk rolle medarbejderes privatliv på arbejdspladsen spiller.

Det er både i den offentlige debat og i forskningen almindeligt at diskutere etik og privatliv ved at henvise til en *ret* til privatliv. Ligesom det er tilfældet med begrebet privatliv, så kan en ret til privatliv imidlertid forstås på flere forskellige måder.

En ret til privatliv kan for det første dreje sig om en juridisk rettighed. Sådanne rettigheder er veletablerede, i både dansk og international lovgivning. De optræder eksempelvis i Grundlovens paragraf 72, og i den europæiske menneskerettighedskonventions artikel 8, som

71 Parent, "Privacy, morality, and the law."; Lundgren, "A Dilemma for Privacy as Control."

beskytter borgernes privatliv i boligen, ved telefonsamtaler samt når de skriver til hinanden (se også afsnit 4 om juridiske rammer for indsamling af data).

I mange diskussioner om retten til privatliv er det imidlertid ikke en juridisk rettighed, som er på spil. Hvis eksempelvis en deltager i en debat kritiserer eksisterende lovgivning ved at henvise til, at den på utilstrækkelig vis sikrer retten til privatliv, så er det klart, at det ikke er den juridiske rettighed, som vedkommende har i tankerne. I sådanne tilfælde er der snarere tale om en moralsk ret til privatliv.

En moralsk ret til privatliv kan forstås på tre væsentligt forskellige måder: som en fundamental rettighed, som en sammenfattende rettighed, eller som en afledt rettighed.

5.3.1. En fundamental ret til privatliv

Den første måde at forstå en moralsk ret til privatliv er som et *grundlæggende* moralsk hensyn. Når den forstås på denne måde, er retten til privatliv ikke begrundet i andre, bagvedliggende moralske hensyn. Det er ganske enkelt moralsk problematisk i sig selv, hvis en person eller en arbejdsplads reducerer medarbejderes privatliv.

En fundamental moralsk ret til privatliv er en stærk rettighed, i den forstand at den ikke er afhængig af andre forhold. Men faktisk fortolkes retten til privatliv nogle gange endnu stærkere. En fundamental moralsk ret til privatliv forstås nemlig ind imellem således, at det moralske hensyn til privatliv *trumfer* andre moralske hensyn.[72] Det vil sige, at det ikke blot er moralsk problematisk, at handle på en måde, som fører til at en person mister privatliv. Sådanne handlinger er grundlæggende *moralsk forkerte*. I faglitteraturen kaldes denne type moralsk rettighed ofte "absolut".

En fundamental moralsk ret til privatliv:

En person har en moralsk ret til privatliv, hvis (og kun hvis) det er grundlæggende moralsk dårligt, at handle på en måde, som reducerer personens privatliv.

Forskellen på en handling, som i én henseende er moralsk problematisk, og en handling, som er moralsk forkert, er, at en handling som i én henseende er moralsk problematisk samlet set kan være moralsk rigtig. Det kræver blot at der er andre hensyn, som vejer tungere end hensynet til privatliv. Hvis hensynet til privatliv derimod trumfer andre hensyn, så er en handling som fører til tab af privatliv moralsk forkert, uanset hvilke grunde som måtte tale for handlingen.

Selvom ideen om en fundamental moralsk ret til privatliv umiddelbart kan virke tiltalende, så findes der stærke argumenter imod, at der skulle findes en sådan moralsk rettighed. Et almindeligt argument er, at en fundamental moralsk rettighed i mange situationer synes at stille for strenge krav, især i den absolutte variant, hvor retten til privatliv trumfer andre grunde.

72 Denne fortolkning af moralske rettigheder, som moralske hensyn der trumfer andre moralske hensyn, forbindes især med den amerikanske retsfilosof Ronald Dworkin. Se Ronald Dworkin, *Taking Rights Seriously* (London: Gerald Duckworth & Co. Ltd., 2005).

Når man tænker efter, så vil de fleste nok mene, at der er situationer, hvor det ikke er moralsk forkert at handle på en måde, som fører til tab af privatliv, hvis der er tilstrækkeligt stærke grunde til at handle på denne måde. Eksempelvis vil mange mene, at en forælder til et lille barn som er blevet væk, kan handle moralsk rigtigt ved at lede efter barnet i en anden persons have, selvom haveejerens privatliv derved reduceres. Den mest naturlige måde at forklare denne konklusion er, at hensynet til at beskytte barnet vejer tilstrækkeligt tungt til, at hensynet til privatliv må vige. Men hvis det er tilfældet, så trumfer hensynet til privatliv ikke andre hensyn.

En fundamental og absolut moralsk ret til privatliv:

En person har en moralsk ret til privatliv, hvis (og kun hvis) det er grundlæggende moralsk *forkert*, at handle på en måde, som reducerer personens privatliv.

Men også den svagere fortolkning af den fundamentale moralske ret til privatliv kan virke for stærk. En sådan rettighed ville medføre, at næsten alle mennesker meget ofte handler på måder, som i hvert fald i én henseende er moralsk dårlig. Når vi eksempelvis færdes i det offentlige rum, og betragter andre, som også færdes i det offentlige rum, så får vi derved information om for eksempel hvordan de ser ud og hvor de befinder sig. Derved reducerer vi disse personers privatliv, men det vil for mange virke besynderligt at sige, at denne måde at reducere personers privatliv på i nogen som helst forstand er moralsk dårlig.[73]

5.3.2. En sammenfattende ret til privatliv

Den anden måde at forstå en ret til privatliv på er som en *sammenfatning* af den moralske status i en konkret situation. Hvis man forstår retten til privatliv på denne måde, så har en person en ret til privatliv, hvis det samlet set vil være moralsk forkert, at reducere vedkommendes privatliv. Her er ideen altså, at man først kigger på alle de moralske hensyn, som påvirker spørgsmålet om, hvorvidt det er moralsk forkert at reducere privatliv på en bestemt måde og vejer dem mod hinanden. Først når man ved at overveje disse hensyn er nået frem til en konklusion om, hvorvidt det er moralsk forkert, at reducere privatliv, kan man afgøre om en person har en ret til privatliv.

En sammenfattende moralsk ret til privatliv:

En person har en moralsk ret til privatliv, hvis (og kun hvis) det samlet set er moralsk forkert, at reducere vedkommendes privatliv.

Udfordringen ved denne måde at forstå privatliv er, at retten til privatliv i denne betydning ikke bidrager til at afklare, hvad der er moralsk rigtigt og forkert. Ofte er det imidlertid netop spørgsmålet om hvorvidt det i en given sammenhæng er moralsk problematisk, at reducere

73 Versioner af denne type kritik af en fundamental ret til privatliv findes blandt andet i Ryberg, "Privacy Rights, Crime Prevention, CCTV, and the Life of Mrs. Aremac."; og Doyle, "Privacy and perfect voyeurism."

en persons privatliv, som vi ønsker at afklare, ved at vurdere om handlingen krænker en ret til privatliv. Når retten til privatliv forstås som en sammenfatning, så kan vi først vide, om der er en ret til privatliv, når vi på anden vis har besvaret dette spørgsmål.

5.3.3. En afledt ret til privatliv

Den tredje måde at forstå en moralsk ret til privatliv, er som et moralsk hensyn der afhænger af et eller flere andre, mere grundlæggende moralske hensyn. I denne betydning har man en moralsk ret til privatliv, når en vis mængde privatliv eller en bestemt form for privatliv, er nødvendig for at tilgodese disse hensyn.[74]

En afledt moralsk ret til privatliv:

En person har en moralsk ret til privatliv, hvis (og kun hvis) der er moralske grunde til ikke at reducere vedkommendes privatliv.

Styrken ved at forstå den moralske ret til privatliv som en afledt rettighed er dels, at det er et meget plausibelt synspunkt, at der kan være moralske grunde til ikke at reducere personers privatliv, og dels at man ved at fokusere på disse grunde kan *forklare*, hvorfor det er moralsk problematisk at reducere en persons privatliv. Udfordringen for en afledt ret til privatliv er oplagt, at man er nødt til at præcisere, hvad det er for yderligere moralske hensyn, som giver grund til at beskytte privatlivet. Retten til privatliv kan komme til at virke på meget forskellig vis, afhængigt af hvilke hensyn, man peger på.

En ret til privatliv kan altså forstås på mindst tre væsentligt forskellige måder. I princippet er de forskellige forståelser forenelige – det kunne godt være tilfældet, at der på samme tid findes både en fundamental og en afledt moralsk ret til privatliv. Men det er vigtigt, når man diskuterer indsamling af medarbejderdata og medarbejderes ret til privatliv, at de ikke bliver forvekslet eller blandet sammen. Det kræver at man i en konkret sammenhæng gør det klart, hvilken af de forskellige betydninger, man har i tankerne. Når vi i de næste afsnit henviser til en moralsk ret til privatliv, er det i udgangspunktet en afledt rettighed, som er på tale.

Hvis retten til privatliv forstås som en sammenfattende eller afledt rettighed er det afgørende at afklare, hvilke etiske hensyn som kan begrunde beskyttelse af medarbejderes privatliv. Derfor introducerer og diskuterer vi i næste afsnit en række af de etiske hensyn, som kan begrunde en sådan ret til privatliv.

74 Et beslægtet synspunkt fungerer som udgangspunkt for en kritik af, at der overhovedet skulle findes en moralsk ret til privatliv. Denne kritik forbindes især med den amerikanske filosof Judith Jarvis Thomson. Se Judith Jarvis Thomson, "The Right to Privacy" *Philosophy & Public Affairs* 4, no. 4 (1975), https://doi.org/10.2307/2265075, http://www.jstor.org/stable/2265075.

5.4. Hvorfor privatliv på arbejdspladsen?

Som vi har set er én måde at forstå en moralsk ret til privatliv på, at en person har en ret til privatliv når der findes moralske grunde til at beskytte personens privatliv. Et afgørende spørgsmål er derfor, hvilke sådanne grunde, man kan tænke sig. Spørgsmålet er blevet diskuteret i en omfattende forskningslitteratur, som peger på en række forskellige etiske hensyn, som potentielt kan begrunde en ret til privatliv.[75] Blandt de mest fremtrædende er:

- Tab af privatliv kan være ydmygende
- Tab af privatliv kan føre til observationsstress
- Tab af privatliv kan gøre det vanskeligt at opretholde eller forme personlige relationer
- Tab af privatliv kan uhensigtsmæssigt afskrække adfærd
- Tab af privatliv kan gøre personer sårbare overfor andre agenter

I de følgende afsnit gennemgår vi kort disse ideer om etisk relevante hensyn, og skitserer hvordan de kan begrunde en ret til privatliv på arbejdsmarkedet.

5.4.1. Privatliv og ydmygelse

Måske den første begrundelse for privatliv som de fleste vil tænke på er, at privatliv kan beskytte personer mod at andre får kendskab til information, som det vil være ydmygende, pinligt eller skamfuldt for den pågældende person, at dele med andre. Oplevelsen af at blive ydmyget gennem afsløring af privat information er typisk i sig selv umiddelbart ubehagelig for personen, men en sådan oplevelse kan også skade selvværd, social status og personlige relationer.

I visse tilfælde kan information være ydmygende, fordi den afslører at en person har handlet på en måde, som andre vil finde kritisabel. Det kan eksempelvis være tilfældet, hvis ny information afslører, at en medarbejder har snydt med udførelsen af sine arbejdsopgaver. I sådanne tilfælde er vi tilbøjelige til at have begrænset sympati med vedkommende. Mange vil mene, at oplevelsen af at blive ydmyget er personens eget ansvar, og at det oplevede ubehag derfor kun i beskedent omfang eller slet ikke kan begrunde en ret til privatliv.

Men det er vigtigt at holde sig for øje, at meget information kan have en karakter, hvor det kan være ydmygende, at andre får kendskab til den, selvom informationen ikke indikerer at personen har gjort noget forkert. Eksempelvis ville langt de fleste nok opleve det som ydmygende, hvis arbejdspladsen indsamlede detaljerede data om toiletbesøg, og delte denne information med ledelse eller kollegaer. I den situation skyldes ubehaget ikke, at det man foretager sig i løbet af et almindeligt toiletbesøg er moralsk problematisk, men alene at det er af så intim karakter, at vi nødigt vil dele information om det med andre.

---

75 For overblik, se eksempelvis Kevin Macnish, "An Eye for an Eye: Proportionality and Surveillance" *Ethical Theory and Moral Practice* 18, no. 3 (2015); Daniel J. Solove, "A Taxonomy of Privacy" *University of Pennsylvania Law Review* 154, no. 3 (2006).

Indsamling af medarbejderdata på arbejdspladsen kan rejse en risiko for ydmygelse på flere måder. Først og fremmest kan indsamlingen, ligesom i eksemplet med data om toiletbesøg ovenfor, være målrettet data, som medarbejdere vil opleve det som pinligt at dele. Men dataindsamling kan også føre til deling af potentielt ydmygende information ved et tilfælde. Et eksempel kunne være en arbejdsplads, som indsamler medarbejderdata om de aktiviteter, som medarbejdere har på internettet via deres arbejdscomputer, og i den forbindelse får adgang til data om en medarbejders lovlige men atypiske seksuelle fetish. Endelig kan indsamling af informationer, som hver for sig ikke har en karakter, som gør, at personer oplever det som ydmygende at dele dem, i nogle tilfælde gøre det muligt at udlede information som har denne karakter. Et eksempel kunne være en arbejdsplads som indsamler en række forskellige medarbejderdata, der tilsammen indikerer, at en medarbejder har et helbredsproblem, som medarbejderen oplever det som pinligt at dele information om.

Visse typer information vil, som i eksemplet med toiletbesøget, i næsten alle tilfælde være information, som medarbejdere oplever det som ydmygende at dele. Men for mange typer information vil det afhænge af kontekst og personlighed, hvordan medarbejdere oplever det, at dele informationen. Det kan derfor være vanskeligt på generelt niveau, at skelne mellem de informationer som har og ikke har denne karakter. Ikke desto mindre er det klart, at ydmygelse kan udgøre en begrundelse for en ret til privatliv, ligesom *risikoen* for ydmygelse kan udgøre en mulig begrundelse for en ret til kontrol over privatliv.

5.4.2. Privatliv og observationsstress

En anden mulig begrundelse for privatliv hviler på den simple betragtning, at det for de fleste mennesker er stressfuldt at føle sig observeret eller overvåget. Det, at udføre en opgave, mens man oplever at blive observeret eller overvåget, har normalt en række både mentale og fysiologiske stresseffekter, som er veldokumenterede i den psykologiske forskning.[76] Men de fleste mennesker vil nok også have erfaret disse effekter i deres eget liv. Mange vil eksempelvis have oplevet, at der er stor forskel på at holde en tale alene foran spejlet, og på at holde den samme tale foran en større forsamling. Indsamling af medarbejderdata risikerer at skabe eller øge observationsstress, simpelthen ved at udbrede eller intensivere oplevelsen

76 "Elisa Giacosa et al., "Stress-inducing or performance-enhancing? Safety measure or cause of mistrust? The paradox of digital surveillance in the workplace" 10.1016/j.jik.2023.100357, *Journal of Innovation & Knowledge* 8, no. 2 (2023), https://doi.org/10.1016/j.jik.2023.100357, https://www.elsevier.es/en-revista-journal-innovation-knowledge-376-articulo-stress-inducing-or-performance-enhancing-safety-measure-S2444569X23000537; N. Backhaus, "Context Sensitive Technologies and Electronic Employee Monitoring: a Meta-Analytic Review", 2019 IEEE/SICE International Symposium on System Integration (2019); M. J. Smith et al., "Employee stress and health complaints in jobs with and without electronic performance monitoring" *Applied Ergonomics* 23, no. 1 (1992), https://doi.org/https://doi.org/10.1016/0003-6870(92)90006-H, https://www.sciencedirect.com/science/article/pii/000368709290006H; Daniel M. Ravid et al., "A meta-analysis of the effects of electronic performance monitoring on work outcomes" *Personnel Psychology* 76, no. 1 (2023), https://doi.org/https://doi.org/10.1111/peps.12514, https://onlinelibrary.wiley.com/doi/abs/10.1111/peps.12514. Et beslægtet og mere komplekst spørgsmål er, om sådanne stresseffekter hæmmer eller gavner præstationer. Her tyder moderne forskning på, at dette både kan afhænge af hvilken type opgave, den pågældende person skal løse, og variere med personlighed. John R. Aiello and Kathryn J. Kolb, "Electronic performance monitoring and social context: Impact on productivity and stress" *Journal of Applied Psychology* 80 (1995), https://doi.org/10.1037/0021-9010.80.3.339; Devasheesh P. Bhave, "The Invisible Eye? Electronic Performance Monitoring and Employee Job Performance" *Personnel Psychology* 67, no. 3 (2014), https://doi.org/https://doi.org/10.1111/peps.12046, https://onlinelibrary.wiley.com/doi/abs/10.1111/peps.12046; P. J. Hills et al., "Being observed caused physiological stress leading to poorer face recognition" *Acta Psychol (Amst)* 196 (May 2019), https://doi.org/10.1016/j.actpsy.2019.04.012."

af at blive observeret eller overvåget på arbejdspladsen. Forskningen peger i den forbindelse på, at observationsstress afhænger af en række faktorer, blandt andet hvor mange data som bliver indsamlet, hvilke data som bliver indsamlet, hvad det erklærede formål med indsamlingen af data er, samt relationer og kultur på arbejdspladsen.

Stress er en form for psykologisk og fysiologisk alarmberedskab. Mange mennesker oplever stress en gang imellem, når der er travlt på arbejdet eller med familien. Hvis en medarbejder oplever stress en gang imellem og i mindre doser, kan det være et begrænset problem. Men stress kan have alvorlige konsekvenser, hvis en person oplever intenst stress eller stress i længere perioder. Medarbejdere vil derfor typisk opleve stærkere negative konsekvenser, jo mere intenst de oplever observationsstress, jo længere perioder de oplever observationsstress, og jo mere stress de i forvejen oplever på arbejdspladsen.

De umiddelbare negative effekter af stress udgør i sig selv et velfærdstab for medarbejderen, som de fleste vil opfatte som etisk dårligt, men de umiddelbare effekter kan også have en række afledte negative konsekvenser, for eksempel for den pågældende medarbejders sociale og kollegiale relationer, for de personer som indgår i disse relationer, og for medarbejderens karriere.

Det er også værd at hæfte sig ved, at observationsstress er en effekt af at *opleve* at data bliver indsamlet. Det betyder både, at hemmelig indsamling af data ikke har en sådan effekt, og at personer kan opleve effekten når de tror, at de bliver observeret eller overvåget, selvom det ikke er tilfældet. Tab af privatliv medfører derfor en risiko for at skabe observationsstress, men tab af kontrol over privatliv kan også have denne effekt, fordi en person som mister denne kontrol kan frygte at blive overvåget eller observeret, også når det ikke er tilfældet.

Når observationsstress og de negative effekter af observationsstress afhænger af mange faktorer, så er der ikke en enkel sammenhæng mellem indsamling af medarbejderdata og moralsk problematisk observationsstress. Men det er klart, at indsamling af medarbejderdata kan risikere at skabe observationsstress. Selvom risikoen for observationsstress må vurderes konkret i den sammenhæng, hvor en arbejdsplads indsamler medarbejderdata, så er denne risiko en mulig begrundelse for en ret til privatliv på arbejdspladsen.

5.4.3. Privatliv og personlige relationer

En tredje mulig begrundelse er, at privatliv kan være nødvendigt for, at personer kan forme og opretholde relationer til andre.[77]

De fleste kender til, at personer deler forskellige dele af deres liv og forskellige sider af deres personlighed med venner, kollegaer, romantiske partnere, og med deres børn. Argumentet for privatliv er, at en del af grundlaget for at man kan have *forskellige* slags relationer er, at vi netop kan dele eller ikke dele forskellige sider af os selv. Det kan for eksempel handle om, at en medarbejder påtager sig en professionel identitet på arbejdspladsen, som på væsentlige måder er anderledes end den identitet vedkommende har i andre sammenhænge. Og det

77 Marmor, "What Is the Right to Privacy?."; Thomas Nagel, "Concealment and Exposure" *Philosophy and Public Affairs* 27, no. 1 (1998); Rachels, "Why privacy is important."

kan være en forudsætning for den professionelle identitet, og for karakteren af de relationer som knytter sig til den, at medarbejderens ledelse, kollegaer og kunder har adgang til nogle personlige informationer, men ikke har adgang til andre. En anden ofte fremført pointe er, at nogle vigtige relationer er baseret på intimitet, hvor personer deler følsom information om sig selv, som de fleste andre ikke har adgang til, og derved skaber sympati og tillid. Dét, at personer besidder en vis mængde privatliv, er således ifølge argumentet forudsætningen for, at personer kan skabe sådanne intime relationer til andre.

En udfordring for begrundelsen er, at det synes at være muligt at have forskellige relationer selvom man mister ganske meget privatliv. Selv hvis det er rigtigt, at en person, som slet ikke havde noget privatliv, ikke ville kunne have forskellige relationer, så følger det ikke, at man mister muligheden for at have meningsfuldt forskellige relationer ved at miste *noget* af sit privatliv. En medarbejder, som opretholder sin professionelle identitet ved at holde dele af sin person og sit liv privat, kunne antageligt fortsætte med dette, selvom ganske mange andre informationer blev indsamlet eller delt med arbejdspladsen. Hvis retten til privatliv begrundes ved at henvise til, at privatliv kan være nødvendigt for at have forskellige, meningsfulde relationer, så risikerer det altså at være en rettighed, som kun gælder i et begrænset sæt af særlige tilfælde.

En måde at udvide rettighedens rækkevidde, er ved at pege på betydningen af *kontrol* over privatliv. I mange tilfælde kan tab af privatliv, selv hvis det ikke fjerner muligheden for at have meningsfuldt forskellige relationer, nok påvirke personers evne til at forme disse relationer. Hvis personers mulighed for at forme relationer på bestemte måder er moralsk vigtig, så kan hensynet begrunde en moralsk ret til privatliv også i alle de situationer, hvor tab af privatliv blot begrænser en persons evne til at forme relationer.

En udfordring for denne variant af argumentet er imidlertid, at det ikke er klart, at personers mulighed for at forme relationer ved at holde information privat altid er moralsk værdifuld. Eksempelvis kunne en erfaren medarbejder, som mobber en ung og usikker kollega, have et stærkt ønske om, at begrænse andres kendskab til dette, for at bevare nære og tillidsfulde relationer til andre kollegaer. Men det er ikke indlysende, at mobberens mulighed for at forme relationer til andre, så de ikke er præget af denne information, er moralsk værdifuld. Sådanne eksempler synes at vise, at det måske kun er når muligheden for at forme relationer tjener et godt formål, at den kan begrunde en ret til privatliv.

5.4.4. Privatliv og afskrækkelse

En fjerde begrundelse for privatliv, som ofte optræder i debatter om overvågning, henviser til at indsamling af medarbejderdata kan have en negativ, afskrækkende effekt på personers adfærd.

Det er en almindelig antagelse i mange sammenhænge, at indsamling af data kan påvirke den måde personer handler. Videoovervågning i det offentlige rum begrundes eksempelvis ofte

med, at overvågning afskrækker potentielle kriminelle fra at begå forbrydelser.[78] Tilsvarende kunne man forvente, at indsamling af data på arbejdspladsen vil afskrække medarbejdere fra eksempelvis svindel, tyveri, brug af arbejdstiden på personlige gøremål, og andre former for uønsket adfærd. I sådanne tilfælde kan den afskrækkende effekt være ønskværdig.

Udgangspunktet for argumentet for privatliv er, at dataindsamling også kan afskrække adfærd, som der er grund til at beskytte eller fremme. I faglitteraturen omtales dette som en "kølende effekt" (eng. "chilling"). Effekten har blandt andet været meget diskuteret i tilknytning til overvågning på internettet og i det offentlige rum. Forskere har i den forbindelse udtrykt bekymring for, at sådan overvågning kan afskrække borgere fra demokratiske aktiviteter, såsom deltagelse i politisk debat på internettet og demonstrationer i det offentlige rum.[79] Ligesom man kan være bekymret for en kølende effekt ved overvågning på nettet, kan man oplagt være bekymret for, at indsamling af data på arbejdspladsen kan afskrække medarbejdere fra aktiviteter, som vi burde beskytte eller støtte. En risiko kunne eksempelvis være, at medarbejdere på en arbejdsplads, som ikke er organiseret i en fagforening, kan afskrækkes fra at forsøge at organisere sig, hvis de frygter at ledelsen vil opdage dette initiativ gennem indsamling af medarbejderdata. Men der kunne også være tale om, at medarbejdere afskrækkes fra almindelige sociale aktiviteter, fra selvstændige initiativer og løsninger, eller fra arbejde som er værdifuldt, men som i mindre grad afspejles i de indsamlede data.

Hvis indsamling af data på en arbejdsplads afskrækker medarbejdere fra adfærd, og der er etisk grund til at beskytte denne adfærd, så kan denne effekt begrunde en ret til privatliv. Retten til privatliv afhænger derfor af både hvornår indsamling af medarbejderdata har en afskrækkende effekt, og af hvor tungtvejende grunde der er, til at beskytte den pågældende adfærd.

5.4.5. Privatliv og sårbarhed

Den sidste begrundelse for en ret til privatliv handler om hvordan kendskab til information kan gøre personer sårbare for andre agenters handlinger. Begrundelsen hviler på to trivielle forhold: at de måder personer handler på ofte afhænger af hvilke personlige informationer om andre de har adgang til, og at nogle måder personer behandler hinanden på gør skade.[80]

Begrundelsen er almindelig både i tilknytning til data og nye teknologier, og i tilknytning til arbejdsmarkedet. Med et banalt eksempel så er informationer om en persons adgangskoder

---

78 Det er værd at bemærke, at det diskuteres i forskningen, om videoovervågning i det offentlige rum faktisk har en kriminalpræventiv effekt. Brandon C. Welsh and David P. Farrington, "Public Area CCTV and Crime Prevention: An Updated Systematic Review and Meta-Analysis" *Justice Quarterly* 26, no. 4 (2009); Gustav Alexandrie, "Surveillance cameras and crime: a review of randomized and natural experiments" *Journal of Scandinavian Studies in Criminology and Crime Prevention* 18, no. 2 (2017), https://doi.org/10.1080/14043858.2017.1387410, https://doi.org/10.1080/14043858.2017.1387410.

79 Det er i denne forbindelse værd at påpege, at ligesom med de kriminalpræventive effekter af overvågning er der ganske svag evidens for at overvågning har en afskrækkende effekt på demokratisk aktivitet. Elizabeth Stoycheff et al., "Privacy and the Panopticon: Online mass surveillance's deterrence and chilling effects" *New Media & Society* 21, no. 3 (2019), https://doi.org/10.1177/1461444818801317, https://journals.sagepub.com/doi/abs/10.1177/1461444818801317; Jonathon W. Penney, "Internet surveillance, regulation, and chilling effects online: a comparative case study" *Internet Policy Review* 6, no. 2 (2017); Jonathon W. Penney, "Chilling effects: Online surveillance and Wikipedia use" *Berkeley Technology & Law Journal* 31, no. 1 (2016).

80 Carrisa Velíz, "The Internet and Privacy" in *Ethics and the Contemporary World*, ed. David Edmonds (Abingdon: Routledge, 2019).

til e-mail- og bankkonti meget følsomme, i den forstand at deling af sådanne informationer gør personen sårbar, fordi den giver andre personer mulighed for at handle på måder, der kan skade personen. Et tilsvarende velkendt eksempel fra arbejdsmarkedet er forbuddet mod at opsøge information om aktuel eller planlagt fremtidig graviditet i forbindelse med ansættelse. Forbuddet er motiveret af, at ansøgeren bliver sårbar for at blive frasorteret, hvis vedkommende deler sådanne informationer.

I en nøddeskal er begrundelsen altså, at der er informationer, som det kan være vigtigt at holde private, fordi deling af informationerne gør personer sårbare for andre agenters handlinger. Forskellige typer information kan gøre en person sårbar på forskellige måder. Udover de allerede nævnte, kunne tab af privatliv eksempelvis gøre medarbejdere sårbare for:

- videredeling af data
- afpresning
- manipulation
- diskrimination
- sammenstilling af data, som afslører ny information.[81]

Tab af privatliv gør ofte en person sårbar overfor videredeling. Det kræver blot, at personens data i en eller anden form kan deles, og at det er dårligt for personen, hvis disse data deles. I den offentlige debat er et kendt eksempel uønsket deling af intime billeder. En person, som deler intime billeder af sig selv med en anden, gør sig selv sårbar for videredeling af disse billeder, fordi det ofte vil være ydmygende, at andre ser dem.

I forlængelse af den første sårbarhed kan tab af privatliv gøre en person sårbar for afpresning. Det er tilfældet når videredeling af en information vil være omkostningsfuld for personen, eksempelvis fordi det vil være ydmygende. I den situation kan personen presses til at handle eller acceptere handlinger, som vedkommende ikke ønsker, under trussel om at informationen deles.

Tab af privatliv kan gøre en person sårbar for manipulation, hvis informationen gør andre i stand til at vurdere hvordan personens individuelle psykologiske profil kan udnyttes til at få vedkommende til at tænke eller handle på bestemte måder.

Tab af privatliv kan gøre en person sårbar for diskrimination hvis informationen selv kan udgøre et grundlag for diskrimination, som i eksemplet med information om aktuel eller planlagt graviditet, eller hvis den kan anvendes til statistisk at forudsige et grundlag for diskrimination.

Endelig kan tab af privatliv gøre en person sårbar for yderligere tab af privatliv, ved at andre analyserer informationen og udleder ny information. Den nye information kan i sig selv være følsom, hvis det eksempelvis er ydmygende, at dele den med andre, eller den kan gøre medarbejderen yderligere sårbar. Med et lidt dramatisk eksempel, så kunne ledelsen på en arbejdsplads sammenstille forskellige informationer om medarbejdere, som hver for sig er trivielle, og derved forsøge at identificere netop den medarbejder, som anonymt har anmeldt arbejdspladsen til Arbejdstilsynet. I dette eksempel gør deling af de oprindelige informationer

81 Se eksempelvis Solove, "A Taxonomy of Privacy."; Macnish, "An Eye for an Eye: Proportionality and Surveillance."

medarbejdere sårbare overfor udledning af ny information, og udledning af denne information gør den medarbejder, som har anmeldt arbejdspladsen, sårbar overfor repressalier fra ledelsen. Den sårbarhed, som handler om at andre kan bruge data til at udlede ny information, er vigtig i diskussionen om nye digitale ledelsesværktøjer på arbejdspladsen, fordi automatiserede beslutningssystemer ofte har netop den funktion at udlede ny information. Et automatiseret beslutningssystem kan eksempelvis bruge indsamlede medarbejderdata til at udlede ny information om medarbejderes præstation, og ledelsen kan efterfølgende bruge denne information til beslutninger om opgavefordeling, lønforhandling, eller endda afskedigelse (se afsnit 3.4).

Sårbarhed handler om risikoen for at andre agenter kan bruge information til at handle på måder som skader en person. Det kan derfor variere fra person til person og fra kontekst til kontekst, hvilke informationer en person bliver sårbar af at dele. Begrundelsens styrke kan også variere, afhængigt af hvor stor risikoen er, og hvor alvorlig skaden vil være. Begrundelsen må altså præciseres og vurderes for netop den type sårbarhed som optræder for bestemte personer i en bestemt kontekst.

Samtidig er det klart, at hensynet til ikke at gøre personer sårbare må afvejes mod andre hensyn, og at det i nogle sammenhænge kan veje tungere end andre. Mange mener eksempelvis, at det er etisk acceptabelt når en arbejdsplads indhenter information om en ansøgers straffeattest, og frasorterer ansøgere på baggrund af denne information. Umiddelbart er der ligesom i eksemplet med information om graviditet tale om, at den pågældende information gør ansøgeren sårbar for at blive frasorteret. Når der intuitivt er forskel på de to situationer kan det skyldes, at vi vurderer, at en arbejdsplads kan have en legitim interesse i at frasortere ansøgere med plettet straffeattest. Omvendt vil mange mene, at arbejdspladser er etisk forpligtede til at påtage sig risikoen for den omkostning, som det udgør, når en medarbejder bliver gravid og tager orlov. I givet fald er der ikke på samme måde et modsatrettet hensyn, som kan begrunde at man gør ansøgere sårbare.

5.5. Hvad er det særlige ved privatliv på arbejdspladsen?

Privatliv kan spille forskellige roller i forskellige sammenhænge. Et oplagt spørgsmål er derfor hvad der karakteriserer privatliv i den særlige sammenhæng, som en arbejdsplads udgør? Her kan fremhæves i hvert fald fire særlige forhold.[82]

For det første giver arbejdspladsen potentielt en enkelt aktør – ledelsen – mulighed for at indsamle store mængder personlige data, som både kan være meget præcise, og afspejle mange forskellige sider af medarbejderens aktiviteter og person. Ved at kombinere data fra forskellige kilder er det i princippet muligt at skabe et både detaljeret og omfattende billede af medarbejderen.

82 Se Mena Angela Teebken, "What Makes Workplace Privacy Special? An Investigation of Determinants of Privacy Concerns in the Digital Workplace" (paper presented at the AMCIS, 2021).

Samtidig er medarbejderen i arbejdssituationen typisk underlagt snævre rammer for sin opførsel. Arbejdspladsen kan stille krav til og sætte grænser for mange forskellige forhold lige fra medarbejderens påklædning over hvad medarbejderen skal foretage sig, og til hvilken attitude medarbejderen skal påtage sig i forhold til arbejdet, kollegaer, og kunder eller klienter.

For det tredje er medarbejderen i arbejdssituationen underlagt en konkret og nærværende trussel om sanktioner, hvis medarbejderen ikke indfrier de forpligtelser og forventninger, som findes på arbejdspladsen. Disse rækker fra mere bløde sociale sanktioner, over formel irettesættelse og tab af lønforhøjelse eller forfremmelse, og til afskedigelse.

Endelig er arbejde og tilknytning til arbejdspladsen for mange en meget væsentlig del af deres liv og personlige identitet. Arbejdet har, udover den økonomiske værdi, ofte en følelsesmæssig værdi, idet den rolle man spiller på arbejdspladsen, og det sociale netværk som man indgår i, er med til at forme både ens selvforståelse og andres forståelse af hvem man er som person.

Tilsammen er disse vilkår med til på nogle måder at gøre medarbejdere særligt udsatte ved tab af privatliv. Mest oplagt er medarbejdere sårbare overfor ledelsens reaktioner. Denne sårbarhed kan videre tænkes at påvirke observationsstress og afskrækkende effekter. Udfordringerne med, at tab af privatliv kan være ydmygende, og kan påvirke medarbejderens personlige relationer, kan tilsvarende forstærkes af den store mængde data, som kan indsamles, samt arbejdets identitetsbærende betydning.

5.6. Hvilken rolle spiller samtykke til indsamling af medarbejderdata?

Vi har i de første afsnit i dette kapitel præsenteret forskellige måder at forstå privatliv og retten til privatliv på. Vi har imidlertid ikke berørt et centralt forhold, som mange intuitivt vil mene kan spille en vigtig rolle for, hvornår det er moralsk problematisk at reducere privatliv: samtykke.

Samtykke er vigtigt, fordi det i mange situationer synes at spille en etisk rolle, om en person har samtykket til at dele personlig information. Eksempelvis er det nok de færreste som vil mene, at man krænker privatlivet, ved at lægge øre til en nær vens betroelser, mens det ville være en grov krænkelse af privatlivet, hvis man fik adgang til den samme information, ved at smuglæse sin vens dagbog. Man kan tilsvarende forestille sig, at samtykke i arbejdssammenhæng kan have den effekt, at indsamling eller deling af medarbejderdata, som ville være moralsk problematisk uden samtykke, kan blive etisk tilladeligt, eller i hvert fald moralsk set mindre problematisk, hvis medarbejdere samtykker til indsamlingen af deres data.[83]

83 Det er vigtigt, også i denne forbindelse, at holde sig forholdet mellem etik og jura for øje. Spørgsmålet er ikke her om samtykke gør en juridisk forskel – det gør det i nogle situationer, om end det måske spiller en begrænset rolle for hvornår arbejdspladser lovligt kan indsamle medarbejderdata (se afsnit 4.2.3). Spørgsmålet er heller ikke om samtykke *bør* være et juridisk krav for lovlig indsamling af data. Spørgsmålet er i udgangspunktet det rent etiske, om og i givet fald hvordan samtykke spiller en rolle for, hvornår indsamling af medarbejderdata er moralsk problematisk (se også afsnit 4.4 om etik og jura).

Der kan være flere forklaringer på, at samtykke spiller en sådan rolle.[84] Én forklaring kan være, at det i almindelighed har gavnlige konsekvenser, når vi gør samtykke til et krav for at udføre visse handlinger. En tidlig variant af dette synspunkt blev forsvaret af den britiske oplysningsfilosof John Stuart Mill.[85] En anden forklaring, som typisk optræder i såkaldt deontologiske moralteorier, kan hævde, at kravet om samtykke beskytter eller respekterer personers autonomi.

Hvis samtykke af en af disse grunde spiller en etisk rolle, så er det indlysende relevant for dataindsamling på arbejdspladsen. En arbejdsplads, som gerne vil indsamle data om medarbejderne, kan i så fald forsøge at undgå eller begrænse krænkelser af privatlivet, ved at bede medarbejdere om at samtykke til indsamlingen af deres data. Kan de udfordringer, som retten til privatliv kunne rejse, løses ved ganske enkelt at indhente samtykke fra medarbejdere til dataindsamlingen?

Mange vil nok umiddelbart nære en vis skepsis overfor ideen om, at man kan løse de etiske udfordringer med privatliv, som indsamling af medarbejderdata rejser, blot ved at indhente samtykke. En sådan skepsis kan begrundes, hvis man kan pege på, at samtykket ikke lever op til en eller flere af de betingelser, som et samtykke skal indfri for at være moralsk relevant.

Den måske mest kendte betingelse er, at den som samtykker skal være en autonom person. Det vil sige, at personen skal være i stand til at træffe selvstændige, reflekterede og ansvarlige beslutninger. Af samme grund antages det i mange sammenhænge, at især mindre børn ikke er i stand til at afgive et relevant samtykke. Når det gælder indsamling af data, vil mange også være tilbøjelige til at sige, at selv hvis et lille barn samtykker, for eksempel i den forstand at det trykker på "accepter" i en app, så ændrer dette samtykke ikke på, om det er etisk eller uetisk at indsamle data, fordi barnet ikke kan give et moralsk relevant samtykke.[86] Medarbejdere på en arbejdsplads vil imidlertid typisk være autonome, og vil derfor kunne indfri den første betingelse. Men moralsk relevant samtykke antages ofte at stille to andre betingelser, som det kan være sværere at leve op til. Et moralsk relevant samtykke, vil nogle hævde, skal være både *informeret* og *frit*.

5.6.1. Informeret samtykke

Udover at den, som samtykker, skal være en autonom person, så vil mange mene, at et moralsk relevant samtykke skal være *informeret*. Betingelsen begrundes ofte ved at pege på, at et samtykke ikke kan være udtryk for den samtykkendes egentlige ønske, hvis vedkommende er fejlinformeret om eller mangler relevant information.

---

84 Spørgsmålet om det etiske grundlag for samtykke er komplekst og meget omfattende behandlet i forskningslitteraturen. For et overblik, Se Nir Eyal, "Informed Consent" in *The Stanford Encyclopedia of Philosophy,* ed. Edward N. Zalta (2019). https://plato.stanford.edu/archives/spr2019/entries/informed-consent/.

85 John Stuart Mill, "On Liberty" in *On Liberty and other writings*, ed. Stefan Collini (Cambridge: Cambridge University Press, 2009).

86 Nogle gange formuleres denne pointe således, at en person slet ikke samtykker, hvis vedkommende ikke lever op til de relevante betingelser. I givet fald vil man sige, at for eksempel et barn gør noget, som minder om at samtykke, men ikke at det samtykker. Vi skelner i stedet for mellem at samtykke, i den brede betydning af at give udtryk for accept, og det at afgive et *moralsk relevant* samtykke.

Betingelsen kendes især fra den medicinske etik, hvor spørgsmål om informeret samtykke til behandling spiller en central rolle. I den medicinske etik antages det således typisk, at hvis en patient skal kunne give moralsk relevant samtykke til en behandling, så skal patienten vide hvad behandlingen indebærer, hvad formålet med behandlingen er, og hvad det forventede resultat af behandlingen er, herunder hvilke risici og bieffekter behandlingen medfører.

Tilsvarende kan man mene, at medarbejdere skal være informeret om dataindsamlingen, formålet, og de forventede konsekvenser, for at kunne afgive et moralsk relevant samtykke. Hvis en arbejdsplads efter at have indhentet samtykke indsamler flere eller andre medarbejderdata, end medarbejderen har forstået, så har medarbejderen ikke samtykket til denne dataindsamling. Og hvis arbejdspladsen anvender disse data på en anden måde eller til andre formål, end medarbejderen har forstået, så har medarbejderen ikke fået mulighed for at overskue konsekvenserne af at samtykke, og har derfor ikke haft muligheden for at tage kvalificeret stilling til dataindsamlingen. I den situation kan det forekomme tvivlsomt, at samtykket skulle kunne ændre på dataindsamlingens moralske karakter.

Betingelsen om at et samtykke skal være informeret kan rejse flere forskellige udfordringer for en arbejdsplads, som ønsker at indsamle medarbejderdata. Konkret kan man pege på udfordringer med

- at sikre, at medarbejdere har forstået den relevante information
- at noget information i praksis eller principielt er uforståelig, samt
- at der kan være relevant information, som der er stærke grunde til *ikke* at kommunikere til medarbejdere.

Den første udfordring er, at det kan være vanskeligt og ressourcekrævende, at formidle kompleks information på en måde, som er forståelig for medarbejderen, og at sikre sig, at medarbejdere har *forstået* informationen. Det er i den forbindelse vigtigt at holde sig for øje, at et samtykke ikke er informeret alene fordi medarbejderen er blevet præsenteret for relevant information. Samtykket er først informeret, når medarbejderen har forstået den pågældende information. Det betyder, at et samtykke næppe kan antages at være informeret i eksempelvis den situation, hvor en arbejdsplads i forbindelse med indhentning af samtykke blot henviser til generel information, som mange medarbejdere vil have svært ved at forstå, og selv må afsætte tid til at sætte sig ind i. På den anden side kan det forekomme rimeligt, at der er grænser for hvor langt kravet om at informere strækker sig. Eksempelvis kan man forestille sig en situation, hvor nogle medarbejdere ikke forstår information om dataindsamling, fordi de ikke gør en indsats for at forstå den, selvom arbejdspladsen har afsat medarbejdertimer til at modtage den, og ressourcer til at præsentere den relevante information på en letforståelig måde. I en sådan situation vil nogle måske være tilbøjelige til at mene, at arbejdspladsen med god ret kan gå ud fra, at medarbejderne har afgivet informeret samtykke, og at det er medarbejderens ansvar, at samtykket reelt ikke er informeret.

En anden udfordring kan være, at noget information er så teknisk kompleks, at det i praksis ikke er muligt at forklare den på en måde, så medarbejdere kan forstå den. Det kan være tilfældet både for information om de tekniske systemer som indsamler data, og de tekniske systemer som behandler den indsamlede data, for eksempel for at levere automatiseret beslutningsstøtte. Sådanne tekniske systemer kan være vanskelige eller umulige at forstå for lægmand undtagen på et meget generelt niveau, og i nogle tilfælde er visse former for

kunstig intelligens så kompleks, at det også for eksperter er umuligt at forstå i detaljer, hvordan systemet fungerer. Dette fænomen kaldes ofte for udfordringen med algoritmisk transparens eller "black-box"-problematikken.[87] Hvis der er tilfælde, hvor relevant information er uforståeligt kompleks, kunne man hævde, at det ikke er muligt for arbejdspladsen at indhente informeret samtykke fra medarbejdere.

I lyset af de udfordringer, som der kan være ved at informere medarbejdere, er det oplagt at spørge *hvor meget* og *hvilken* information en medarbejder skal have, for at kunne samtykke til dataindsamling? Hvis man ikke kan svare på disse spørgsmål, så kan man ikke vurdere, om et samtykke er informeret eller ej.

Spørgsmålene om hvor meget og hvilken information en person skal have, for at kunne give informeret samtykke, er især behandlet i medicinsk etik. Et princip som i lidt forskellige varianter forsvares i den nyere litteratur er, at en person skal have al relevant information, og at information er relevant, hvis den vil have eller burde have betydning for patientens (eller medarbejderens) beslutning om at samtykke.[88] Ideen er altså, at informationer, som vil give en person mere eller mindre lyst til at samtykke, er relevant. Men også at information kan være relevant, hvis det er information som en person rationelt eller etisk set burde lægge vægt på, selv hvis vedkommende faktisk ignorerer den. I praksis vil det nok ofte være vanskeligt, at vurdere hvornår information har denne karakter, og der vil med stor sandsynlighed være tilfælde hvor selv ledere og medarbejdere, som er enige om principperne, er uenige om, hvilken information som er relevant.

Den tredje udfordring er, at der kan være information, som arbejdspladsen eller andre har stærke grunde til at tilbageholde. Det kan eksempelvis være tilfældet, når relevant information om et system til at indsamle eller anvende data udgør en forretningshemmelighed. I den situation vil udvikleren af systemet have stærke grunde til at modsætte sig, at medarbejdere (eller andre) informeres om hvordan systemet virker. Det kan også være tilfældet når deling af information vil underminere formålet med at indsamle data. Hvis eksempelvis en arbejdsplads indsamler medarbejderdata for at evaluere medarbejderes præstationer, så kan detaljeret information om hvordan data indsamles og behandles gøre det muligt for medarbejdere, at ændre deres adfærd på en måde, så præstationsmålingen bliver uprøcis (se også afsnit 6.2 om gaming-effekter).

I de tilfælde hvor der er grunde til at tilbageholde relevant information opstår en konflikt mellem behovet for at give medarbejdere den information, som er nødvendig, for at de kan afgive informeret samtykke, og behovet for at tilbageholde information, som eksempelvis vil afsløre forretningshemmeligheder eller underminere formålet med dataindsamlingen. Ét muligt synspunkt i den situation er, at informeret samtykke er strengt nødvendig for etisk indsamling af medarbejderdata. Dette synspunkt medfører, at arbejdspladsen bør vælge

87 En omfattende men lettilgængelig diskussion af problemet findes i Christoph Molnar, *Interpretable machine learning. A guide for making black box models explainable* (2019). https://christophm.github.io/interpretable-ml-book/

88 Se Joseph Millum and Danielle Bromwich, "Informed Consent: What Must Be Disclosed and What Must Be Understood?" *American Journal of Bioethics* 21, no. 5 (2021); Tom Walker, "Informed Consent and the Requirement to Ensure Understanding" *Journal of Applied Philosophy* 29, no. 1 (2011); Arnon Keren and Ori Lev, "Informed Consent, Error and Suspending Ignorance: Providing Knowledge or Preventing Error?" *Ethical Theory and Moral Practice* 25, no. 2 (2022).

enten at lade være med at indsamle data, eller at afgive den følsomme information, som er nødvendig for informeret samtykke. Et andet muligt synspunkt er, at det i sådanne situationer er nødvendigt at afveje de modstridende hensyn mod hinanden, og at det i hvert fald i princippet er muligt, at afvejningen kan falde sådan ud, at dataindsamling er moralsk tilladelig, selvom medarbejderen ikke har givet moralsk relevant samtykke. Hvilket af sådanne synspunkter man bør anlægge afhænger i hvert fald til dels af, hvad man mener at det etiske grundlag er, for kravet om informeret samtykke.

5.6.2. Frit samtykke

Den sidste almindelige betingelse for et moralsk relevant samtykke er, at samtykket er frit. Det betyder, at samtykket er udtryk for den samtykkende persons vilje, uden at denne vilje har været udsat for forstyrrende indflydelse.

Betingelsen har spillet en stor rolle i diskussioner om henholdsvis seksuelt samtykke og udnyttelse (eng. "exploitation"). I begge sammenhænge er den centrale pointe, at et samtykke kan diskvalificeres, hvis en person giver udtryk for samtykke under pres fra konsekvenserne ved at lade være. Med et klassisk eksempel, så vil de færreste sige, at en person i nogen moralsk relevant forstand samtykker, hvis vedkommende giver sine ejendele til en røver, der snerrer "Pengene eller livet!".[89] Det skyldes ikke, at man ikke kan samtykke til at give andre sine ejendele. Det kunne en person gøre i andre situationer, eksempelvis af medlidenhed i mødet med en tigger. Det skyldes heller ikke, at offeret ikke har et valg. I en vis forstand vælger offeret at give røveren sine ejendele, idet offeret har et alternativ i form af valget om at lade livet. Når truslen intuitivt diskvalificerer samtykket er den mest indlysende forklaring, at offerets vilje bliver underlagt en så forstyrrende indflydelse, i form af det massive pres som truslen udgør, at man ikke meningsfuldt kan sige, at beslutningen er udtryk for offerets frie vilje.

Kritikere har i forbindelse med samtykke til indsamling af medarbejderdata påpeget, at det kan være vanskeligt for medarbejdere at afgive frit samtykke, hvis konsekvenserne af at nægte samtykke er eller risikerer at være alvorlige for medarbejderen. (se afsnit 4.2.3 om muligheden for juridisk gyldigt samtykke). Eksempelvis kan en ansøger til en stilling blive oplyst om, at arbejdspladsen indsamler bestemte medarbejderdata, og anmodet om at samtykke til indsamlingen. Hvis dette samtykke er en forudsætning for ansættelse, så har nægtelse af samtykke potentielt store konsekvenser for ansøgeren. I den situation kan man diskutere, om personer er i stand til at afgive frit samtykke.

Hvis et samtykke skal være frit, for at være moralsk relevant, så er det oplagt at spørge hvad der skal til, for at et samtykke er frit. Det viser sig at være et ganske vanskeligt spørgsmål, at besvare. De fleste vil være enige om, at der ikke er tale om frit samtykke i visse situationer, som i eksemplet med røveren. Men det er langt mindre klart, hvor grænsen går for, at et valg er underlagt tilstrækkelig meget forstyrrende indflydelse, til at et samtykke bliver ufrit. Hvis eksempelvis en arbejdsplads tilbyder medarbejdere en bonus, hvis de samtykker til indsamling

89 A. Wertheimer, *Coercion* (Princeton University Press, 2014).

af medarbejderdata, men i øvrigt lader det være valgfrit, så giver arbejdspladsen medarbejdere en tilskyndelse til at samtykke, som kan øve indflydelse på medarbejderens vilje. Mange vil ikke desto mindre nok vurdere, at medarbejdere i den situation kan give frit samtykke til indsamling af data. Men hvad så hvis arbejdspladsen gør adgang til lønforhøjelser og forfremmelse afhængig af, at medarbejdere samtykker til indsamling af medarbejderdata? Eller lader forstå, at selvom manglende samtykke ikke i sig selv er grundlag for afskedigelse, så vil villighed til at samtykke til indsamling af medarbejderdata være et kriterium, som vil indgå i overvejelserne, hvis arbejdspladsen på et tidspunkt skal gennemføre en fyringsrunde? På et tidspunkt bliver omkostningen ved manglende samtykke antageligt så stor, at vi ligesom i eksemplet med røveren vil være tilbøjelige til at mene, at samtykket ikke længere er frit.

6. Etiske udfordringer ved anvendelse af automatiserede beslutningssystemer på arbejdspladsen – fejlvurderinger, bias og adfærd

I det forrige kapitel har vi kigget på, hvilke etiske udfordringer det kan rejse, når en arbejdsplads indsamler medarbejderdata. Når en arbejdsplads har indsamlet medarbejderdata, kan arbejdspladsen ønske at bruge disse data på mange måder. Ofte vil arbejdspladsen ønske at bruge data til at hjælpe med at træffe beslutninger. Arbejdspladser har i årtier anvendt data til at udarbejde analyser, som kunne informere ledelsen. Inden for det seneste årti er det imidlertid blevet stadig mere almindeligt, at arbejdspladser bruger forskellige former for kunstig intelligens til at analysere medarbejderdata, og levere automatiseret beslutningsstøtte eller fuldt automatiserede beslutninger (se afsnit 3.4).

Udbredelsen af automatiserede beslutningssystemer på arbejdspladsen har ikke være ukontroversiel. Teknologien har været genstand for intens kritisk diskussion, idet blandt andet forskere, borgerrettighedsorganisationer, og faglige bevægelser har givet udtryk for bekymringer.[90] Dette kapitel giver et overblik over en gruppe af centrale dataetiske udfordringer ved automatiserede beslutningssystemer på arbejdspladsen som knytter sig til fejlagtige beslutninger, bias og adfærdseffekter.

Indledningsvis præsenterer kapitlet den helt elementære udfordring, at automatiserede beslutningssystemer kan begå fejl. Denne pointe er udgangspunktet for flere af de andre udfordringer. I den forbindelse diskuterer vi også risikoen for, at automatiserede beslutningssystemer kan have negativ effekt på medarbejdertilfredshed og -produktivitet, samt risikoen for såkaldte "gaming"-effekter, hvor medarbejdere reagerer på indsamling og anvendelse af data på en måde, som underminerer arbejdspladsens grundlæggende mål. Efterfølgende præsenterer kapitlet den måske mest diskuterede udfordring for automatiserede beslutningssystemer, som består i at disse systemer kan have algoritmisk bias. Vi diskuterer i den forbindelse både hvordan man kan definere algoritmisk bias på flere forskellige måder, hvordan algoritmisk bias kan opstå, og hvad der kan gøre algoritmisk bias etisk problematisk. Afslutningsvis diskuterer kapitlet muligheder og udfordringer ved at anvende automatiseret beslutningsstøtte, hvor der er "et menneske i kredsløbet". Det peger i den forbindelse på

90 Se eksempelvis: Merve Hickok and Nestor Maslej, "A policy primer and roadmap on AI worker surveillance and productivity scoring tools" *AI and Ethics* 3, no. 3 (2023), https://doi.org/10.1007/s43681-023-00275-8, https://doi.org/10.1007/s43681-023-00275-8; Giacosa et al., "Stress-inducing or performance-enhancing? Safety measure or cause of mistrust? The paradox of digital surveillance in the workplace."; AI Now Institute, *Algorithmic Management: Restraining Workplace Surveillance*; Adams-Prassl, "Regulating algorithms at work: Lessons for a 'European approach to artificial intelligence'."; Megan Fritts and Frank Cabrera, "AI recruitment algorithms and the dehumanization problem" *Ethics and Information Technology* 23, no. 4 (2021), https://doi.org/10.1007/s10676-021-09615-w, https://doi.org/10.1007/s10676-021-09615-w; Ulrich Leicht-Deobald et al., "The Challenges of Algorithm-Based HR Decision-Making for Personal Integrity" *Journal of Business Ethics* 160, no. 2 (2019), https://doi.org/10.1007/s10551-019-04204-w, https://doi.org/10.1007/s10551-019-04204-w; Ifeoma Ajunwa, Kate Crawford, and Jason Schultz, "Limitless worker surveillance" *California Law Review* (2017).

risikoen for støj i menneskelige beslutninger og en række særlige bias, som kan optræde, blandt andet automatiseringsbias og algoritmisk aversionsbias.

6.1. Beslutningssystemer begår fejl

Den amerikanske forfatter Cathy O'Neil indleder sin indflydelsesrige bog *"Weapons of Math Destruction"* med historien om læreren Sarah Wysocki, som i 2011 blev fyret fra sit arbejde ved en skole i Washington D.C., USA.[91] Byen havde indført et nyt system, IMPACT, til evaluering af lærere i det offentlige skolevæsen. Bystyret havde også truffet en beslutning om hvert år at afskedige de lærere, som scorede lavest. Wysocki var nyuddannet lærer, og fik fremragende evalueringer af sin skoleleder og forældrene til de børn, hun underviste. Men ved slutningen af skoleåret fik hun så ringe en score af IMPACT, at hun havnede blandt de 5% af lærerne, som scorede dårligst. Derfor blev hun ligesom godt 200 af sine kollegaer fyret.

Wysockis historie kan give anledning til kritik på flere måder. Mange vil nok blive provokeret af ideen om en politik, hvor en arbejdsplads hvert år automatisk afskediger de medarbejdere, som klarer sig dårligst, selv hvis arbejdspladsen kunne lave præcise evalueringer af medarbejderne. Kritikere kunne måske pege på, at det kunne være, at selv de fem procent af medarbejderne, som præsterede relativt dårligst, gjorde et fremragende stykke arbejde. Det kunne også være, at nogle medarbejdere underpræsterede fordi de i en periode kæmpede med personlige udfordringer, som arbejdspladsen burde tage hensyn til, eller på grund af dårlig ledelse eller ringe arbejdsforhold. Men et oplagt kritikpunkt i Wysockis tilfælde var også, at evalueringen netop ikke var præcis. Wysocki var efter alt at dømme en dygtig og samvittighedsfuld lærer, som klarede sig godt. Alligevel blev hun af et automatiseret beslutningsstøttesystem vurderet som en af de allerdårligst præsterende medarbejdere. Hvordan kunne systemet begå en sådan fejl? Hvorfor laver automatiserede beslutningssystemer i det hele taget fejl?

Et automatiseret beslutningssystem virker typisk ved, at systemet forsøger at vurdere en målegenskab.[92] Systemets vurdering er baseret på en matematisk model for, hvordan andre egenskaber eller værdier statistisk set påvirker målegenskaben. Eksempelvis kan et system forsøge at vurdere, hvilken medarbejder som bedst vil løse en opgav, ved at kigge på hvilke egenskaber medarbejderne har, og hvordan disse egenskaber statistisk set kvalificerer en medarbejder til at løse opgaven.

I nogle situationer er det muligt for et system at lave vurderinger med meget stor præcision, men i realistiske situationer vil alle beslutningssystemer begå fejl fra tid til anden. Visse systemer kan have en meget lav fejlrate, mens andre vil ramme ved siden af temmelig ofte. Hvor høj eller lav præcision et system har afhænger af flere forskellige faktorer.

---

91 Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Crown/Archetype, 2016)

92 Man skelner teknisk mellem klassificeringssystemer og regressionssystemer. Et system til klassificering vurderer om en medarbejder har en målegenskab, for eksempel "Er medarbejderen i risiko for stress?". Et system til regression vurderer hvad værdien er for en målegenskab, for eksempel "Hvordan præsterer medarbejderen på en skala fra 1 til 10?" Vi fokuserer i diskussionen nedenfor på klassificering, da det er lettest at illustrere de dataetiske udfordringer med eksempler på sådanne systemer.

Systemets præcision afhænger først og fremmest af mængden og kvaliteten af de træningsdata, som læringsalgoritmen træner systemet på. Jo mere relevant data, jo bedre muligheder har læringsalgoritmen for at finde statistiske sammenhænge og lave en præcis model af dem. Omvendt medfører lav kvalitet i træningsdata, at det bliver vanskeligt for læringsalgoritmen at træne en præcis model.

Præcision afhænger også af hvilken type model systemet anvender. Forskellige læringsalgoritmer træner systemer med forskellige typer modeller. En bestemt type model kan være mere eller mindre egnet til en bestemt type vurderinger – en model kan være god til at repræsentere sammenhængene i én situation, men ikke så god i andre situationer. Eksempelvis er nogle modeller enkle, mens andre er mere komplekse. Enkle modeller er typisk lette at træne, og det er let at forstå hvordan modellen virker, men enkle modeller kan have vanskeligt ved at repræsentere komplekse sammenhænge.

Endelig afhænger præcision også af hvad det er, som systemet forsøger at vurdere – nogle vurderinger er ganske enkelt mere vanskelige end andre. Sammenhængene i systemets model udtrykker statistiske sandsynligheder. I realistiske situationer er målegenskaben eller målværdien kun i begrænset omfang bestemt af de statistiske sammenhænge, som det kan lade sig gøre at modellere. Der er så at sige et element af uundgåelig uforudsigelighed i automatiserede beslutningssystemers vurderinger. Det kan eksempelvis betyde, at to medarbejdere, som er identiske i alle de henseender systemet kan modellere, alligevel viser sig at være meget forskellige – den ene præsterer eksempelvis bedre end den anden, uden at man ved at kigge på data kan forklare hvorfor. Nogle slags vurderinger kan være meget uforudsigelige, mens andre slags vurderinger vil være mindre uforudsigelige. Systemer der foretager vurderinger, der i høj grad er uforudsigelige, har naturligt tendens til højere fejlrate, end systemer der foretager vurderinger af forhold, som er lette at forudsige.

6.1.1. Hvad gik der galt for Sarah Wysocki?

Hvorfor begik systemet en fejl i den historie om Sarah Wysocki, som vi indledte dette kapitel med at præsentere? O'Neil præsenterer to grunde, som kan forklare resultatet.[93] IMPACT evaluerede lærere ved at sammenligne elevernes testresultater fra slutningen af ét skoleår med resultaterne ved slutningen af det næste skoleår. Ved at korrigere for visse sociodemografiske faktorer kunne systemet vurdere, hvordan de pågældende elever burde udvikle sig, hvis de modtog undervisning af en gennemsnitlig lærer, og sammenligne dette med elevernes faktiske udvikling. Hvis eleverne klarede sig bedre end forventet, vurderede systemet derfor, at læreren præsterede bedre end den gennemsnitlige lærer. Hvis eleverne omvendt klarede sig dårligere end forventet, vurderede systemet, at læreren præsterede ringere end gennemsnittet.

O'Neil peger på, at systemets vurdering af den enkelte lærers præstation for det første var baseret på et meget lille datagrundlag. Hver lærer blev vurderet på baggrund af resultater fra sine elever, men hver elevs resultater vil forventeligt variere på grund af utallige faktorer, som systemets model ikke kunne repræsentere. En elev, hvis forældre er blevet skilt, kan have

93 O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*.

svært ved at koncentrere sig om skolen. En anden elev, som er begyndt at få privat lektiehjælp, kan pludselig klare sig bedre. Hvis hver lærer underviste hundredevis eller tusindvis af elever hvert år, så ville sådanne tilfældigheder udjævnes statistisk. Men fordi hver lærer kun underviste en lille gruppe elever, kunne sådanne tilfældigheder få stor indflydelse på, hvor godt lærerens elevgruppe klarede sig. Elevernes præstationer afhang af en masse faktorer, som ikke kunne måles, og vurdering af den enkelte lærers præstation blev derfor upålidelig.

Den anden grund, som O'Neil peger på, er at Wysockis femte-klasses elever kom fra fjerde-klasses indskoling med andre lærere. De lærere, som stod for undervisning af eleverne i fjerde klasse, var underlagt stærke incitamenter til at producere gode testresultater til afgangsprøven ved slutningen af fjerde klasse. Dårlige resultater ved denne prøve kunne, ligesom det skete for Wysocki, føre til at den pågældende lærer blev fyret. Samtidigt havde Washington D.C. indført høje kontantbonusser til lærere, som producerede gode testresultater. Intentionen med incitamenterne var selvsagt at motivere lærerne til at levere god undervisning. Men i praksis var konsekvensen med stor sandsynlighed også, at nogle lærere manipulerede med afgangsprøven, for at forbedre deres elevers resultater (se afsnit 6.2 om "gaming"-effekter). Wysocki modtog en stor gruppe elever, som havde scoret højt ved deres afsluttende prøver i fjerde-klasse, men som viste sig at halte langt bagefter i undervisningen. Ved slutningen af femte klasse scorede disse elever langt under hvad IMPACT forventede på baggrund af deres resultater fra fjerde klasse, og Wysocki kom til at ligne en lærer, som præsterede dårligt. Systemet blev øjensynligt fodret med fejlbehæftede data, fordi evalueringerne skabte skæve incitamenter.

6.1.2. Hvad er det etiske problem ved fejl i beslutningssystemer?

Hvornår udgør det et etisk problem, hvis et beslutningssystem begår fejl? Og hvor stort er problemet? Svarene på disse spørgsmål er mere komplekse, end man måske umiddelbart kunne tro, blandt andet fordi de afhænger af mindst tre forhold:

- Hvilke typer beslutninger systemet behandler
- Hvordan systemets vurdering anvendes
- Hvordan man etisk vurderer konsekvenserne af fejl

Først og fremmest afhænger problemets omfang af den type beslutning, som systemet støtter eller udfører. Automatiserede beslutningssystemer anvendes i dag i mange forskellige sammenhænge. En af udfordringerne for udvikling og implementering af automatiserede beslutningssystemer er, at det i nogle af disse sammenhænge er ekstremt vigtigt at undgå fejl. For et system, som diagnosticerer patienter eller styrer en bil, kan selv relativt små fejl være livsfarlige. Men et beslutningssystems fejl kan også have trivielle konsekvenser. Når der eksempelvis slipper en spam-mail igennem filtret på en e-mail-konto, så medfører det typisk blot at en person er nødt til manuelt at slette den. Når et beslutningssystem anvendes til at støtte eller erstatte menneskelige ledelsesbeslutninger på en arbejdsplads, så spiller det tilsvarende en afgørende rolle, hvilken beslutning der er tale om. En så afgørende beslutning som afskedigelse af en medarbejder, som i Sarah Wysockis tilfælde, har indlysende større konsekvenser end eksempelvis en beslutning om hvilken af to medarbejdere, som skal tildeles en rutineopgave.

For det andet afhænger problemets omfang af, hvilken rolle systemets vurdering spiller. Ved en fuldt automatiseret beslutning har systemets fejl afgørende betydning. Hvis arbejdspladsen anvender automatiseret beslutningsstøtte, kan fejlen have mindre eller ingen betydning, fordi et menneske har mulighed for at underkende systemets (fejl)vurdering, og træffe den korrekte beslutning (se afsnit 6.6 om et menneske i kredsløbet).

For det tredje, og mest afgørende, så afhænger problemets omfang af, hvordan man etisk vurderer den pågældende konsekvens. Forskellige typer beslutninger kan have etisk set forskellige konsekvenser, og de moralske forpligtelser til at undgå eller skabe visse konsekvenser kan variere fra kontekst til kontekst.

Hvis eksempelvis en arbejdsplads anvender et system til at identificere de bedst kvalificerede ansøgere til en stilling, så kan systemet begå en fejl ved ikke at prioritere en af de bedst kvalificerede ansøgere. Derved går arbejdspladsen potentielt glip af en dygtig medarbejder, og ansøgeren går glip af muligheden for at få stillingen. For arbejdspladsen kan fejlen være omkostningsfuld, hvis den fører til at man ansætter en medarbejder som præsterer dårligere. Men fejlen kan også være triviel, hvis arbejdspladsen ansætter en anden medarbejder, som præsterer på samme niveau som (eller bedre end) den frasorterede ansøger ville have præsteret. For ansøgeren kan fejlen på lignende vis være omkostningsfuld, hvis ansættelse ville være en væsentlig forbedring, og triviel, hvis eksempelvis vedkommende umiddelbart efter ansættes i et andet job, som er lige så godt. Samtidig har fejlen den effekt, at den begunstiger andre ansøgere. Hvis fejlen fører til at ansøgeren mister stillingen, så fører den også til at en anden ansøger *får* stillingen. Den omkostning, som fejlen påfører den ene ansøger medfører således samtidig at en anden ansøger tilsvarende gavnes.

For endeligt at vurdere de etiske konsekvenser af en fejl i prioritering af ansøgere, er man således nødt til også at spørge, hvilke etiske forpligtelser en arbejdsplads har til at behandle ansøgere til en stilling på bestemte måder. Findes der eksempelvis en meritokratisk pligt at prioritere ansøgere efter kvalifikationer, og ansætte den bedst kvalificerede? Ville det være etisk problematisk, hvis en arbejdsplads valgte at besætte en opslået stilling, ved at trække lod mellem de indkomne ansøgninger?

Spørgsmålet, om en arbejdsplads er forpligtet til at være meritokratisk i forbindelse med ansættelse, viser sig i forskningslitteraturen at være både komplekst og kontroversielt.[94] I denne sammenhæng er det tilstrækkeligt at pege på den type overvejelser, som det kan kræve at tage stilling til, hvor stort et etisk problem det udgør, når et beslutningssystem begår en fejl. Fordi beslutningssystemer kan anvendes i forbindelse med mange forskellige typer beslutninger, må de relevante overvejelser gøres i tilknytning til det konkrete beslutningssystem og de specifikke beslutninger, som det støtter eller udfører.

Anvendelse af automatiserede beslutningssystemer på arbejdspladsen har potentiale til at effektivisere og optimere mange processer. Ideelt fører anvendelsen af sådanne systemer til både hurtigere, billigere og bedre beslutninger. Men alle beslutningssystemer kan begå fejl, og sådanne fejl kan være etisk problematiske. Et centralt element i en dataetisk evaluering af et

94 Shlomi Segall, "Should the Best Qualified Be Appointed?*" Journal of Moral *Philosophy* 9, no. 1 (2012), https://doi.org/https://doi.org/10.1163/174552411X592149, https://brill.com/view/journals/jmp/9/1/article-p31_4.xml.

automatiseret beslutningssystem på en arbejdsplads er derfor analyse af systemets præcision, af hvilke typer fejl, som systemet kan begå, samt af hvordan disse fejl etisk set må vurderes.

6.2. Medarbejdertilfredshed, produktivitet og gaming-effekter

Når en arbejdsplads introducerer et automatiseret beslutningssystem, er formålet typisk at forbedre arbejdsgange og resultater. Som vi har set ovenfor, kan automatiserede beslutningssystemer begå fejl. Sådanne fejl kan dels begrænse systemets evne til at realisere de potentielle fordele, og dels påføre medarbejdere, arbejdspladsen som hele, eller andre interessenter ulemper. Udover risikoen for fejl kan anvendelse af et automatiseret beslutningssystem have en række uønskede bivirkninger. En første sådan effekt er, at indsamling og anvendelse af medarbejderdata kan føre til negative reaktioner blandt medarbejdere, eksempelvis i form af tab af medarbejdertilfredshed og deraf faldende produktivitet. En anden mulig effekt er, at indsamling og anvendelse af medarbejderdata kan skabe såkaldt "gaming", hvor medarbejdere tilpasser deres adfærd til de datapunkter som måles, snarere end de mål, som arbejdspladsen ultimativt har. I dette afsnit skitserer vi disse to risici.

Effekten af at indsamle og anvende medarbejderdata, især i forbindelse med præstationsmålinger, har været et vigtigt forskningsfelt i årtier. Interessen for dette felt synes kun at være intensiveret i takt med, at arbejdspladser har fået adgang til øgede muligheder for dataindsamling og nye former for automatiserede beslutningssystemer. En almindelig hypotese har været, at indsamling og anvendelse af medarbejderdata, udover de fordele som anvendelsen af data ville medføre, også kunne skabe ønskværdige reaktioner, eksempelvis ved at afskrække uønsket adfærd. Sat på spidsen kunne man forestille sig, at en fordel ved øget indsamling og anvendelse af medarbejderdata ville være, at medarbejdere i mindre grad dovnede, sjuskede eller endda svindlede på arbejdet. Derved ville indsamling og anvendelse af medarbejderdata af sig selv hjælpe med at forbedre medarbejderes præstation og arbejdspladsens effektivitet.

Visse nyere studier rejser imidlertid tvivl ved, om indsamling af medarbejderdata har denne effekt. I nogle studier fører elektronisk præstationsmåling ligefrem til et fald i medarbejders præstationer.[95] Et metastudie af elektronisk præstationsmåling, som omfatter 94 studier med sammenlagt ca.23.000 medarbejdere, når frem til, at studiernes resultater peger i forskellige retninger, men ikke samlet set giver evidens for, at elektronisk præstationsmåling har en positiv effekt på medarbejderes præstation.[96] Der kan være en række grunde til, at effekten på præstation er nul eller ligefrem negativ, herunder tab af motivation og øget stress (se afsnit 5.4.2

95 Se eksempelvis Allison Brown Yost et al., "Reactance to Electronic Surveillance: a Test of Antecedents and Outcomes" *Journal of Business and Psychology* 34, no. 1 (2019), https://doi.org/10.1007/s10869-018-9532-2, https://doi.org/10.1007/s10869-018-9532-2; Angela J. Martin, Jackie M. Wellen, and Martin R. Grimmer, "An eye on your work: How empowerment affects the relationship between electronic surveillance and counterproductive work behaviours" *The International Journal of Human Resource Management* 27, no. 21 (2016), https://doi.org/10.1080/09585192.2016.1225313.

96 Ravid et al., "A meta-analysis of the effects of electronic performance monitoring on work outcomes."; Thomas Kalischko and René Riedl, "Electronic performance monitoring in the digital workplace: conceptualization, review of effects and moderators, and future research opportunities" *Frontiers in psychology* 12 (2021)

om observationsstress). En yderligere forklaring kan være, at medarbejdere strategisk ændrer adfærd på måder, som underminerer arbejdspladsens egentlige mål, såkaldt "gaming".

Gaming er den effekt, at indsamling og anvendelse af medarbejderdata kan skabe en bevidst eller ubevidst ændring i adfærd, som på den ene side er en rationel tilpasning til de nye incitamenter, men som på den anden side må vurderes som et tilbageskridt for arbejdspladsens egentlig mål. Det underliggende princip formuleres ofte som Goodharts lov: "Jo mere en kvantitativ social indikator anvendes til at træffe sociale beslutninger, jo mere vil den være under pres for at blive korrumperet, og jo mere tendens vil den have til at forstyrre og korrumpere den sociale proces, som det er meningen at den skal overvåge."[97]

Gaming:

En ændring i medarbejderes adfærd, som kan forklares som en rationel reaktion på indsamling og anvendelse af medarbejderdata, men som underminerer formålet med indsamling af data eller arbejdspladsens ultimative mål.

Sådanne gaming-effekter er veldokumenterede i en række sektorer, inklusive forskning, sundhed og uddannelse.[98] For en arbejdsplads, som indsamler og anvender medarbejderdata, er udfordringen, at dette kan skabe stærke incitamenter for medarbejderne til at ændre adfærd, for eksempel således at de vurderes mere positivt.[99] Sådanne adfærdsændringer kan være gavnlige, hvis de for eksempel fører til et bedre arbejdsmiljø, mere effektivt arbejde, eller arbejde af højere kvalitet. Men adfærdsændringerne kan være uhensigtsmæssige, hvis de eksempelvis skader arbejdspladsens egentlige mål.[100] Det kan ofte være tilfældet, fordi de data som indsamles sjældent direkte repræsenterer arbejdspladsens mål. Selv hvis en arbejdsplads eksempelvis forsøger at måle medarbejderes præstationer, så vil man ofte kun være i stand til at måle datapunkter, som statistisk hænger sammen med, eller antages at hænge sammen med præstation. Med et lidt simpelt eksempel, så kunne en arbejdsplads måle på hvor hurtigt medarbejdere løser opgaver, med henblik på at fremme effektivitet i opgaveløsningen, og derved utilsigtet skabe en kultur, hvor medarbejdere løser opgaver så hurtigt som muligt uden hensyn til kvaliteten. Et mere konkret eksempel findes i historien om den fyrede lærer Sarah Wysocki. En del af forklaringen på, at hun modtog elever, som var fagligt svagere

97 C. A. E. Goodhart, "Problems of Monetary Management: The UK Experience" in *Monetary Theory and Practice: The UK Experience* (London: Macmillan Education UK, 1984).

98 Se eksempelvis Peter Weingart, "Impact of bibliometrics upon the science system: Inadvertent consequences?" Scientometrics 62 (2005); Gwyn Bevan and Christopher Hood, "What's Measured is What Matters: Targets and Gaming in the English Public Health Care System" *Public Administration* 84, no. 3 (2006), https://doi.org/10.1111/j.1467-9299.2006.00600.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9299.2006.00600.x; James L. Perry, Trent A. Engbers, and So Yun Jun, "Back to the Future? Performance-Related Pay, Empirical Research, and the Perils of Persistence" *Public Administration Review* 69, no. 1 (2009), https://doi.org/10.1111/j.1540-6210.2008.01939_2.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6210.2008.01939_2.x. For et (kritisk) overblik, se Jerry Muller, *The tyranny of metrics* (Princeton University Press, 2018). Et nyligt systematisk review, som finder omfattende gaming-effekter er Monica Franco-Santos and David Otley, "Reviewing and Theorizing the Unintended Consequences of Performance Management Systems" *International Journal of Management Reviews* 20, no. 3 (2018), https://doi.org/10.1111/ijmr.12183, https://onlinelibrary.wiley.com/doi/abs/10.1111/ijmr.12183.

99 Joshua Kroll et al., "Accountable Algorithms" *University of Pennsylvania Law Review* 165, no. 3 (2017), https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3.

100 Zachary C. Lipton, The Mythos of Model Interpretability, 2017, arXiv.

end deres testresultater indikerede, kan have været at systemet skabte et stærkt incitament til at forbedre elevers præstationer i tests, snarere end de underliggende færdigheder, som kunne overføres til næste skoleår, såkaldt "teaching to the test".[101]

Det er værd at understrege, at de negative effekter, som er beskrevet i dette afsnit, er et resultat af indsamling og anvendelse af medarbejderdata, og derfor kan optræde uanset om arbejdspladsen bruger et digitalt ledelsesværktøj og automatiserede beslutningssystemer, eller en analog, menneskelig indsamling og anvendelse af medarbejderdata. Men udbredelsen af digitale ledelsesværktøjer, som kan indsamle og anvende store mængder medarbejderdata, gør det vigtigere end nogensinde, at holde sig disse risici for øje. Det er også vigtigt at være opmærksom på, at ligesom med risikoen for fejl er det et åbent og empirisk spørgsmål, om indsamling og anvendelse af medarbejderdata i en konkret sammenhæng vil have positive eller negative sideeffekter, eksempelvis i form af tab af motivation eller gaming. Givet at sådanne effekter risikerer at underminere de positive effekter af at indsamle og anvende data, er det vigtigt at tage forholdsregler for at forebygge at de indtræffer, at undersøge om de alligevel forekommer, og at inddrage dem, hvis de forekommer, i den dataetiske evaluering af et automatiseret beslutningssystem.

6.3. Algoritmisk bias i beslutningssystemer

Beslutningssystemer kan, som vi har set ovenfor, begå fejl. Vi har også diskuteret det forhold, at nogle fejl etisk set er værre end andre. En beslægtet problematik knytter sig til risikoen for, at automatiserede beslutningssystemer har algoritmisk bias. I dette afsnit introducerer vi først algoritmisk bias som overordnet begreb, og dernæst fire forskellige måder algoritmisk bias kan optræde på i et automatiseret beslutningssystem. Afslutningsvis kigger vi på de udfordringer der kan være ved, at forsøge at undgå algoritmisk bias i et automatiseret beslutningssystem.

Bias:

En tendens til systematisk afvigelse fra de forventede eller ønskede måder at virke på.

En bias betyder almindeligvis en tendens til at forstå eller vurdere noget i et særligt lys. Begrebet er typisk negativt ladet, på den måde at man forudsætter, at en forståelse eller vurdering bliver forstyrret eller misvisende under indflydelse af bias.[102] Med en mere teknisk definition, kan bias forstås som en tendens til systematiske afvigelser fra en forventet eller ønsket måde at virke på.[103] Med et simpelt eksempel, så har en almindelig terning en bias, hvis den slår 6 langt oftere

101 Louis Volante, "Teaching to the Test: What Every Educator and Policy-Maker Should Know" *Canadian Journal of Educational Administration and Policy* (2004).

102 *Den danske ordbog* præsenterer eksempelvis to nært beslægtede betydninger af bias som henholdsvis: "[En] misvisende fremstilling af undersøgelsesresultater, målelige størrelser el.lign. som især skyldes metodiske fejl eller ubevidste præferencer", og "skævhed eller misforhold der skyldes forudfattede meninger og forestillinger". *Den danske ordbog* (2023): https://ordnet.dk/ddo/ordbog?query=bias.

103 Thomas Kelly, *Bias: A Philosophical Study* (Oxford University Press, 2022).

end den slår 1, fordi vi forventer at terningen har lige stor sandsynlighed for alle resultater. En HR-medarbejder har tilsvarende en bias, hvis vedkommende systematisk lægger større vægt på mandlige ansøgeres kvalifikationer end på kvindelige ansøgeres kvalifikationer, fordi vi forventer, at kvalifikationer vurderes uafhængigt af køn.

I psykologi- og adfærdsforskningen har man de seneste årtier dokumenteret slående eksempler på såkaldte kognitive bias. En kognitiv bias kan bredt defineres som en systematisk tendens til at afvige fra rationel tænkning eller adfærd.[104] Eksempelvis er konfirmationsbias en tendens til at vurdere ny information forskelligt afhængigt af, om den støtter eller udfordrer en persons eksisterende synspunkter. Ankereffekter er det pudsige fænomen, at information kan have stor betydning for en persons vurdering af faktuelle forhold, hvis man har denne information præsent i bevidstheden, når man skal tage stilling, uanset om informationen er relevant eller ej (se også afsnit 6.6.3 om konfirmationsbias og ankereffekter ved automatiseret beslutningsstøtte). Og status quo bias er en tendens til at vurdere værdien af en mulighed forskelligt, afhængigt af om den opfattes som status quo eller en ændring fra status quo.[105]

Et andet vigtigt fokus har det seneste årti været studiet af såkaldte implicitte bias, hvor forskning har dokumenteret den forskel, som ubevidste værdiladede opfattelser af personers egenskaber kan gøre for menneskers tænkning og adfærd.[106] Et meget omtalt eksempel er testen af implicitte associationer, som blandt andet er brugt til at vise, at mange personer har sværere ved at knytte positivt ladede begreber til medlemmer af etniske minoritetsgrupper end til personer fra etniske majoritetsgrupper.

Et automatiseret beslutningssystem har ikke de kognitive bias eller implicitte bias som mennesker kan have. Ikke desto mindre har der det seneste årti har været et voksende og intenst fokus på såkaldt "algoritmisk bias" i automatiserede beslutningssystemer.[107] Ligesom bias mere generelt, kan algoritmisk

Algoritmisk bias:

Et automatiseret beslutningssystems tendens til systematiske afvigelser fra de forventede eller ønskede måder at virke på.

104 For en klassisk introduktion til kognitive bias, se Daniel Kahneman and Amos Tversky, eds., *Choices, Values, and Frames* (New York: Cambridge University Press, 2009).

105 William Samuelson and Richard Zeckhauser, "Status quo bias in decision making" journal article, *Journal of Risk and Uncertainty* 1, no. 1 (1988), https://doi.org/10.1007/bf00055564, http://dx.doi.org/10.1007/BF00055564.

106 For et overblik, Se Michael Brownstein, "Implicit Bias" in *The Stanford Encyclopedia of Philosophy,* ed. Edward Zalta (2015). http://plato.stanford.edu/archives/spr2015/entries/implicit-bias/.

107 For overblik, Se Alycia N. Carey and Xintao Wu, "The statistical fairness field guide: perspectives from social and formal sciences" *AI and Ethics* 3, no. 1 (2023), https://doi.org/10.1007/s43681-022-00183-3, https://doi.org/10.1007/s43681-022-00183-3; Shira Mitchell et al., "Algorithmic Fairness: Choices, Assumptions, and Definitions" *Annual Review of Statistics and Its Application 8*, no. 1 (2021), https://doi.org/10.1146/annurev-statistics-042720-125902, https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-042720-125902; Hoda Heidari et al., "A moral framework for understanding fair ml through economic models of equality of opportunity" (paper presented at the Proceedings of the conference on fairness, accountability, and transparency, 2019); Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and machine-learning* (2019), https://fairmlbook.org/; Alexandra Chouldechova and Aaron Roth, "The Frontiers of Fairness in Machine Learning" *arXiv e-prints* (2018). https://ui.adsabs.harvard.edu/\#abs/2018arXiv181008810C; Sina Fazelpour and David Danks, "Algorithmic bias: Senses, sources, solutions" *Philosophy Compass* 16, no. 8 (2021), https://doi.org/https://doi.org/10.1111/phc3.12760, https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12760. Algoritmisk bias er i dansk kontekst og med særligt fokus på offentlige myndigheder behandlet fra et menneskeretsligt perspektiv i Marya Akhtar et al., *Når algoritmer sagsbehandler*, Institut for Menneskerettigheder (2021), https://menneskeret.dk/files/media/document/Algoritmer_8.K.pdf.

bias bredt forstås som et systems tendens til systematiske afvigelser fra en forventet eller ønsket måde at virke på.

Diskussionen af algoritmisk bias har især knyttet sig til systematiske afvigelser på tværs af grupper defineret ved særligt følsomme personkategorier, som køn, etnicitet, seksualitet, religion, eller handicap., for eksempel ved at systemet behandler personer med en sådan egenskab anderledes end andre personer. Man kan i sådanne tilfælde tale om, at systemet har bias mod personer med denne egenskab. Ligesom for bias i dagligdagstale er begrebet ofte negativt ladet på den måde, at det forudsættes at algoritmisk bias er problematisk, eksempelvis fordi den stiller én gruppe af medarbejdere dårligere end andre medarbejdere.

Meget af de seneste års interesse for algoritmisk bias udspringer af en prominent debat om anvendelsen af beslutningsstøttesystemet COMPAS i det amerikanske retsvæsen. COMPAS er designet til at vurdere en persons risiko for recidivisme, dvs. fornyet, fremtidig kriminalitet. Det anvendes i USA blandt andet i forbindelse med myndigheders beslutninger om varetægtsfængsling og prøveløsladelse. I 2016 offentliggjorde tænketanken Pro Publica en analyse, som hævdede, at COMPAS havde en bias, der medførte at sorte og hvide amerikanere systematisk blev forskelsbehandlet.[108]

Pro Publica præsenterede dokumentation for, at systemet havde to kontroversielle tendenser til at behandle de to grupper forskelligt. I gruppen af personer, som *ikke* efterfølgende begik en forbrydelse, blev sorte amerikanere langt oftere (fejl)vurderet til at have en høj risiko end hvide amerikanere. Og i gruppen af personer som efterfølgende *faktisk* begik en forbrydelse, blev hvide amerikanere langt oftere (fejl)vurderet til at have lav risiko end sorte amerikanere. COMPAS var godt nok fair i den forstand, at det begik nogenlunde lige mange fejl, når det vurderede de to grupper, men systemet havde tendens til at begå én type fejl, når det vurderede den ene gruppe, og en anden type fejl, når det vurderede den anden.

En omfattende litteratur har efterfølgende diskuteret en række forskellige måder, hvorpå et system kan manifestere bias mod relevante grupper.[109] På et generelt plan kan man skelne mellem fire typer algoritmisk bias:

- Bias ved anvendelse af følsomme variable
- Bias i fordelingen af positive og negative vurderinger
- Bias i vurderingernes præcision

---

108 Julia Angwin et al., "Machine Bias" *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

109 Se eksempelvis Deborah Hellman, "Measuring Algorithmic Fairness" *Virginia Law Review* 106, no. 4 (2020); Jon Kleinberg et al., "Discrimination in the Age of Algorithms" *arXiv e-prints* (2019). https://ui.adsabs.harvard.edu/\#abs/2019arXiv190203731K; Solon Barocas and Andrew D. Selbst, "*Big Data*'s Disparate Impact" *California Law Review* 104, no. 3 (2016 2016), https://doi.org/10.2139/ssrn.2477899, https://www.ssrn.com/abstract=2477899; Cynthia Dwork et al., "Fairness Through Awareness" arXiv:1104.3913 [cs] (2011), http://arxiv.org/abs/1104.3913; Sorelle A. Friedler et al., "A comparative study of fairness-enhancing interventions in machine learning" (Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, Association for Computing Machinery, 2019); Barocas, Hardt, and Narayanan, *Fairness and machine-learning*; Chouldechova and Roth, "The Frontiers of Fairness in Machine Learning."; Moritz Hardt, Eric Price, and Nathan Srebro, "Equality of Opportunity in Supervised Learning" *arXiv:1610.02413 [cs]* (2016), http://arxiv.org/abs/1610.02413; Sam Corbett-Davies and Sharad Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning" *arXiv preprint arXiv:1808.00023* (2018). Et nyligt overblik findes i Carey and Wu, "The statistical fairness field guide: perspectives from social and formal sciences."

- Bias i hvilken type fejl, systemet begår

Nedenfor præsenterer vi hver af disse forskellige typer bias. Efterfølgende diskuterer vi i de næste afsnit først hvordan algoritmisk bias kan opstå og om man kan undgå bias, og dernæst to prominente forklaringer på, at algoritmisk bias kan være moralsk problematisk, som handler om henholdsvis mulighedsulighed og skade.

6.3.1. Bias og følsomme variable

Et automatiseret beslutningssystem anvender data til at foretage en vurdering. Konkret optræder disse data som variable i systemets model - en del af den matematiske funktion, hvor en medarbejders data kan indsættes. Hvis eksempelvis modellen anvender "anciennitet" som variabel, så indeholder modellen et led af funktionen, hvor værdien for den enkelte medarbejders anciennitet kan sættes ind. Et af de afgørende skridt, når man udvikler et automatiseret beslutningssystem, er at vælge hvilke data, modellen potentielt skal anvende som variable. Og den måske mest enkle måde, hvorpå et automatiseret beslutningssystem systematisk kan behandle personer forskelligt, er ved at systemet anvender data om en personkarakteristik som variabel.[110]

I en bred fortolkning, af denne form for algoritmisk bias, kan beslutningssystemer i princippet have algoritmiske bias for alle de variable, som systemet anvender. Eksempelvis kan et system, som forsøger at identificere den bedst kvalificerede ansøger til en stilling, lægge vægt på ansøgeres uddannelse. Når det gør det, så forskelsbehandler systemet i den trivielle forstand, at det systematisk behandler ansøgere forskelligt afhængigt af hvilken uddannelsesbaggrund de har. Det kan imidlertid lyde underligt, at sige, at systemet har en bias mod ansøgere med nogle uddannelser. Fokus i debatten om algoritmisk bias er typisk på de særligt følsomme personkarakteristika, som også er i centrum i diskriminationslovgivningen.

Dette fokus kan forklares ved at pege på, at det i mange tilfælde ikke er en uønsket eller uventet afvigelse, når et system behandler personer forskelligt ved at anvende variable om for eksempel uddannelse, og derfor ikke udgør det, vi normalt forstår ved en bias. Danmark, og de omgivende Europæiske lande, er i dag præget af liberale normer som tilsiger, at personer ikke må behandles forskelligt på baggrund af særligt følsomme personkarakteristika. Systematisk forskelsbehandling af sådanne grupper er således normalt en uventet eller uønsket afvigelse. Et automatiseret beslutningssystem behandler imidlertid i udgangspunktet alle data ens – en læringsalgoritme skelner ikke af sig selv mellem variable, som karakteriserer følsomme grupper, og andre variable. Det betyder, at en læringsalgoritme lige så godt kan træne et system til at lægge vægt på eksempelvis køn, etnicitet eller religion, som på andre data. Hvilke data et system lægger vægt på afhænger alene af hvilke data det trænes på, og hvilke begrænsninger udvikleren eventuelt sætter (se afsnit 6.4.2 nedenfor, om "blinding" af læringsalgorimer).

I Danmark introducerede Styrelsen for Arbejdsmarked og Rekruttering i 2017 en opdateret udgave af et automatiseret beslutningssystem, som blev anvendt af danske jobcentre til at

110 Nina Grgic-Hlaca et al., "Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning" (Association for the Advancement of Artificial Intelligence, 2018).

vurdere nylediges risiko for langtidsledighed.[111] Systemet var et såkaldt prædiktivt beslutningstræ, som i en række beslutningsknudepunkter filtrerede personer til forskellige forgreninger af træet, for til sidst at placere dem i en demografisk gruppe med en vis statistisk sandsynlighed for langtidsledighed.

Algoritmisk bias ved anvendelse af følsomme variable:

Et automatiseret beslutningssystems tendens til systematisk at behandle personer fra en følsom gruppe anderledes end personer fra andre grupper, fordi systemet anvender data om medlemskab af gruppen som variabel.

Systemet blev kritiseret i den offentlige debat, da det viste sig, at mindst et af beslutningspunkterne anvendte "ikke-vestlig herkomst" som variabel, således at personer med "vestlig herkomst" og personer med "ikke-vestlig herkomst" blev sorteret til forskellige forgreninger i træet. I praksis havde dette den konsekvens, at nyledige med "ikke-vestlig herkomst" i nogle tilfælde kunne klassificeres som havende høj risiko for langtidsledighed, selvom tilsvarende personer med "vestlig herkomst" blev klassificeret som havende lav risiko for langtidsledighed. Dette blev af mange opfattet som en algoritmisk bias, der i praksis udgjorde en form for negativ direkte forskelsbehandling på baggrund af etnicitet.[112]

Bias i form af anvendelse af data om medlemskab af en særligt følsom gruppe som variabel kan imidlertid være relativt sjælden. Det skyldes at der som nævnt findes vidt udbredte liberale normer, og at disse normer er kodificeret i både dansk og europæisk diskriminationslovgivning. Mange måder at anvende medlemskab af en følsom gruppe som variabel er således omfattet af juridiske forbud mod direkte diskrimination.[113] Automatiserede beslutningssystemer kan derfor have større risiko for andre algoritmiske bias, som kan opstå uden at systemet anvender data om særligt følsomme personkarakteristika.

6.3.2. Bias og vurderinger

Et automatiseret beslutningssystem kan have en algoritmisk bias, i den forstand at det systematisk behandler personer fra forskellige grupper på forskellige måder, uden at systemet anvender eller har adgang til data om medlemskab af de pågældende grupper. Det kan ske når et system anvender data, som statistisk hænger sammen med medlemskab af en gruppe (se 6.4 nedenfor om hvordan algoritmisk bias kan opstå). Der er imidlertid flere måder, at

111 MPLOY, *Evaluering af projekt "Samtaler og indsats der modvirker langtidsledighed"*, Styrelsen for Arbejdsmarked og Rekruttering, (2018), https://star.dk/media/8004/evaluering-af-projekt-samtaler-og-indsats-der-modvirker-langtidsledighed.pdf; Therese Moreau and Frederik Kulager, "Vi har skilt jobcentrenes algoritme ad" *Zetland*, June 10, 2021, https://www.zetland.dk/historie/sOMVZ7qG-aOz9m93B-a30b8; Akhtar et al., *Når algoritmer sagsbehandler*.

112 Jens Bostrup, "Tellis oplevelse på jobcenteret fører nu til en optrapning i kampen mod diskriminerende algoritmer" *Politiken* 2021, https://politiken.dk/viden/Tech/art8140892/Tellis-oplevelse-p%C3%A5-jobcenteret-f%C3%B8rer-nu-til-en-optrapning-i-kampen-mod-diskriminerende-algoritmer.

113 Det er værd at bemærke, at Ligebehandlingsnævnet i 2022 behandlede en klage over STARs system til vurdering af risici for langtidsledighed, og nåede frem til, at systemet ikke ulovligt diskriminerede på baggrund af etnicitet. Se Ligebehandlingsnævnets afgørelse om Etnisk oprindelse, No. 9518 (Ligebehandlingsnævnet 2022).

definere hvad det vil sige, at systemet behandler grupper forskelligt, og derfor flere måder, et system kan siges at have algoritmisk bias.

En første måde at definere algoritmisk bias handler om systemets tendens til at give personer fra forskellige grupper forskellige vurderinger.[114] Et system har bias i denne forstand, hvis det systematisk vurderer medlemmer af én gruppe anderledes end medlemmer af andre grupper. Et eksempel kunne være et system til beslutningsstøtte i forbindelse med ansættelse, som systematisk vurderer mandlige ansøgere som bedre kvalificerede end kvindelige ansøgere.

Algoritmisk bias i fordelingen af vurderinger:

Et automatiseret beslutningssystems tendens til systematisk at give forskellige vurderinger til personer fra forskellige grupper.

Demografiske forskelle er ofte i fokus i debatter om bias og forskelsbehandling, eksempelvis når vi taler om, at topposter i politik eller erhvervsliv fortsat er ulige fordelt mellem mænd og kvinder. Mange kritikere har imidlertid fremført, at demografisk bias i automatiserede beslutningssystemer kan afspejle virkelige forskelle mellem grupperne.[115] Hvis det i en bestemt situation faktisk var sådan, at mandlige ansøgere i gennemsnit var bedre kvalificerede end kvindelige ansøgere, så ville et system med høj præcision systematisk vurdere mænd og kvinder forskelligt. Systemet ville derfor have demografisk bias i den forstand, som her er på tale. I en sådan situation, kan en kritiker hævde, gør den demografiske bias imidlertid ikke af sig selv et automatiseret beslutningssystem moralsk problematisk (se afsnit 6.5 nedenfor om hvad der kan være moralsk problematisk ved algoritmisk bias).[116]

6.3.3. Bias og præcision

En anden måde at definere algoritmisk bias på handler om systemets tendens til at producere mere eller mindre *præcise* vurderinger. Et system kan eksempelvis være meget præcist, når det anvendes til at vurdere medlemmer af én gruppe, og mindre præcist, når det anvendes til at vurdere medlemmer af en anden gruppe.

I faglitteraturen finder man en lang række forskellige måder at definere et systems præcision.[117] Den mest enkle er *overordnet præcision*, hvor man blot måler hvor mange af systemets

114 Denne form for algoritmisk bias kaldes i faglitteraturen ofte for "demografisk bias". Se Dwork et al., "Fairness Through Awareness."; Matt J. Kusner et al., "Counterfactual Fairness" *arXiv e-prints* (2017). https://ui.adsabs.harvard.edu/\#abs/2017arXiv170306856K.

115 Se eksempelvis Barocas, Hardt, and Narayanan, *Fairness and machine-learning*.

116 Bemærk, at der i en sådan situation kunne være grund til alligevel at forsøge at undgå demografisk bias ved at sikre ligelig repræsentation. Hvis eksempelvis kvinders gennemsnitligt lavere kvalifikationer skyldes, at de er blevet diskrimineret imod, så kunne man hævde, at det vil være uretfærdigt, at lade effekten af denne diskrimination påvirke deres muligheder for ansættelse. Og hvis en ligelig repræsentation har gavnlige effekter, for eksempel ved at skabe rollemodeller, som piger og unge kvinder kan spejle sig i, så kunne denne reaktions-kvalifikation tale for at prioritere kvindelige ansøgere, som i andre henseender har lavere kvalifikationer. Se også afsnit 6.5 nedenfor.

117 Se eksempelvis Barocas, Hardt, and Narayanan, *Fairness and machine-learning*; Carey and Wu, "The statistical fairness field guide: perspectives from social and formal sciences."

vurderinger, som er korrekte. Denne tilgang til præcision kan yderligere fokuseres ved at skelne mellem præcision for systemets positive vurderinger, såkaldt *positiv prædiktiv værdi*, og præcision for systemets negative vurderinger, såkaldt *negativ prædiktiv værdi*.

Algoritmisk bias i fordelingen af præcision:

Et automatiseret beslutningssystems tendens til systematisk at lave mere præcise vurderinger for personer fra én gruppe, end for personer fra andre grupper.

Hvis en arbejdsplads eksempelvis anvender automatiseret beslutningsstøtte til at vurdere, om medarbejdere har indfriet kriterier for en lønforhøjelse, så kan systemet vurdere, at en medarbejder *har* indfriet kriterierne (positiv vurdering), eller at en medarbejder *ikke har* indfriet kriterierne (negativ vurdering). I de tilfælde hvor systemet fejlagtigt vurderer, at en medarbejder har indfriet kriterierne, selvom dette ikke er tilfældet, er fejlvurderingen en "falsk positiv". Omvendt er fejlvurderingen en "falsk negativ", i de tilfælde hvor systemet fejlagtigt vurderer, at en medarbejder ikke har indfriet kriterierne, selvom dette faktisk er tilfældet. Hvis arbejdspladsen er mest optaget af at sikre, at medarbejdere som har fortjent en lønforhøjelse også får den, så kan det være mest interessant for arbejdspladsen at kigge på systemets negative prædiktive værdi. Den fortæller hvor mange af de negative vurderinger, som er korrekte (og hvor mange, som er fejlagtige). Hvis arbejdspladsen omvendt er mest optaget af, at kun de medarbejdere, som har fortjent en lønforhøjelse også får den, så kan det være mere interessant at kigge på den positive prædiktive værdi, som fortæller hvor mange af systemets positive vurderinger, som er korrekte.

Positiv prædiktiv værdi:

$$\frac{\text{Antallet af personer, som både er } \textit{vurderet} \text{ positivt og } \textit{er} \text{ positive}}{\text{Det samlede antal personer, som er vurderet positivt}}$$

Negativ prædiktiv værdi:

$$\frac{\text{Antallet af personer, som både er } \textit{vurderet} \text{ negativt og } \textit{er} \text{ negative}}{\text{Det samlede antal personer, som er vurderet negativt}}$$

I mange sammenhænge kan det virke intuitivt, at et system bør have lige god overordnet præcision, positiv prædiktiv værdi og/eller negativ prædiktiv værdi for forskellige grupper, det vil sige, at et system ikke har algoritmisk bias i præcision. Hvis et system har lavere overordnet præcision, så laver det ganske enkelt flere fejl for den ene gruppe end for den anden. Og hvis eksempelvis et system har højere negativ prædiktiv værdi for mænd end for kvinder, så betyder det, at de negative vurderinger af kvinder oftere er fejlagtige. I eksemplet ovenfor ville det for arbejdspladsen betyde, at de i mindre grad kunne stole på systemets vurderinger af de kvindelige medarbejdere, end på vurderinger af de mandlige medarbejdere. Og for de kvindelige medarbejdere, som fik afslag på en lønforhøjelse, ville det betyde, at de oftere havde fortjent den, end deres mandlige kollegaer, som fik afslag på en lønforhøjelse.

Ikke desto mindre har kritikere fremført, at det ikke er tilstrækkeligt, at et system undgår algoritmisk bias i præcision. Det kan være nødvendigt, hævder sådanne kritikere, at kigge på hvordan systemet vurderer henholdsvis den positive og negative klasse, snarere end på vurderingernes prædiktive værdi. Derved får man blik for hvilke *typer* fejl et system har tendens til at begå (se også afsnit 6.4.4 nedenfor om at undgå bias).

6.3.4. Bias og fejltyper

De typer præcision, som er beskrevet ovenfor, kigger på hvor mange af et systems positive vurderinger, som bliver givet til personer med den pågældende egenskab, og hvor mange negative vurderinger, som bliver givet til personer uden den pågældende egenskab. I begge tilfælde er spørgsmålet altså, hvad sandsynligheden er for, at en *vurdering* er korrekt. Men det er også muligt at spørge, hvor mange af de *personer*, der henholdsvis har og ikke har en egenskab, som modtager en positiv og negativ vurdering.

Når man eksempelvis spørger, hvor mange af de personer, som besidder den relevante egenskab (den "positive klasse"), som modtager en (korrekt) positiv vurdering, så måler man "sand positiv raten" (også kaldet sensitivitet; eng. "sensitivity" eller "recall"). Når man tilsvarende spørger, hvor mange af de personer som ikke besidder den relevante egenskab, som modtager en negativ vurdering, så måler man "sand negativ raten" (også kaldet specificitet; eng. "specificity").

Sand positiv rate:

$$\frac{\text{Antallet af personer, som både er } \textit{vurderet} \text{ positivt og } \textit{er} \text{ positive}}{\text{Antallet af personer, som er positive}}$$

Sand negativ rate

$$\frac{\text{Antallet af personer, som både er } \textit{vurderet} \text{ negativt og } \textit{er} \text{ negative}}{\text{Antallet af personer, som er negative}}$$

Skiftet i perspektiv fra præcision for positive og negative vurderinger til præcision for den positive og negative klasse er vigtigt, men kan være vanskeligt at begribe. Det kan også være fristende at tænke, at de to perspektiver på præcision må følges ad. En variant af eksemplet med vurdering af fortjent lønforhøjelse, som vi ovenfor har benyttet, kan illustrere forskellen, og hvordan de to perspektiver kan belyse forskellige kvaliteter ved et system.

Vi forestiller os, at en arbejdsplads benytter et automatiseret beslutningssystem til at vurdere, hvorvidt medarbejdere har gjort sig fortjent til lønforhøjelse. Arbejdspladsen har 100 ansatte, hvoraf halvdelen er mænd og halvdelen kvinder. Heraf har 16 kvinder og 12 mænd gjort sig fortjent til lønforhøjelse. Systemet er imidlertid ikke fejlfrit. Det har en overordnet præcision på 0.8 for begge grupper, en positiv prædiktiv værdi på 0.75 for begge grupper, og næsten samme negative prædiktive værdi for de to grupper (0.82 og 0.84). Men systemet har meget forskellig sensitivitet.

Figur: I alt 12 kvindelige medarbejdere vurderes positivt. Heraf har 9 fortjent lønforhøjelse, mens 3 ikke har. Også 4 mandlige medarbejdere vurderes positivt, men heraf har kun 3 fortjent en lønforhøjelse, mens 1 ikke har. Tilsvarende får i alt 38 kvinder afslag på lønforhøjelse, hvoraf 7 havde fortjent den, mens kun 9 af de i alt 46 mænd, som får afslag, havde fortjent en lønforhøjelse.

Af de 16 kvinder, som reelt fortjener en lønforhøjelse, vurderes 9 positivt. Hvis arbejdspladsen baserer sine beslutninger på systemets vurderinger, så får mere end halvdelen af de kvinder, som har gjort sig fortjent til en lønforhøjelse, faktisk en lønforhøjelse. Blandt de 12 mænd, som reelt fortjener en lønforhøjelse, er det derimod kun 3 som vurderes positivt. Det betyder, at kun en fjerdedel af de fortjenstfulde mænd får en lønforhøjelse.[118] Selvom systemet har samme overordnede præcision og positive prædiktive værdi, så har det tendens til langt oftere, at lave falsk negative fejl, når det vurderer mænd, end når det vurderer kvinder.[119] Man kan sige, at systemet har algoritmisk bias i fordelingen af fejltyper.

Algoritmisk bias i fordelingen af fejltyper:

Et automatiseret beslutningssystems tendens til systematisk at lave forskellige typer fejl for personer fra forskellige grupper.

I sådanne situationer vil mange intuitivt mene, at det er rimeligt at kræve, at systemet har samme sand negativ rate og sand positiv rate på tværs af relevante grupper. Desværre er det typisk umuligt på samme tid at undgå både bias i præcision og bias i fordelingen af fejltyper. Det kigger vi på i næste afsnit, hvor vi diskuterer hvordan algoritmisk bias kan opstå, og hvad man kan gøre for at undgå det.

118 Bemærk at sandsynlighed her fortolkes som frekvensen i en gruppe, eller anderledes udtrykt, som sandsynligheden for, at et tilfældigt valgt medlem af gruppen, bliver vurderet positivt.

119 Systemet har derfor også en tendens, om end mindre stærk, til at lave flere falsk positive fejl, når det vurderer kvinder, end når det vurderer mænd. Kvinder, som ikke fortjener en lønforhøjelse, har gennemsnitligt 9% chance for at blive vurderet positivt. Mænd, som ikke fortjener en lønforhøjelse, har gennemsnitligt 3% chance for alligevel at blive vurderet positivt.

6.4. Hvordan opstår algoritmisk bias?

Algoritmisk bias, i de forskellige varianter som vi ovenfor har præsenteret, kan opstå på flere forskellige måder. Først og fremmest kan bias opstå i kraft af den model, som udvikleren vælger at anvende. Bias kan også opstå i træningsfasen ved systematiske fejl i træningsdata, eller ved at udvikleren fokuserer på et særligt sæt af træningseksempler eller variable. I dette afsnit kigger vi på disse forskellige kilder til bias, inden vi diskuterer hvilke muligheder der findes for at reducere eller undgå algoritmisk bias.

6.4.1. Bias, modeltyper og gruppeforskelle

Kernen i et automatiseret beslutningssystem er en matematisk model, som repræsenterer sammenhænge mellem dét, som systemet forsøger at vurdere, og de data, som systemet fodres med. Forskellige læringsalgoritmer træner systemer med forskellige typer modeller, fra logistiske regressionsmodeller, over prædiktive beslutningstræer, og til dybe neurale netværk.[120] Nogle er enkle og lette at fortolke, andre mere komplekse og vanskelige at forstå. Ofte viser det sig i en konkret situation, at nogle modeltyper fungerer bedre end andre. En central opgave for udvikleren af et beslutningssystem er derfor, at finde den modeltype, der egner sig bedst til netop den vurdering, som systemet skal foretage.

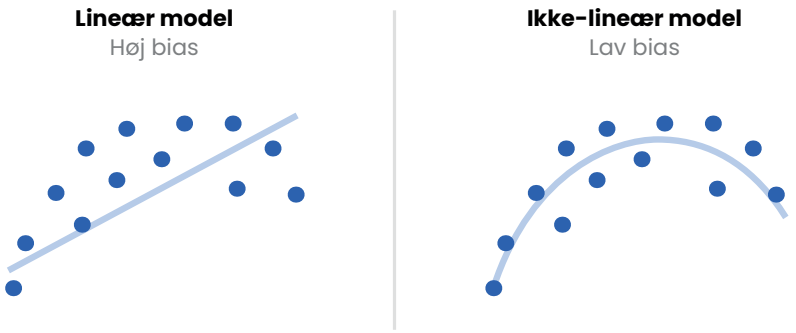
Modelbias:

Et automatiseret beslutningssystems indbyggede tendenser til at repræsentere sammenhænge i data på måder, der skaber systematiske forskelle i systemets vurderinger.

Én måde at forklare, at en modeltype kan fungere bedre i nogle situationer end i andre, er ved at pege på, at alle modeller har indbyggede bias, i den specielle forstand, at enhver model har tendenser til at repræsentere sammenhænge i data på bestemte måder, og at dette kan skabe systematiske forskelle i systemets vurderinger. Modeller generaliserer fra træningsdata for at finde og repræsentere sammenhænge, men den specifikke måde en bestemt type model generaliserer og repræsenterer sammenhænge giver modellen systematiske tendenser i hvordan den behandler data. Et enkelt eksempel er, at visse (meget) simple modeller ikke kan repræsentere komplekse sammenhænge i data, for eksempel ikke-lineære sammenhænge og interaktioner mellem variable. Hvis der faktisk er sådanne komplekse sammenhænge i de data, som systemet baserer sin vurdering på, så vil et system med en simpel model få en tendens til systematisk at begå fejl, der varierer med effekten af de komplekse sammenhænge.

120 Et overblik over en række af de mest almindelige modeltyper kan findes i Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning* (Springer, 2017). https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12_toc.pdf; Gareth James et al., *An Introduction to Statistical Learning* (New York: Springer, 2013).Overblik fokuseret på forskellige elementer af dataetik findes i eksempelvis Molnar, *Interpretable machine learning. A guide for making black box models explainable.*og Barocas, Hardt, and Narayanan, *Fairness and machine-learning.*

Modelbias behøver ikke at føre til forskelsbehandling på tværs af følsomme personkategorier som køn, etnicitet, eller religion. Men hvis der er sammenfald mellem forskelle i data for de pågældende grupper og modellens bias, så kan konsekvensen være, at systemet virker forskelligt for grupperne.

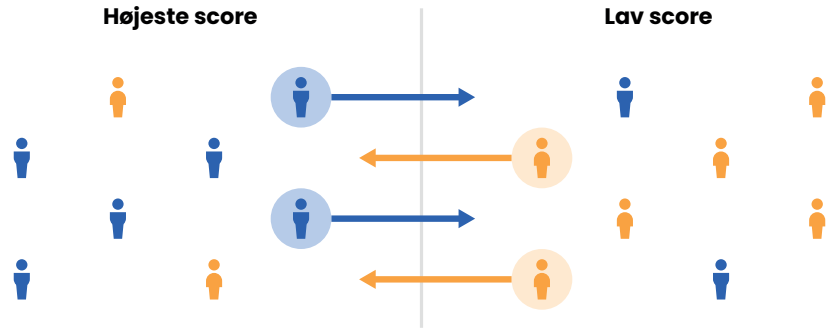


Sådanne forskelle i data mellem grupper kan også føre til algoritmisk bias på andre måder. Det kan eksempelvis være tilfældet, hvis der forskel på basisraten, eller der er større varians i data for én gruppe end for en anden. At der er forskel på basisraten betyder, at den egenskab, systemet skal vurdere, er mere almindelig i en gruppe end i andre grupper. Det kan umiddelbart føre til forskelle i hvordan systemet vurderer personer fra de to grupper (se afsnit 6.3.2 ovenfor), men kan også give systemet tendens til at begå forskellige typer fejl. At der er forskelle i varians betyder, at der for én gruppe er mange og/eller stærke sammenhænge mellem data og det, som systemet forsøger at vurdere, mens der er få og/eller svage sammenhænge i data for andre grupper. En sådan forskel vil have den effekt, at det er nemmere at vurdere medlemmer af den ene gruppe, end det er at vurdere medlemmer af den anden gruppe, uanset hvordan udvikleren træner beslutningssystemet.

Som illustration kan vi igen forestille os en arbejdsplads, som ønsker at bruge automatiseret beslutningsstøtte til at vurdere ansøgere på baggrund af uddannelse. Vi kan også forestille os, at det viser sig, at der for kvinder er en tæt sammenhæng mellem hvilke karakterer en ansøger har fået under uddannelse og ansøgerens præstation som medarbejder, mens der for mandlige ansøgere er en svag eller slet ingen sammenhæng mellem disse forhold. I denne situation vil det være vanskeligt at træne en model, som kan vurdere mandlige ansøgere, men relativt let at træne en model for kvindelige ansøgere. Et beslutningssystem vil derfor have en tendens til højere præcision for den ene gruppe end for anden. En sådan tendens vil være en konsekvens af systematiske forskelle mellem grupperne, og kan derfor være vanskelig at forhindre. Men det er dog muligt at sikre, at træning af modellen ikke forstærker den pågældende bias. En bias kan eksempelvis blive forstærket, hvis læringsalgoritmen har tendens til at træne en "doven" model – en model, der prioriterer korrekt vurdering af de eksempler, som er lettest at vurdere – men graden af dovenskab varierer afhængigt af modeltype, ligesom det er muligt at tilrettelægge træningen således at man modvirker tendensen.[121]

121 Samuel James Bell and Levent Sagun, "Simplicity Bias Leads to Amplified Performance Disparities" (Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago, IL, USA, Association for Computing Machinery, 2023).

6.4.2. Bias og fejlagtige træningsdata



Bias i et automatiseret beslutningssystem kan også opstå ved systematiske fejl i de data, som systemet trænes på. Sådanne systematiske fejl kan skyldes enten at der er bias i den praksis som generer data, for eksempel fordi den diskriminerer mod nogle grupper, eller at der er bias i registreringen af data, for eksempel fordi der ofte bliver begået fejl i registreringen af data der knytter sig til nogle grupper.[122] Uanset årsag skaber de systematiske fejl en bias i data.

Bias i data:

Systematiske fejl i de data, som et automatiseret beslutningssystem trænes på og/eller anvender, således at en eller flere grupper misrepræsenteres.

Bias i data kan angå den målegenskab, som systemet forsøger at vurdere, de variable som statistisk hænger sammen med denne målegenskab, eller begge dele. En udvikler kan eksempelvis forsøge at udvikle et system til automatisk prioritering af ansøgere i forbindelse med ansættelse, og i den forbindelse anvende evalueringer af præstation som målestok for, hvor succesfulde tidligere ansøgere har vist sig at være, når de blev ansat (målegenskaben). Men hvis kvinder historisk er blevet diskrimineret på arbejdspladsen, ved systematisk at have modtaget dårligere præstationsmålinger end mænd for den samme indsats, vil denne diskrimination afspejle sig i de historiske data.[123] Træningsdata vil i så fald systematisk misrepræsentere kvinders faktiske præstationer. Et system trænet på sådanne data vil lære, at kvindelige ansøgere i gennemsnit bliver mindre succesfulde medarbejdere end mandlige ansøgere, selvom dette ikke er tilfældet.

122 Et meget omdiskuteret eksempel er politiets overpatruljering af kvarterer, hvor etniske minoriteter er overrepræsenterede blandt beboerne, hvilket fører til at disse minoriteter overrepræsenteres i data om interaktioner med politiet, herunder arrestationer og kriminalitet. Se A.G. Ferguson, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement* (NYU Press, 2017). Danielle Ensign et al., "Runaway Feedback Loops in Predictive Policing" (1st Conference on Fairness, Accountability and Transparency, arXiv, 2017).

123 Når vi her og andre steder beskriver eksempelvis diskrimination af kvinder på arbejdsmarkedet som en hypotetisk situation, er det alene fordi det optræder i sammenhæng med et fiktivt eksempel. Det er ikke udtryk for den opfattelse, at kvinder ikke oplever diskrimination på arbejdsmarkedet.

En udvikler kan forsøge at forhindre denne form for bias ved at "blinde" læringsalgoritmen.[124] At blinde læringsalgoritmen betyder, at udvikleren sikrer, at data om følsomme personkategorier ikke optræder som variable i træningsdata (se også afsnit 6.3.1 om bias ved anvendelse af følsomme variable). Hvis læringsalgoritmen eksempelvis ikke har data om historiske ansøgeres køn, kan den ikke træne et system, som bruger køn som variabel, og i kraft deraf vurderer ansøgere forskelligt på baggrund af køn.

Når man kan blinde læringsalgoritmen, og derved forhindre at systemet anvender følsomme variable, kunne man måske håbe, at udfordringen var løst. Det har imidlertid vist sig, at data om følsomme personkategorier ofte hænger tæt sammen med andre data, og at maskinlæring er meget effektiv til at opdage og udnytte disse sammenhænge, til at opnå samme forskelsbehandling i praksis. I litteraturen kaldes dette fænomen for redundant indkodning (eng. "redundant encoding").[125]

I forlængelse af eksemplet ovenfor, så kan man forestille sig, at ansøgere lister fritidsaktiviteter på deres CV, og at der er statistiske forskelle på, hvilke fritidsaktiviteter mænd og kvinder dyrker – med et stereotypt eksempel fordi mænd oftere spiller fodbold og kvinder oftere dyrker yoga. Hvis det er tilfældet, så kan læringsalgoritmen opdage, at visse fritidsaktiviteter statistisk hænger sammen med højere præstation og visse med lavere præstation. Denne sammenhæng i de historiske data kan i princippet udelukkende skyldes kombinationen af forskelle i fritidsaktiviteter på tværs af køn og den historiske diskrimination af kvinder. Ved at træne et system, som vurderer ansøgere med typisk mandlige fritidsaktiviteter højere, og ansøgere med typisk kvindelige fritidsaktiviteter lavere, kan udvikleren ende med et system, som i praksis forskelsbehandler på næsten samme måde, som det ville gøre, hvis det anvendte køn som variabel. Det kan især være tilfældet, hvis flere forskellige variable statistisk hænger sammen med køn.[126]

Metoden med at blinde læringsalgoritmen vil altså i mange tilfælde ikke forhindre, at et system reproducerer en bias i data. Men oven i købet kan blinding af læringsalgoritmen i nogle tilfælde forhindre systemet i at modvirke bias i data. Det kan især være tilfældet, når der er bias i de data, som systemet skal bruge som variable. Hvis vi fortsætter eksemplet med et automatiseret beslutningssystem til prioritering af ansøgere, så kan man forestille sig, at det anvender data om ansøgernes karakterer på job-relevante uddannelser. Vi kan også forestille os, at kvinder diskrimineres i uddannelsessystemet, således at kvindelige ansøgere gennemsnitligt har fået lavere karakterer end mandlige studerende med de samme talenter. Hvis læringsalgoritmen er blindet, og ikke skelner mellem køn, så vil den lære en gennemsnitlig sammenhæng på tværs af køn mellem ansøgeres karakterer og præstationer som medarbejdere. Kvindelige

---

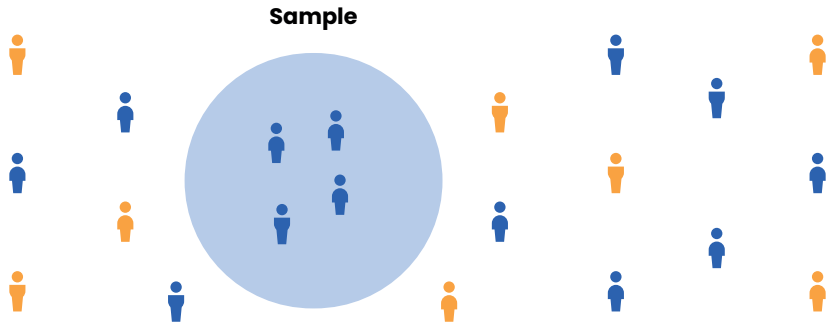
124 Dwork et al., "Fairness Through Awareness."; Niki Kilbertus et al., "Blind Justice: Fairness with Encrypted Sensitive Attributes" *arXiv:1806.03281 [cs, stat]* (2018), http://arxiv.org/abs/1806.03281.

125 Dwork et al., "Fairness Through Awareness."; Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley, "Does mitigating ML's impact disparity require treatment disparity?" (32nd Conference on Neural Information Processing Systems, 2018).

126 Et virkeligt og meget omtalt eksempel angik Amazons udvikling af en algoritme til prioritering af ansøgere, som måtte skrinlægges, da den udviste netop en sådan tendens til at reproducere historisk diskrimination mod kvinder ved at anvende statistiske forskelle mellem mandlige og kvindelige ansøgere. Se Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women" in *Ethics of data and analytics* (Auerbach Publications, 2022); Jeremias Adams-Prassl, Reuben Binns, and Aislinn Kelly-Lyth, "Directly Discriminatory Algorithms" *The Modern Law Review* 86, no. 1 (2023), https://doi.org/https://doi.org/10.1111/1468-2230.12759, https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-2230.12759.

ansøgere vil i den situation stå dårligere end mandlige ansøgere, fordi systemet ikke tager højde for, at data om kvindelige ansøgeres karakterer undervurderer deres faktiske talenter.[127] Hvis læringsalgoritmen derimod har adgang til data om køn, så vil den kunne lære, at der er forskellige sammenhænge mellem karakterer og præstation på jobbet for mænd og kvinder – kvinder med lavere karakterer klarer sig lige så godt, som mænd med højere karakterer. Læringsalgoritmen ville på den baggrund kunne træne et system, som kompenserer for den diskrimination, som kvinder har været udsat for.[128]

6.4.3. Bias, træningseksempler og variable



Bias kan også opstå ved, at udvikleren anvender et træningssæt, som systematisk overrepræsenterer visse grupper og underrepræsenterer andre grupper. Det vil sige at eksemplerne, som optræder i træningsdata, ikke afspejler diversiteten i den situation, som systemet skal anvendes på. Ofte vil dette betyde, at systemet bliver bedre til at vurdere de grupper som er overrepræsenteret, og dårligere til at vurdere de grupper, som er underrepræsenteret. Et meget omtalt eksempel på dette er systemer til ansigtsgenkendelse, som i mange tilfælde har haft vanskeligt ved at genkende etniske minoriteters ansigter, blandt andet fordi de fortrinsvis var trænet på billeder af personer fra de grupper, som udgør den etniske majoritet i vestlige lande.[129]

---

127 Vi antager i eksemplet også, at mandlige og kvindelige medarbejdere i gennemsnit klarer sig lige godt, at deres præstation vurderes ens, samt at der for begge er den samme, enkle sammenhæng mellem præstation på uddannelsen og præstation i job, således at bedre præstation på uddannelsen forudsiger bedre præstation i jobbet

128 Muligheden for at træne et system, som kan repræsentere denne forskel i data forudsætter en modeltype, som kan repræsentere interaktioner mellem variable, og introducerer til gengæld en bias i form af anvendelse af en følsom variabel. Se Lipton, Chouldechova, and McAuley, "Short Does mitigating ML's impact disparity require treatment disparity?"

129 Joseph P Robinson et al., "Face recognition: too bias, or not too bias?" (paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020); Fabio Bacchini and Ludovica Lorusso, "Race, again: how face recognition technology reinforces racial discrimination" *Journal of Information, Communication and Ethics in Society* 17, no. 3 (2019), https://doi.org/10.1108/JICES-05-2018-0050, https://doi.org/10.1108/JICES-05-2018-0050; Patrick Grothar, Mei Ngan, and Kayee Hanaoka, *Face Recognition Vendor Test (FRVT) - Part 3: Demographic Effects*, National Institute of Standards and Technology (U.S. Department of Commerce, 2019).

I eksemplet med et system til prioritering af ansøgere, så kunne der være visse data, som hang tæt sammen med præstation for mænd, og andre data, som hang tæt sammen med præstation for kvinder. I sådanne tilfælde vil læringsalgoritmen have brug for en stor mængde eksempler på både mænd og kvinder, for at lære at repræsentere begge slags sammenhænge præcist. Hvis træningsdata overrepræsenter mænd og underrepræsenter kvinder, så kan systemet få høj præcision i vurderingen af mandlige ansøgeres fremtidige præstation, men lav præcision i vurderingen af kvindelige ansøgeres fremtidige præstation.[130]

På lidt beslægtet vis kan bias opstå ved, at udvikleren vælger at inkludere visse træningsdata, men undlader at inkludere andre. Hvis der, ligesom i eksemplet ovenfor, er visse data, som hænger tæt sammen med præstation for mænd, men andre data, som hænger tæt sammen med præstation for kvinder, så er det essentielt at begge typer data inkluderes. Selv hvis træningsdata har lige mange mænd og kvinder som eksempler, kan læringsalgoritmen ikke træne en model, som repræsenterer de sammenhænge, der ikke findes i træningsdata. Hvis træningssættet fokuserer på de data, som hænger tæt sammen med præstation for mænd, men ikke for kvinder, så vil systemet blive bedre til at vurdere mandlige ansøgeres fremtidige præstation, end til at vurdere kvindelige ansøgeres fremtidige præstation.

6.4.4. Kan man undgå bias?

Vi har i de foregående afsnit set, at et automatiseret beslutningssystem kan have flere forskellige former for algoritmisk bias, og at disse bias kan opstå på flere forskellige måder. Et almindeligt synspunkt er, at det vil være bedst, hvis et automatiseret beslutningssystem undgår alle former for algoritmisk bias. Ind imellem formuleres synspunktet endda stærkere, som et krav om, at arbejdspladser kun bør benytte automatiserede beslutningssystemer, hvis de er fri for bias.

Den kritiske tænketank *ProPublica* forsvarede en variant af dette synspunkt i debatten om det amerikanske system COMPAS, som vurderede straffede og sigtedes risiko for recidivisme.[131] *ProPublica* anerkendte, at COMPAS var "kalibreret". At et system er kalibreret vil sige, at det har (nogenlunde) samme præcision for forskellige befolkningsgrupper på tværs af systemets vurderinger. At COMPAS var kalibreret betød således, at borgere som fik den samme risiko-score, havde samme risiko for recidivisme, uanset om de var sorte eller hvide. *ProPublica* fremførte, at selvom dette var en vigtig egenskab ved systemet, så var det ikke tilstrækkeligt. Systemet havde nemlig en anden tendens. I gruppen af borgere, som efterfølgende viste sig at være lovlydige, blev flere sorte end hvide borgere fejlvurderet som havende høj risiko. Og

130 En relateret komplikation er, at det i nogle tilfælde kan forbedre systemets præcision, hvis systemet kan skelne mellem ansøgere på baggrund af gruppetilhørsforhold, for eksempel fordi modellen så kan give vægt til netop de sammenhænge, som er relevante for henholdsvis mandlige og kvindelige ansøgere. Dette forudsætter imidlertid anvendelsen af følsomme variable til at behandle personer forskelligt (se afsnit 6.3.1 ovenfor). Se Lipton, Chouldechova, and McAuley, "Short Does mitigating ML's impact disparity require treatment disparity?".

131 Jeff Larson et al., "How We Analyzed the COMPAS Recidivism Algorithm" *ProPublica* (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm; Angwin et al., "Machine Bias."Se også William Dieterich, Christina Mendoza, and Tim Brennan, *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*, NorthPointe (2016).

i gruppen af borgere, som efterfølgende viste sig at begå ny kriminalitet, blev flere hvide end sorte borgere fejlvurderet som havende lav risiko. Lidt forenklet, så havde systemet samme præcision for de to grupper, men tendens til at begå forskellige typer fejl: når sorte borgere blev fejlvurderet var det ofte en overvurdering af deres risiko; når hvide borgere blev fejlvurderet var det ofte en undervurdering af deres risiko. ProPublicas pointe var, at det er lige så vigtigt at undgå denne form for bias i fejltyper.

Desværre er det i praksis umuligt at undgå alle former for bias. Det er ofte muligt at reducere eller undgå bestemte former for bias, for eksempel ved at sikre præcise træningsdata, der repræsenterer diversiteten i den gruppe af medarbejdere, som systemet skal anvendes på, og som indeholder alle de relevante variable (se afsnit 6.4.3 ovenfor). Der findes også en voksende mængde tekniske metoder til at måle og reducere bias, både i trænings- og anvendelsesfasen.[132] Men der vil ofte være algoritmiske bias, som det er vanskeligt at reducere, og for visse algoritmiske bias har reduktion af én bias den konsekvens, at man forstærker en anden bias.

Udfordringen med, at reduktion af nogle bias uundgåeligt forstærker andre bias, er demonstreret i såkaldte umulighedsteoremer for algoritmisk bias. To forskergrupper producerede i kølvandet på COMPAS-debatten uafhængigt matematiske beviser for, at det kun i helt særlige situationer kan lade sig gøre på samme tid at undgå bias i præcision, forstået som forskelle i positiv eller negativ prædiktiv værdi, og bias i fordeling af fejltyper, forstået som forskelle i sand positiv rate (sensitivitet) og sand negativ rate (specificitet).[133] De helt særlige situationer, hvor det kan lade sig gøre, er når systemet enten er fejlfrit eller når alle grupper har samme basisrate. Fejlfrie systemer er så godt som uopnåelige i praksis. At basisraten er den samme betyder, at den egenskab, som systemet skal vurdere, er lige fordelt på tværs af grupperne. Det ville være tilfældet i debatten om COMPAS, hvis sorte og hvide amerikanere gennemsnitligt havde samme sandsynlighed for at recidivere. Der findes imidlertid ofte statistiske forskelle mellem de grupper, som er defineret af særligt følsomme personkarakteristika. Der er således næppe tvivl om, at blandt andet racisme og ulige socioøkonomiske vilkår giver sorte borgere i USA en gennemsnitligt højere sandsynlighed for at recidivere, end hvide borgere i USA.[134] Tilsvarende kan der være statistiske forskelle på medarbejdere fra forskellige grupper på en arbejdsplads på grund af deres vilkår på arbejdspladsen eller i samfundet som hele. Når grupper er forskellige er et automatiseret beslutningssystem, som begår en vis mængde fejl, nødt til at have en af de to typer algoritmisk bias.

---

132 For overblik, Se Friedler et al., "Short A comparative study of fairness-enhancing interventions in machine learning."; Jiawei Chen et al., "Bias and debias in recommender system: A survey and future directions" *ACM Transactions on Information Systems* 41, no. 3 (2023).

133 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores" *arXiv e-prints* (2016). https://ui.adsabs.harvard.edu/\#abs/2016arXiv160905807K; Alexandra Choulde-chova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments" *Big Data* 5, no. 2 (2017), https://ui.adsabs.harvard.edu/\#abs/2016arXiv161007524C; Mitchell et al., "Algorithmic Fairness: Choices, Assumptions, and Definitions."

134 Det er værd at bemærke, at det er et vigtigt men vanskeligt spørgsmål, hvor stor forskellen faktisk er. Det skyldes blandt andet at der kan være forskelle i hvor ofte kriminalitet begået af henholdsvis hvide og sorte borgere registreres. Når dette forekommer bliver resultatet, at den ene gruppe alene af denne grund fremstår som mere kriminelle end den anden gruppe. Se Ensign et al., "Short Runaway Feedback Loops in Predictive Policing.".

Hvorfor kan et system ikke undgå bias? De matematiske beviser er relativt tekniske, men en forlængelse af eksemplet, som vi ovenfor har brugt, kan illustrere udfordringen (se afsnit 6.3.4). I eksemplet anvendte arbejdspladsen et automatiseret beslutningssystem til at vurdere, om medarbejdere fortjente en lønforhøjelse. Systemet havde lige god overordnet præcision (0.8), og positiv prædiktiv værdi (0.75) for de to grupper. Men det havde forskellige sand negativ rate for mænd (0.97) og kvinder (0.91), og *meget* dårligere sand positiv rate for mænd (0.25) end for kvinder (0.56). Et automatiseret beslutningssystem kan ikke ændre på, om medarbejdere faktisk har gjort sig fortjent til en lønforhøjelse. For at skabe bedre balance i mænds og kvinders sand positiv rate kunne udviklere derimod forsøge at træne systemet til korrekt at klassificere nogle af de ni positive mænd, som blev fejlvurderet. Hvis det ikke er muligt blot at forbedre systemets overordnede præcision, så må det øgede antal korrekte vurderinger af fortjenstfulde mænd modsvares af et tilsvarende antal nye fejlvurderinger af mænd, som ikke fortjener en lønforhøjelSe Systemet trænes, så tærsklen for hvornår en mand bliver vurderet som fortjenstfuld er lavere, hvorved både nogle mænd som fortjener det, og nogle mænd som ikke fortjener det, bliver klassificeret positivt.

Figur: I alt 12 kvindelige medarbejdere vurderes positivt. Heraf har 9 fortjent lønforhøjelse, mens 3 ikke har. Også 12 mandlige medarbejdere vurderes positivt, men heraf har kun 7 fortjent en lønforhøjelse, mens 5 ikke har. Tilsvarende får i alt 38 kvinder afslag på lønforhøjelse, hvoraf 7 havde fortjent den, mens kun 5 af de i alt 38 mænd, som får afslag, havde fortjent en lønforhøjelse.

Hvis udviklere eksempelvis har held til at træne systemet, så det klassificerer i alt syv af de tolv fortjenstfulde mænd korrekt, får systemet næsten samme sand positiv rate for mænd og kvinder (0.58 for mænd mod 0.56 for kvinder). Systemet bevarer sin overordnede præcision, fordi det samtidig laver fire nye fejlvurderinger, hvor mænd som ikke fortjener lønforhøjelse nu klassificeres positivt. Er bias derved forhindret? Nej – for ved denne ændring har systemet fået ulige positiv prædiktiv værdi (0.58 mod 0.75). Systemet er altså blevet mindre præcist for positivt vurderede mænd end for positivt vurderede kvinder, og arbejdspladsen har derfor mindre grund til at stole på en positiv vurdering, når den angår en mandlig medarbejder, end

når den angår en kvindelig. Effekten er den samme, hvis man forsøger at ændre på systemets klassifikation af kvindelige medarbejdere: Ved at skrue på det ene parameter, ændrer man også på de andre, og det kan ikke lade sig gøre samtidigt, at skabe balance for dem alle.

Meget af debatten om algoritmisk bias har derfor handlet om hvilke typer bias man bør undgå. Dette spørgsmål kigger vi på i næste afsnit, som handler om hvad der kan gøre algoritmisk bias moralsk problematisk.

6.5. Hvornår er algoritmisk bias uetisk?

Det kan i mange tilfælde være vanskeligt eller endda umuligt, at undgå algoritmisk bias i et automatiseret beslutningssystem. Hvad betyder det for en arbejdsplads, som ønsker at bruge automatiserede beslutningssystemer? Skal man helt lade være med at bruge automatiserede beslutningssystemer? Eller kan det omvendt være lige meget hvilket system man bruger, hvis alle systemer uundgåeligt har algoritmisk bias? Hvis algoritmisk bias gør, at det nogle gange er etisk og nogle gange uetisk, at bruge et automatiseret beslutningssystem, hvornår har et system så for meget algoritmisk bias? Og hvis man skal vælge mellem forskellige slags algoritmiske bias, for eksempel mellem bias med hensyn til præcision og bias med hensyn til fejltyper, hvilken bias bør man så foretrække?

Alle disse spørgsmål afhænger af, hvad der i sidste ende er moralsk problematisk ved algoritmisk bias. Siden COMPAS-debatten i 2016 har de etiske udfordringer ved algoritmisk bias været genstand for intens offentlig og akademisk debat (se afsnit 6.4.4 ovenfor om COMPAS). Der findes derfor i faglitteraturen i dag en række forskellige bud på, hvorfor og hvornår algoritmisk bias er moralsk problematisk. I dette afsnit præsenterer vi to prominente forklaringer, som repræsenterer grundlæggende forskellige tilgange til at besvare spørgsmålet: En som handler om, at algoritmisk bias kan skabe ulighed i personers muligheder, og en som handler om, at algoritmisk bias kan gøre skade.

6.5.1. Algoritmisk bias og mulighedsulighed

Den første og måske mest omdiskuterede forklaring handler om, at algoritmisk bias kan skabe eller forstærke uligheder i individers muligheder. En indflydelsesrig strømning i moderne etik og politisk filosofi argumenterer for, at det er grundlæggende uretfærdigt, hvis sociale strukturer eller individers handlinger skaber ulige muligheder for, at personer kan leve gode liv.[135]

135 Teorien forbindes især med den amerikanske politiske filosof John Rawls, men er udviklet og forsvaret af en lang række senere forskere. John Rawls, *A Theory of Justice* (Oxford: Oxford University Press, 1999). Se også Ronald Dworkin, *Sovereign Virtue – The Theory and Practice of Equality* (Cambridge: Harvard University Press, 2000); Gerald Allan Cohen, "On the Currency of Egalitarian Justice" in *On the Currency of Egalitarian Justice and Other Essays in Political Philosophy*, ed. Michael Otsuka (Princeton: Princeton University Press, 2011); Richard J. Arneson, "Equality and equal opportunity for welfare" *Philosophical Studies* 56, no. 1 (1989).For overblik over hvordan teorien er forsøgt anvendt på algoritmisk bias, se Carey and Wu, "The statistical fairness field guide: perspectives from social and formal sciences."; Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich, "Fairness as equality of opportunity: normative guidance from political philosophy" *arXiv preprint arXiv:2106.08259* (2021).

Ideen om hvad det vil sige, at personer har lige eller ulige muligheder, kan imidlertid fortolkes på lidt forskellige måder.

Mulighedslighed:

Det er moralsk problematisk, at handle på en måde, som bevarer, skaber eller forstærker ulighed i de muligheder, som personer har for at leve gode liv.

En første måde at forstå lige muligheder handler om, at personer ikke møder formelle barrierer. Såkaldt "formel mulighedslighed" kræver blot at individer, som i den relevante henseende har de samme egenskaber, også har de samme muligheder. Moralsk problematisk ulighed indtræffer, hvis personer som har de samme relevante karakteristika, alligevel behandles forskelligt. Et automatiseret beslutningssystem giver personer ulige muligheder i denne henseende, hvis det vurderer personer forskelligt, selvom de har de samme værdier for alle relevante variable. Det kan eksempelvis være tilfældet, hvis en arbejdsplads anvender automatiseret beslutningsstøtte til at prioritere ansøgere, og systemet giver kvindelige ansøgere en negativ vægt i vurderingen, selvom der ingen sammenhæng er mellem køn og kvalifikationer.

Der findes imidlertid stærke argumenter for, at formel lighed må betragtes som et sympatisk men også utilstrækkeligt princip. Det skyldes, at personers karakteristika ikke kun er resultatet af individuelle valg, men også af en kombination af medfødte talenter og sociale omstændigheder. En kvindelig ansøger til en stilling kan eksempelvis have svagere kvalifikationer end en mandlig ansøger, *fordi* den mandlige ansøger har nydt stærkere sociale privilegier. I et konservativt, patriarkalsk samfund kan forskellen være dramatisk, for eksempel ved at kvinden hele sit liv er blevet diskrimineret relativt til manden, og af den grund har haft langt ringere muligheder for at udvikle sine kvalifikationer. Når det er tilfældet, hævder mange fortalere for mulighedslighed, sikrer formel lighed ikke en retfærdig fordeling af goder. Automatiseret beslutningsstøtte, som prioriterer ansøgere baseret på deres givne kvalifikationer, vil tværtimod reproducere og forstærke de uligheder, som uligheden i sociale privilegier har skabt.[136]

Man kunne måske indvende, at udfordringerne for formel lighed først og fremmest udgør et argument for at sikre en ligelig fordeling af sociale privilegier, eksempelvis således at mænd ikke systematisk har bedre muligheder for at udvikle professionelle kvalifikationer end kvinder. Indvendingen ændrer imidlertid ikke på, at formel lighed er utilstrækkelig i de situationer, hvor ulighed i sociale privilegier faktisk har ført til ulige kvalifikationer, ligesom den ikke tager højde for, at formel og social lighed ikke kompenserer for ulighed i fordeling af naturlige talenter.

En stærkere fortolkning af mulighedslighed hævder således, at personer har lige muligheder i den forstand som garanterer en retfærdig fordeling af goder, når det kun er deres individuelle

136 Bemærk at dette er en anden udfordring, end den som vi ovenfor beskrev i afsnittet om fejlagtige data. En person kan blive fejlvurderet, hvis historisk diskrimination har ført til misvisende data. Men pointen her er, at selv hvis data korrekt afspejler en persons kvalifikationer, kan de kvalifikationer en person har være et resultat af diskrimination eller uretfærdige, sociale strukturer

valg, som afgør hvordan deres liv udfolder sig.[137] Uligheder, som skyldes individuelle valg, er derfor moralsk uproblematiske – for eksempel forskelle i kvalifikationer, som skyldes forskellige beslutninger om uddannelse eller karriere – men ikke uligheder, der skyldes forskelle i naturlige talenter eller sociale privilegier, som den enkelte ikke er herre over.[138]

Hvornår er algoritmisk bias moralsk problematisk, hvis man antager den stærke variant af teorien om mulighedsulighed? Det korte svar er, at algoritmisk bias er moralsk problematisk, når den skaber, bevarer eller forstærker ulighed i personers muligheder for at leve gode liv. Teorien giver således principielt mulighed for en præcis og utvetydig etisk evaluering af algoritmisk bias. Svaret på spørgsmålet om, hvornår et automatiseret beslutningssystem har så meget algoritmisk bias, at vi har grund til ikke at bruge systemet er, at dette er tilfældet, når den algoritmiske bias skaber mere ulighed i muligheder, end alternative måder at træffe beslutningen på. Og svaret på spørgsmålet om, hvordan vi skal prioritere mellem forskellige slags algoritmiske bias er, at vi skal vælge de bias, som skaber mindst ulighed i individers muligheder.

Teorien om mulighedslighed rejser imidlertid også to væsentlige komplikationer, når den anvendes til at forklare det etisk problematiske ved algoritmisk bias. Den første komplikation er, at mulighedslighed handler om *individers* muligheder, mens algoritmisk bias er udtryk for statistiske forskelle på tværs af grupper.[139] To grupper kan gennemsnitligt være ens, selvom individerne i de to grupper har meget forskellige muligheder. Hvis mulighedslighed handler om *individers* muligheder, så er fraværet af algoritmisk bias på gruppeniveau øjensynligt ikke tilstrækkeligt.[140]

Den anden komplikation er, at teorier om mulighedslighed handler om individers muligheder set over et helt liv, mens algoritmisk bias handler om muligheder i en isoleret valgsituation. Et automatiseret beslutningssystem til for eksempel lønforhøjelse eller ansættelse, som stiller én person bedre, og en anden person ringere, kan isoleret betragtet have algoritmisk bias mod den anden person. Men hvis bias i denne konkrete situation tjener til at udligne forskelle mellem de to set over deres samlede liv, så vil der i et ulighedsperspektiv ikke være tale om en moralsk problematisk form for bias – tværtimod. Om en algoritmisk bias er moralsk problematisk eller ej afhænger således, ifølge teorien om mulighedslighed, ikke kun af hvordan det automatiserede beslutningssystem påvirker personer, men også af, hvordan disse personer i øvrigt er stillet, og hvor meget af deres situation, som er og ikke er et resultat af individuelle valg.

---

137 Den stærke variant af mulighedslighed behandles i faglitteraturen under den engelske betegnelse "luck egalitarianism", for at understrege, at den sigter mod at udjævne de forskelle mellem individers muligheder som skyldes rent held, eksempelvis at en person er født med usædvanligt få eller usædvanligt mange naturlige talenter, eller i en ressourcestærk versus en ressourcesvag familie. Betegnelsen er vanskelig at oversætte til dansk, og vi holder her fast i at henvise til den bredere teori om mulighedslighed.

138 Stærke varianter af teorien om mulighedsulighed er anvendt på algoritmisk bias af blandt andet Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning."; Heidari et al., "A moral framework for understanding fair ml through economic models of equality of opportunity. Et mere generelt forsvar af mulighedslighed som forklaring af, hvad der kan være moralsk problematisk ved forskelsbehandling findes i Shlomi Segall, "What's so bad about Discrimination?" *Utilitas* 24, no. 1 (2012).

139 Se Clinton Castro and Michele Loi, "The Fair Chances in Algorithmic Fairness: A Response to Holm" *Res Publica* 29, no. 2 (2023), https://doi.org/10.1007/s11158-022-09570-3, https://doi.org/10.1007/s11158-022-09570-3.

140 For generelle argumenter om at moralske hensyn i forbindelse med diskrimination og forskelsbehandling knytter sig til individer, ikke grupper, se Kasper Lippert-Rasmussen, "Discrimination and the Aim of Proportional Representation" Politics, Philosophy & Economics 7 (2008); Kasper Lippert-Rasmussen, "Algorithm-based sentencing and discrimination" *Sentencing and Artificial Intelligence* (2022).

De to komplikationer viser, at der ikke er nogen simpel relation mellem algoritmisk bias og mulighedsulighed, for eksempel på den måde, at bias med hensyn til præcision nødvendigvis er moralsk problematisk, fordi det forstærker ulighed i muligheder. Mulighedslighed fokuserer på individer, ikke på grupper, og på muligheder over et helt liv, ikke i en isoleret situation. Det kan gøre det vanskeligt at anvende teorien til at vurdere algoritmisk bias i en konkret situation. Det kan eksempelvis være vanskeligt at skelne mellem de dele af en persons kvalifikationer, som er resultatet af individuelle valg, og de dele som skyldes held (i bred forstand). Hvilke kvalifikationer ville en kvindelig medarbejder med beskedne sociale privilegier men store naturlige talenter have fået, hvis hun hverken havde været heldig med hensyn til talenter eller uheldig med hensyn til sociale privilegier? Det vil i mange situationer være svært at give præcise og ukontroversielle svar på sådanne spørgsmål, men hvis man vil tage moralsk hensyn til mulighedslighed, når man evaluerer algoritmisk bias i et automatiseret beslutningssystem, så er dette en udfordring, som man er nødt til efter bedste evne at løse.[141]

6.5.2. Algoritmisk bias og skade

En anden forklaring på, hvad der kan være moralsk problematisk ved algoritmisk bias, er den enkle ide, at algoritmisk bias er moralsk problematisk, når den forårsager skade.[142] Det er almindeligt accepteret, at det er moralsk problematisk at skade personer, så den centrale opgave for forklaringen er, at vise hvordan algoritmisk bias kan forårsage skade.

Skadesprincippet:

Det er moralsk problematisk, at handle på en måde, som skader personer.

En første del af forklaringen kan pege på, at der er situationer, hvor algoritmisk bias er udtryk for, at et automatiseret beslutningssystem i utilstrækkelig grad tager hensyn til forskelle mellem grupper. Når eksempelvis et system til diagnosticering af hudkræft fortrinsvis trænes på data fra patienter med lys hud, så kan systemet i kraft deraf få lavere præcision for patienter med mørk hud. Denne ulighed i præcision, som kunne forhindres ved bedre tilrettelagt træning, fører til at mørklødede patienter udsættes for en unødvendig og alvorlig helbredsrisiko.[143]

---

141 Et indflydelsesrigt forsøg på at løse den praktiske udfordring på et generelt niveau er udviklet af den amerikanske økonom John Roemer. Se John E. Roemer, "A Pragmatic Theory of Responsibility for the Egalitarian Planner" *Philosophy & Public Affairs* 22, no. 2 (1993), http://www.jstor.org/stable/2265444.

142 Se M. Altman, A. Wood, and E. Vayena, "A Harm-Reduction Framework for Algorithmic Fairness" *IEEE Security & Privacy* 16, no. 3 (2018), https://doi.org/10.1109/MSP.2018.2701149; Corbett-Davies and Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning."; Hoda Heidari et al., "Fairness behind a veil of ignorance: A welfare analysis for automated decision making" *Advances in neural information processing systems* 31 (2018); Frej Klem Thomsen, "Algorithmic indirect discrimination, fairness and harm" *AI and Ethics* (2023), https://doi.org/10.1007/s43681-023-00326-0, https://doi.org/10.1007/s43681-023-00326-0.Forklaringen kan trække på bredere teorier om, at diskrimination er moralsk problematisk, når og hvis det gør skade. Se Kasper Lippert-Rasmussen, *Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination* (Oxford: Oxford University Press, 2013).

143 Eksemplet er ikke hypotetisk. Se Adewole S. Adamson and Avery Smith, "Machine Learning and Health Care Disparities in Dermatology" *JAMA Dermatology* 154, no. 11 (2018), https://doi.org/10.1001/jamadermatol.2018.2348, https://doi.org/10.1001/jamadermatol.2018.2348.

En anden del af forklaringen kan pege på, at personer kan have forskellig sårbarhed overfor resultatet af en vurdering, og at et system, som ikke tager højde for dette, risikerer at underprioritere præcision for de vigtigste vurderinger. Hvis eksempelvis et automatiseret beslutningssystem vurderer både børn og voksne, og eventuelle fejl i vurderingen udgør en langt større risiko for børn end for voksne, så vil et system, som ikke tager højde for dette, gøre mere skade end et system, som prioriterer at undgå fejl i vurderingen af børn.

Endelig kan en tredje del af forklaringen pege på, at det kan have bredere sociale effekter, hvis et system behandler grupper forskelligt. I hvert fald én bekymring ved det system, som anvendte etnisk herkomst til at vurdere nylediges risiko for langtidsledighed var, at systemet derved risikerede at stigmatisere en i forvejen sårbar gruppe (se afsnit 6.3.1 ovenfor). Tilsvarende kunne et system, som prioriterer ansøgere, og har bias mod kvindelige ansøgere, risikere derved at forstærke sociale kløfter mellem mænd og kvinder, på arbejdspladsen og i samfundet som hele.

Forklaringen om, at algoritmisk bias er moralsk problematisk, når den gør skade, har nogle af de samme styrker og komplikationer som teorien om mulighedsulighed. Ligesom ved mulighedslighed giver forklaringen principielt mulighed for en præcis og utvetydig etisk evaluering af algoritmisk bias. Og ligesom ved mulighedslighed, kan man besvare spørgsmålet om, hvornår et automatiseret beslutningssystem har så meget algoritmisk bias, at vi har grund til ikke at bruge systemet, ved at slutte, at dette er tilfældet, når den algoritmiske bias gør mere skade, end alternative måder at træffe beslutningen på. Svaret på spørgsmålet om, hvordan vi skal prioritere mellem forskellige slags algoritmiske bias bliver tilsvarende, at vi skal vælge de bias, som gør mindst skade. Men fordi forklaringen handler om det moralsk problematiske ved at skade individer, er der, ligesom ved mulighedslighed, ikke nogen simpel sammenhæng mellem algoritmisk bias på gruppeniveau og uetiske automatiserede beslutningssystemer. Det kan også være vanskeligt at fastslå, hvordan en konkret vurdering eller en algoritmisk bias vil påvirke forskellige individer. Fortalere for teorien vil imidlertid, ligesom ovenfor, kunne svare, at dette er vanskeligheder som vi må løse, snarere end grunde til at tro, at teorien skulle være en fejlagtig analyse af hvilke moralske hensyn, som er på spil.

6.6. Et menneske i kredsløbet

I de foregående afsnit har vi diskuteret udfordringer ved automatiserede beslutningssystemer, inklusiv de fejl et system kan begå og de algoritmiske bias, et system kan have. En almindelig måde at forsøge at tackle nogle af de udfordringer, som automatiserede beslutningssystemer kan rejse, er at anbefale eller kræve, at arbejdspladser anvender automatiseret beslutnings*støtte* snarere end fuldt automatiserede beslutninger. Ved automatiseret beslutningsstøtte træffes den endelige beslutning af en menneskelig beslutningstager. Den vurdering, som systemet leverer, indgår i beslutningsgrundlaget, men der er et "menneske i kredsløbet" (eng. "human-in-the-loop").

Et krav om "menneske i kredsløbet" kan motiveres på flere forskellige måder. Et almindeligt argument er, at det gør det lettere at placere et ansvar for beslutningen, og dermed at holde

personer ansvarlige, hvis der bliver truffet en moralsk problematisk beslutning.[144] Et andet argument er, at det kan gøre en forskel for de personer, som bliver berørt af beslutningen, om den er truffet af et menneske eller af en kunstig intelligens. Nogle personer vil måske føle sig mere trygge ved en menneskelig beslutning, eller opleve det som udtryk for manglende anerkendelse, hvis beslutningen træffes fuldautomatiseret.

Et tredje argument hævder, at det i nogle situationer er for risikabelt, at anvende fuldt automatiserede beslutninger. Automatiserede beslutningssystemer vil, som vi har set, realistisk set altid have en vis fejlrate. Især i situationer, hvor beslutningen har store konsekvenser for de berørte, er det vigtigt, at begrænse eller undgå sådanne fejl. Hvis den endelige beslutning overlades til et menneske, så er det muligt for denne person, at identificere og rette systemets fejl. Forhåbningen kan være, at de to systemer supplerer hinanden. Hvert system fungerer som et sikkerhedsnet, der griber forkerte beslutninger – nogle af de fejl, som slipper igennem det ene system, fanges af det andet system, og vice versa. I en nøddeskal er argumentet således, at ved vigtige beslutninger bør der være et menneske i kredsløbet, fordi præcisionen - alt andet lige – er højere ved automatiseret beslutningsstøtte, end ved fuldt automatiserede beslutninger.[145]

Det er imidlertid ikke indlysende, at menneskelige beslutningstagere vil korrigere deres egne fejlagtige vurderinger i lyset af information fra automatiseret beslutningsstøtte, og heller ikke klart, at menneskelige beslutningstagere vil være i stand til at identificere og korrigere de fejl, som et beslutningssystem begår. Det skyldes en kombination af, at menneskelige beslutninger kan være præget af støj, effekten af generelle kognitive bias, og to særlige bias, som kan optræde i forbindelse med anvendelse af automatiserede beslutningssystemer. I dette afsnit diskuterer vi de særlige udfordringer, som automatiseret beslutningsstøtte på arbejdspladsen kan rejse.

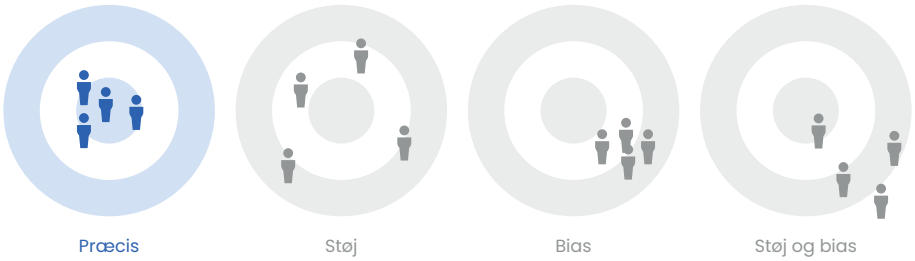
6.6.1. Støj

En væsentlig udfordring for menneskelige beslutninger er, at der ofte er meget støj i beslutningerne. Det betyder, at beslutninger truffet af forskellige mennesker i den samme sag varierer betydeligt, og endda at den samme person kan træffe forskellige beslutninger om

144 Se eksempelvis Christian List, "Group Agency and Artificial Intelligence" *Philosophy & Technology* 34, no. 4 (2021), https://doi.org/10.1007/s13347-021-00454-7, https://doi.org/10.1007/s13347-021-00454-7; MSI-AUT, *A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework*, Council of Europe (2018), https://rm.coe.int/draft-study-of-the-implications-of-advanced-digital-technologies-inclu/16808ef255.

145 Det er i den forbindelse værd at bemærke, at der kan være en tilsvarende begrundelse for at anvende automatiseret beslutningsstøtte, snarere end rent menneskelige beslutninger: Systemet vil være i stand til at korrigere nogle af de fejl, som menneskelige beslutningstagere begår. Denne begrundelse møder en beslægtet udfordring: Hvis menneskelige beslutninger er fejlfrie, hvis systemet til beslutningsstøtte laver de *samme* fejl som menneskelige beslutningstagere, eller hvis menneskelige beslutningstagere aldrig ændrer deres beslutninger i lyset af systemets vurdering, så kan anvendelsen af automatiseret beslutningsstøtte ikke føre til bedre beslutninger, end rent menneskelige beslutninger.

ensartede sager. Beslutninger præget af støj er underlagt en vis grad af tilfældighed, som fører til uens beslutninger, og begrænser deres præcision.[146]



Figur: Effekten af henholdsvis støj og bias illustreret ved evnen til at ramme centrum af en skydeskive.

Udfordringen med støj i menneskelige beslutninger kan illustreres af to slående studier. I et studie fra 2017 under ledelse af den amerikanske datalog John Kleinberg, trænede forskere et automatiseret beslutningssystem til vurderinger i forbindelse med beslutninger om varetægtsfængsling i staten New York, USA.[147] Systemet vurderede risikoen for, at en sigtet ville stikke af eller begå en forbrydelse, hvis vedkommende blev sat på fri fod. Ved at sammenligne systemets vurderinger med data om menneskelige dommeres beslutninger, kunne forskergruppen beregne den overordnede præcision for henholdsvis automatiserede og menneskelige beslutninger. Resultatet var først og fremmest, at det automatiserede beslutningssystem var betragteligt mere præcist. Men forskerne foretog også en økonometrisk analyse af *hvorfor* systemet var mere præcist. Analysen afslørede, at den afgørende forskel ikke var indflydelse fra menneskelige bias, men derimod støj. Vel var der systematiske forskelle på dommernes beslutninger – nogle dommere havde langt højere tendens til at varetægtsfængsle end andre dommere – men på tværs af sådanne forskelle, var den faktor som begrænsede dommernes præcision, at de øjensynligt reagerede på en lang række forhold, som *ikke* havde nogen statistisk forbindelse til risici, og som derfor udgjorde støj i beslutningerne. Støjens effekt var ikke uvæsentlig: Forskergruppen anslog, at anvendelse af det automatiserede beslutningssystem kunne reducere antallet af varetægtsfængslede med 42%, uden at øge risikoen for forbrydelser og undvigelser. Sat på spidsen, så førte støj i de menneskelige beslutninger til, at knap halvdelen af de varetægtsfængslede sad i fængsel uden grund.

Et ældre, klassisk studie demonstrerede om muligt endnu mere slående effekten af støj. Lewis Goldberg undersøgte i slutningen af 1960'erne eksperters evne til at forudsige studerendes

146 Udfordringen med støj i beslutninger er de senere år især fremhævet af adfærdsforskerne Daniel Kahneman, Olivier Sibony og Cass Sunstein. Se Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein, *Noise: A Flaw in Human Judgment* (London: William Collins, 2021).

147 Jon Kleinberg et al., "Human Decisions and Machine Predictions" *NBER Working paper series* (2017), http://www.nber.org/papers/w23180.

akademiske præstation.[148] På baggrund af et stort datasæt med eksperternes vurderinger udviklede Goldberg statistiske modeller, som kunne forudsige eksperternes vurdering af de studerendes præstation. Målet for disse modeller var altså ikke at forudsige hvordan de studerende klarede sig, men at forudsige hvordan hver af eksperterne ville vurdere de studerende, også når disse vurderinger var fejlagtige. Studiet viste for det første, at det var muligt at lave modeller, som med rimelig høj præcision kunne forudsige, hvordan eksperterne ville vurdere en studerende. Men Goldbergs studie undersøgte også et andet spørgsmål: Hvad var mest præcist til at vurdere de studerende – eksperterne, eller de modeller som forsøgte at forudsige eksperternes vurderinger?

Det slående resultat var, at modellerne var bedre til at vurdere de studerende end eksperterne, selvom modellerne var trænet til at imitere eksperternes vurdering. Det skyldtes, at de enkle modeller anvendte data konsistent, og derved eliminerede den støj, som påvirkede eksperternes vurderinger.[149]

I forbindelse med automatiseret beslutningsstøtte er udfordringen med beslutninger præget af støj ikke blot at støj skaber fejl, men også at støj gør det vanskeligt, at identificere de situationer, hvor menneskelige beslutninger er fejlagtige. Hvis en menneskelig vurdering og vurderingen fra et automatiseret beslutningssystem er forskellige, så skal den menneskelige beslutningstager afgøre, hvilken af de to vurderinger som er fejlagtig. Når en vurdering er præget af bias, så vil det ofte være muligt at identificere de situationer, hvor vurderingen er usikker. Når vurderinger er præget af støj, så optræder fejlene derimod usystematisk. Støj gør det altså vanskeligt at vide, hvornår man kan og ikke kan stole på den menneskelige vurdering.

6.6.2. Automatiserings- og algoritmisk aversionsbias

To særlige bias kan også spille en vigtig rolle for kvaliteten af automatiseret beslutningsstøtte: Automatiseringsbias og algoritmisk aversionsbias.

Automatiseringsbias er en psykologisk tendens til at have for høj tillid til automatiserede beslutningssystemer, og til at foretage hurtigere, mindre grundige vurderinger af en opgave, som et sådant system har behandlet, end man egentlig burde. Automatiseringsbias er veldokumenteret,

Automatiseringsbias:

En psykologisk tendens til at lægge for meget vægt på et automatiseret beslutningsstøttesystems vurderinger.

148 Lewis Goldberg, "Man Versus Model of Man: A Rationale, plus some Evidence, for a Method of Improving on Clinical Inferences" *Psychological Bulletin* 73, no. 6 (1970).

149 Resultatet er bekræftet siden hen. Kahneman, Sibony og Sunstein refererer eksempelvis et endnu mere radikalt resultat fra et nyere studie, som sammenligner HR-eksperters vurdering af kandidater til en toplederstilling med *tilfældigt* trænede modeller. Studiet demonstrerer, at selv disse modellers vurderinger er bedre end de menneskelige eksperters, fordi modellerne ikke er præget af støj. Se Martin C. Yu and Nathan R. Kuncel, "Pushing the Limits for Judgemental Consistency: Comparing Random Weighting Schemes with Expert Judgments" *Personal Assessments and Decisions* 6, no. 2 (2020).

især i sundhedsvæsenet og for piloter.[150] Risikoen ved automatiseringsbias er, at menneskelige beslutningstagere ikke, eller kun i begrænset omfang, korrigerer de fejl, som beslutningssystemet begår. I værste fald bliver automatiseret beslutningsstøtte under indflydelse af automatiseringsbias til en mindre effektiv fuldt automatiseret beslutning i forklædning.

Algoritmisk aversionsbias:

En psykologisk tendens til, at lægge for lidt vægt på et automatiseret beslutningsstøttesystems vurderinger.

I visse situationer kan beslutningstagere i stedet ligge under for en anden bias, som groft sagt har den modsatte effekt. Såkaldt algoritmiske aversionsbias optræder i situationer, hvor personer irrationelt tillægger et systems vurderinger for *lidt* vægt.[151] Algoritmisk aversionsbias udløses typisk, når beslutningstageren har førstehåndskendskab til, at systemet er fejlbarligt. I typiske eksempler ignorerer en beslutningstager systemets vurdering, og anvender alene sin selvstændige vurdering til at træffe en beslutning, selvom systemets vurderinger dokumenterbart er mere præcise, og beslutningstageren ved at dette er tilfældet. Algoritmisk aversionsbias kan således føre til, at menneskelige beslutningstagere ikke, eller kun i begrænset omfang, bruger information fra den automatiserede beslutningsstøtte til at korrigere de fejl, som de selv begår.

6.6.3. Konfirmationsbias og ankereffekter i automatiseret beslutningsstøtte

Vi har tidligere nævnt det faktum, at menneskelige beslutninger ofte kan være præget af kognitive bias (se afsnit 6.3). Den tredje udfordring knytter sig til, hvordan to af de mest velkendte sådanne bias kan øve indflydelse på beslutningen i automatiseret beslutningsstøtte.

Når en arbejdsplads anvender automatiseret beslutningsstøtte, kan det groft sagt foregå på to måder. Beslutningstageren kan vurdere sagen *inden* vedkommende får information om systemets vurdering, eller *efter* at vedkommende får information om systemets vurdering. I det første tilfælde kan beslutningstageren justere sin vurdering i lyset af den nye information, som systemets vurdering udgør. I det andet tilfælde kan beslutningstageren anvende systemets vurdering som en del af grundlaget for sin vurdering.

---

150 D. Lyell and E. Coiera, "Automation bias and verification complexity: a systematic review" *Journal of the American Medical Informatics Association* 24, no. 2 (2017), https://doi.org/10.1093/jamia/ocw105; Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt, "Automation bias: a systematic review of frequency, effect mediators, and mitigators" *Journal of the American Medical Informatics Association* 19, no. 1 (2011), https://doi.org/10.1136/amiajnl-2011-000089 %J *Journal of the American Medical Informatics Association*, https://doi.org/10.1136/amiajnl-2011-000089.

151 Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey, "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err" *Journal of Experimental Psychology: General* 144, no. 1 (2015); Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey, "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them" *Management Science* 64, no. 3 (2018), https://doi.org/10.1287/mnsc.2016.2643, https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2016.2643; Andrew Prahl and Lyn Van Swol, "Understanding algorithm aversion: When is advice from automation discounted?" *Journal of Forecasting* 36, no. 6 (2017), https://doi.org/10.1002/for.2464, https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2464; Jason W. Burton, Mari-Klara Stein, and Tina Blegind Jensen, "A systematic review of algorithm aversion in augmented decision making" 33, no. 2 (2020), https://doi.org/10.1002/bdm.2155, https://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.2155.

Hvis beslutningstageren vurderer sagen *før* vedkommende får indsigt i systemets vurdering, så risikerer dette at udløse konfirmationsbias. Konfirmationsbias er en ubevidst tendens til at tillægge information forskellig vægt, afhængigt af om informationen understøtter eller strider mod en forudgående overbevisning.[152]

En konfirmationsbias kan i forbindelse med automatiseret beslutningsstøtte bestå i, at systemets vurdering tillægges forskellig vægt afhængigt af, om den bekræfter eller strider mod beslutningstagerens indledende vurdering. Den kan derfor føre til, at beslutningstagere ikke i tilstrækkelig grad korrigerer deres fejlagtige vurderinger i lyset af systemets vurdering.

Konfirmationsbias:

En psykologisk tendens til at give ny information forskellig vægt, afhængigt af, om den bekræfter eller strider mod eksisterende synspunkter.

Når man risikerer at udløse konfirmationsbias, ved at give beslutningstagere systemets vurdering *efter* at de har formet deres egen, så kan det være fristende at konkludere, at vi bør foretrække den anden tilgang, hvor beslutningstageren får systemets vurdering *inden* vedkommende foretager sin egen vurdering. Udfordringen for den anden tilgang er, at den risikerer at skabe en anker-effekt.

En anker-effekt er en ubevidst tendens til at lade en information påvirke en vurdering, alene fordi den pågældende information optræder i bevidstheden umiddelbart inden man foretager vurderingen. Anker-effekter kan optræde selvom man ved, at den pågældende information er komplet irrelevant for den vurdering, man skal foretage, men kan naturligvis også forekomme, når der er en tæt relation mellem information og vurdering.[153]

Anker-effekt:

En psykologisk tendens til at lægge for meget vægt på information, alene fordi denne information optræder i bevidstheden, når man skal foretage en vurdering.

Det virker oplagt, at hvis en arbejdsplads anvender automatiseret beslutningsstøtte, og giver systemets vurdering til beslutningstageren *inden* vedkommende har foretaget sin egen vurdering, så kan det udløse en anker-effekt. En anker-effekt vil i så fald bestå i, at systemets vurdering ubevidst tillægges vægt på en måde, så den former den menneskelige vurdering langt mere, end den burde, og mere end beslutningstageren er sig bevidst. Det kan føre til,

152 Konfirmationsbias er veldokumenteret i mange forskellige sammenhænge, men er i de senere år blandt andet blevet studeret og dokumenteret i retsvæsenet. Se Eric Rassin, "Context effect and confirmation bias in criminal fact finding" 25, no. 2 (2020), https://doi.org/10.1111/lcrp.12172, https://onlinelibrary.wiley.com/doi/abs/10.1111/lcrp.12172; Moa Lidén, Minna Gräns, and Peter Juslin, "'Guilty, no doubt': detention provoking confirmation bias in judges' guilt assessments and debiasing techniques" *Psychology, Crime & Law* 25, no. 3 (2019), https://doi.org/10.1080/1068316X.2018.1511790, https://doi.org/10.1080/1068316X.2018.1511790. For en klassisk introduktion til fænomenet, se Raymond S Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises" *Review of general psychology* 2, no. 2 (1998).

153 For et klassisk eksempel fra retsvæsenet, se Birte Englich and Thomas Mussweiler, "Sentencing under uncertainty: Anchoring effects in the courtroom" *Journal of Applied Social Psychology* 31 (2001).

at mennesker kun i begrænset omfang korrigerer de fejl, som det automatiserede beslutningssystem begår.

6.6.4. Det værste af begge verdener

Når menneskelige beslutninger kan være præget af støj og kognitive bias, og automatiseret beslutningsstøtte kan udløse automatiserings- og algoritmisk aversionsbias, så kunne man håbe at trøste sig med, at menneskelige beslutninger trods alt undgår de udfordringer med algoritmisk bias, som automatiserede beslutningssystemer møder (se afsnit 6.4.4). Det ville imidlertid være forlorent håb. Menneskelige beslutninger, som anvender de samme data til at træffe den samme beslutning som et automatiseret beslutningssystem, vil møde de samme udfordringer. Det skyldes, at algoritmisk bias er en effekt af matematiske sammenhænge mellem de anvendte data og vurderingen, som optræder uanset om det er et menneske eller et automatiseret beslutningssystem, som foretager en vurdering.

Ideelt fører automatiseret beslutningsstøtte til færre fejlagtige beslutninger. I lyset af de udfordringer, som automatiseret beslutningsstøtte kan møde, tyder meget imidlertid på, at det er et åbent spørgsmål, om automatiseret beslutningsstøtte vil have denne fordel i en given situation. I ekstreme tilfælde kunne beslutningsstøtte tænkes, at kombinere det værste af begge verdener. Hvis mennesker har tendens til at fastholde deres egen vurdering i netop de tilfælde, hvor mennesker begår fejl, og tendens til at ændre deres vurdering i netop de tilfælde, hvor beslutningssystemer begår fejl, så vil automatiseret beslutningsstøtte føre til *flere* fejl, end både rent menneskelige beslutninger og fuldt automatiserede beslutninger.[154] I en sådan situation supplerer mennesker og automatiserede beslutningssystemer ikke hinanden som to sikkerhedsnet, der fanger hver sit sæt fejl. De supplerer snarere hinanden som to huller i en båd, hvor forskellige slags fejl kan strømme igennem.

Der kan, som indledningsvis nævnt, være flere forskellige grunde til at foretrække automatiseret beslutningsstøtte frem for automatiserede beslutninger. Men om automatiseret beslutningsstøtte fører til flere fejl, færre fejl, de samme fejl eller andre fejl end alternativerne, er i sidste ende et empirisk spørgsmål, som det kan være relevant at afklare, ved at teste hvad fejlraten er for henholdsvis menneskelige beslutninger, automatiserede beslutninger, og automatiseret beslutningsstøtte.

154 Teknisk set er det tilstrækkeligt at en af disse tendenser optræder, og at den er stærkere end en eventuel tendens, som fører til færre fejl. Det kunne godt være tilfældet eksempelvis, at mennesker har tendens til at korrigere deres vurdering i de tilfælde, hvor mennesker begår fejl og beslutningssystemer rammer rigtigt. Det vil i givet fald føre til færre fejl. Men hvis mennesker også har tendens til at ændre deres vurdering i de tilfælde, hvor mennesker rammer rigtigt mens beslutningssystemer begår fejl, og denne tendens er stærkere end den første, så vil den samlede effekt være, at automatiseret beslutningsstøtte fører til flere fejl.

7. Tre illustrationer af dataetiske udfordringer ved indsamling og anvendelse af medarbejderdata på arbejdspladsen

I dette kapitel illustrerer vi, hvordan de dataetiske hensyn, som vi har introduceret i de tidligere kapitler, kan informere en dataetisk analyse. Vi præsenterer tre eksempler, hvor en arbejdsplads indsamler og anvender medarbejderdata. For hvert eksempel diskuterer vi nogle af de dataetiske udfordringer, man bør være opmærksom på, og hvordan man kan forstå dem i lyset af de indsigter, som vi har udviklet i løbet af denne rapport.

7.1. Opgaveløsning

En lagerhal anvender et digitalt system til at fordele pakningsopgaver til medarbejderne, og registrerer i den forbindelse hvilke opgaver, den enkelte medarbejder løser. Arbejdspladsen udvikler på baggrund af disse data et måltal for, hvor mange pakningsopgaver en medarbejder forventes at løse på en arbejdstime. Dette måltal sammenlignes med data om hver enkelt medarbejders aktuelle effektivitet. Disse data annonceres på en storskærm i lagerhallen. Formålet er at forbedre effektiviteten, ved at gøre det klart for medarbejdere, hvordan deres arbejdsindsats svarer til arbejdspladsens forventninger.

Opgaveløsning er et eksempel på en arbejdsplads, som indsamler og analyserer medarbejderdata for at måle medarbejderes præstation, og forsøger at bruge disse præstationsmålinger som direkte incitament til at forbedre arbejdsindsatsen. Ved at koble indsamling af medarbejderdata direkte til et stærkt incitament skaber arbejdspladsen risiko for observationsstress og gaming-effekter, og øger omkostningerne ved de fejl, systemet kan risikere at begå.

7.1.1. Observationsstress

I *Opgaveløsning* skal medarbejdere konstant forholde sig til målinger af hver enkelt medarbejders arbejdsindsats. Disse oplysninger deles offentligt, således at den enkelte medarbejder kan sammenligne sin præstation med andre medarbejderes, men hver medarbejder ved også, at både ledelse og kollegaer kan sammenligne vedkommendes præstation med andres. Denne bevidsthed om, at arbejdsindsats vil blive kendt og sammenlignet på tværs af arbejdspladsen kan i sig selv skabe et stærkt psykologisk pres for at præstere. Samtidig kan medarbejdere frygte, hvilke konsekvenser det vil have, hvis de vurderes til at præstere under gennemsnittet. Selv hvis der ikke findes en officiel politik om, at medarbejderes præstation vil have betydning for eksempelvis lønforhandlinger eller afskedigelser, så kan medarbejdere have grund til at antage, at de kan spille en sådan rolle. Denne forventning kan forstærke det psykologiske pres for at præstere.

Det psykologiske pres for at præstere, som arbejdspladsen skaber, kan påvirke oplevelsen af, at få indsamlet medarbejderdata. Arbejdspladsen indsamler konsekvent data om medarbejderes løsning af opgaver på lagerhallen. Kombinationen af konstant dataindsamling og det psykologiske pres fra anvendelsen af de indsamlede data kan skabe observationsstress, hvor medarbejdere føler sig overvåget på en måde, der negativt påvirker deres mentale og fysiske helbred.

7.1.2. Fejl i målingen af præstationer

En anden udfordring knytter sig til, at præstationsmålingen kan være misvisende, fordi de data lagerhallen bruger som målestok for præstation, ikke præcist repræsenterer medarbejderes indsats. I *Opgaveløsning* tager målingen eksempelvis ikke højde for, at nogle pakningsopgaver kan være svære, mens andre kan være lette. En pakke kan ligge forskellige steder i lagerhallen, som det kan tage kortere eller længere tid at komme til, og den kan variere i størrelse, vægt, og hvordan den skal håndteres, for eksempel hvis den er særligt skrøbelig.

Lagerhallen anvender et automatiseret beslutningssystem til at fordele opgaver til medarbejderne. Hvis systemet fordeler opgaverne tilfældigt, kan der derfor opstå tilfældige forskelle i medarbejdernes målte præstation, som skyldes tilfældige forskelle i hvilke opgaver, de er blevet tildelt. En medarbejder, som har været uheldig at blive tildelt en stribe vanskelige opgaver, vil ufortjent blive vurderet som dårligere præsterende, end en kollega, som har været mere heldig med fordelingen af opgaver. Det kan i sig selv være ubehageligt for en medarbejder, at blive udstillet på arbejdspladsen, som en der præsterer dårligt, men i situationer hvor det er ufortjent, vil mange opleve det som grundlæggende uretfærdigt og frustrerende. Sådanne fejl kan have endnu større konsekvenser, hvis præstationsmålingen har indflydelse på ansættelsesforhold som løn og afskedigelse.

Situationen kan være endnu værre, hvis opgaverne ikke fordeles tilfældigt, fordi det betyder at der kan opstå systematiske forskelle i, hvilke opgaver medarbejdere tildeles. Det kan for eksempel være tilfældet, hvis lagerhallen arbejder i vagthold, og der er forskelle i hvilke opgaver, medarbejdere behandler på forskellige tidspunkter af døgnet, eller hvis medarbejdere behandler opgaver fra forskellige grupper kunder. Nogle medarbejder kan i den situation opleve, at de systematisk får sværere eller lettere opgaver, end deres kollegaer. Hvis målingen ikke tager højde for, at nogle opgaver er lettere, mens andre er mere tidskrævende, vil systemet systematisk vurdere nogle medarbejderes præstation højere, og andre medarbejderes præstation lavere, alene fordi der konsekvent er forskel på de opgaver, som medarbejderne bliver tildelt.

7.1.3. Gaming-effekter

De stærke incitamenter, som arbejdspladsen skaber ved at indsamle og anvende data i Opgaveløsning, risikerer også at skabe gaming-effekter, hvor medarbejdere tilpasser deres adfærd til de data som måles, snarere end til arbejdspladsens egentlige mål.

Hvis eksempelvis en pakke er fejlmærket, så kan medarbejderen vælge mellem at afsende den forkerte pakke til kunden, og få lukket en arbejdsopgave, eller at bruge tid på at rette

mærkningen, finde den rigtige pakke, opdatere systemet, og lukke opgaven med afsendelse af den rigtige pakke. For lagerhallen er den sidste løsning klart at foretrække – det betyder, at kunden modtager den korrekte pakke – men for medarbejderen skaber præstationsmålinger et stærkt incitament til at vælge det første – derved løser medarbejderen på papiret en opgave langt hurtigere. I værste fald kan incitamentet motivere medarbejdere til at snyde. Hvis eksempelvis en bestemt pakningsopgave er meget tidskrævende at udføre, kan medarbejdere være fristet til at registrere pakken som bortkommet og annullere opgaven, for at kunne fokusere på opgaver, som er hurtigere at løse.

For lagerhallen kan der være mange situationer, hvor medarbejdere må vælge mellem at udføre opgaver bedst muligt, eller at udføre flest mulig opgaver. Når systemet alene måler på antallet af udførte opgaver, vil medarbejdere have en stærk tilskyndelse til at "game", for at få den bedst mulige præstationsmåling. Arbejdspladsen kan forsøge at håndtere denne udfordring, ved at indføre mere nuancerede præstationsmålinger, men det kan være vanskeligt at måle præstationer på en måde, som fuldstændig undgår at skabe de skæve incitamenter, der danner grundlag for gaming-effekter.

7.2. Rekruttering

En telemarketingvirksomhed anvender maskinlæring til at analysere sammenhænge mellem målinger af tidligere og nuværende ansattes præstation på den ene side, og deres historiske ansøgninger til jobbet på den anden side. Analysen udvikler en statistisk model for, hvilke træk ved en ansøgning, som karakteriserer medarbejdere, som efterfølgende klarer sig henholdsvis mere og mindre godt. Et automatiseret beslutningssystem anvender modellen til at vurdere ansøgninger til nye opslåede stillinger, og prioriterer dem afhængigt af, hvordan modellen statistisk forudsiger, at de vil klare sig. Arbejdspladsens HR-afdeling bruger disse prioriteringer, når de vælger hvilke ansøgere, som skal kaldes til samtale.

Rekruttering er et eksempel på en virksomhed, som benytter sig af medarbejderdata og automatiseret beslutningsstøtte for at forbedre virksomhedens evne til at ansætte de bedst kvalificerede medarbejdere. Ved at træne på historiske data risikerer virksomheden imidlertid, at skabe moralsk problematisk algoritmisk bias, og ved at anvende beslutningsstøtte risikerer den, at skabe automatiseringsbias.

7.2.1. Algoritmisk bias

Virksomhedens formål med at anvende automatiseret beslutningsstøtte i *Rekruttering* er, at blive bedre til at ansætte de bedste ansøgere. En motivation kan i den sammenhæng være, at reducere effekten af menneskelige bias og støj i beslutningerne. Virksomheden kan håbe, at systemet kan hjælpe med at finde frem til højt kvalificerede ansøgere, som ellers ville blive overset. En vigtig risiko i den forbindelse er, at virksomheden træner systemet på sine historiske data for ansøgninger og ansættelser. Det betyder, at systemet uundgåeligt vil være begrænset af de beslutninger, som tidligere er blevet truffet, inklusive de bias, som har påvirket sådanne beslutninger.

Hvis tidligere ansættelser eksempelvis har været præget af en præference for mandlige ansøgere, eller ansøgere med etnisk majoritetsbaggrund, så kan systemet lære, at de bedste medarbejdere har haft træk, som karakteriserer disse grupper. Det vil give systemet tendens til at prioritere sådanne ansøgere, og derved reproducere de bias, som præger træningsdata.

Der kan også være forskelle mellem grupperne, som påvirker hvordan de bør vurderes. Det kan eksempelvis være sådan, at de træk ved en ansøgning, som karakteriserer dygtige mandlige medarbejdere, i mindre grad karakteriserer dygtige kvindelige medarbejdere, mens der omvendt er træk ved en ansøgning som i høj grad karakteriserer dygtige kvindelige medarbejdere, men i mindre grad dygtige mandlige medarbejdere. Hvis mandlige medarbejdere er overrepræsenterede i systemets træningsdata, fordi flertallet af medarbejderne historisk har været mænd, så kan systemet blive bedre til at lære at vurdere mandlige ansøgere, ved at bruge de træk ved ansøgninger, som karakteriserer dygtige mandlige medarbejdere. Omvendt kan systemet få svært ved at vurdere kvindelige ansøgere, fordi det ikke har data til at lære at bruge de træk ved en ansøgning, som i højere grad karakteriserer dygtige kvindelige medarbejdere. Resultat vil være en tendens til lavere præcision for kvindelige ansøgere end for mandlige medarbejdere. En sådan forskel i præcision kan i sig selv være problematisk, fordi den giver mandlige og kvindelige ansøgere forskellige muligheder for at opnå ansættelse, men den kan også give virksomheden incitament til at fravælge kvindelige ansøgere, fordi vurderinger af kvindelige ansøgere viser sig, at være mindre pålidelige.

7.2.2. Automatiseringsbias

Telemarketingvirksomheden har udviklet et beslutningsstøttesystem, for at hjælpe med at finde de bedste ansøgere. Virksomheden har valgt at bruge systemets vurdering som information til den leder, som skal træffe beslutningen – der er et "menneske i kredsløbet" – snarere end at lade systemets vurdering afgøre, hvilke ansøgere virksomheden ansætter. Denne beslutning kan være motiveret af en formodning om, at mennesker vil være i stand til at korrigere nogle af de fejl, som systemet laver, når det skal vurdere ansøgere. En afgørende udfordring i den forbindelse er, at mennesker kan have tendens til at lægge for meget vægt, på systemets vurdering – såkaldt automatiseringsbias. Når det er tilfældet, så tror virksomheden, at den bruger beslutningsstøtte, og at de fejl, som systemet begår vil blive fanget, men i praksis kan beslutningerne være så godt som fuldt automatiserede.

Det er i den forbindelse værd at holde sig for øje, hvordan beslutningsstøtte kan virke forskelligt på menneskelige beslutninger, afhængigt af, om systemet har prioriteret ansøgeren højt eller lavt. Når systemet prioriterer en ansøger lavt, kan det være en svær beslutning, at trodse systemet. Hvis det viser sig, at medarbejderen præsterer dårligt, så har man ansat vedkommende *på trods* af systemets lave vurdering. Når systemet prioriterer en ansøger højt, kan det være lettere, at trodse systemet, og nægte vedkommende ansættelse. Det skyldes, at virksomheden ikke kan måle præstation for de ansøgere, som den vælger ikke at ansætte. En beslutning om *ikke* at ansætte en ansøger, er derfor ikke en beslutning, som umiddelbart kan registreres som en fejl, selv hvis man fravælger en ansøger, som ville være blevet til en fantastisk medarbejder.

De ulige incitamenter, som beslutningsstøtte kan skabe, kan spille sammen med algoritmiske og menneskelige bias. Hvis systemet for eksempel har bias på den måde, at det har

tendens til at foretrække mandlige ansøgere, kan det være svært for den menneskelige beslutningstager, at trodse systemets vurdering, og ansætte en kvindelig ansøger, som er blevet vurderet lavt. Omvendt vil en menneskelig beslutningstager, som selv har bias til fordel for mænd, uden væsentlig risiko kunne trodse systemets vurdering af en kvindelig ansøger, som er blevet vurderet højt.

7.3. Trivsel

En kommune indsamler data om hvilke arbejdsopgaver den enkelte medarbejder får tildelt og hvordan opgaverne løses, herunder antallet og længden af møder, antallet af indkommende og udgående e-mails, antallet og længden af telefonsamtaler, og antallet og længden af arbejdssessioner i forskellige systemer. Et automatiseret beslutningssystem anvender disse data til at vurdere medarbejderes trivsel, og forsøger at identificere medarbejdere som er i risiko for stress, udbrændthed og lav motivation. Disse medarbejdere modtager besked om, at de er blevet vurderet som værende i risiko, og at deres leder er blevet gjort opmærksom på risikoen.

Trivsel er et eksempel på en arbejdsplads, der anvender et automatiseret beslutningssystem til at identificere medarbejdere, som har forhøjet risiko for stress og beslægtede former for mistrivsel. Formålet er, at gøre det lettere for en leder at gribe ind, og hjælpe medarbejderen, for eksempel ved at ændre på arbejdsopgaver eller arbejdsprocesser, før stress eller mistrivsel får alvorlige konsekvenser. De gode hensigter til trods risikerer kommunens indsamling og anvendelse af medarbejderdata, at rejse mange af de dataetiske udfordringer, som vi har set i de foregående eksempler, inklusive risici for fejl, observationsstress, gaming-effekter, og algoritmisk bias. Oveni disse udfordringer kan der i *Trivsel* være grund til skepsis overfor, om kommunens medarbejdere er i stand til, at afgive et frit samtykke til indsamling og anvendelse af medarbejderdata.

7.3.1. Fejl, stress, gaming, og bias

Beslutningssystemet i *Trivsel* benytter en række forskellige medarbejderdata, som hver for sig kun i begrænset omfang hænger sammen med risiko for mistrivsel. Samtidig vil mange faktorers betydning variere fra medarbejder til medarbejder, på måder der er vanskelige at måle, og der må forventes at være risikofaktorer, som systemet ikke har data på. Systemet vil derfor forventeligt begå en vis mængde fejl. Det kan dels klassificere nogle medarbejdere som havende høj risiko, selvom disse medarbejdere faktisk trives fint (falsk positiv). Det kan også overse nogle medarbejdere, som faktisk mistrives, og klassificere dem som havende lav risiko (falsk negativ). Begge typer fejl kan være problematiske. En medarbejder som trives fint, kan opleve det som stødende, at blive klassificeret som i høj risiko for mistrivsel. Medarbejderen kan også være bekymret for, at denne vurdering vil gøre et negativt indtryk på ledelsen, som kan have konsekvenser for medarbejderens karriere. Omvendt kan de medarbejdere, som overses af systemet, mistrives uden at få den hjælp, som de har brug for. Ligesom i *Rekruttering* kan denne risiko forstærkes af automatiseringsbias, hvor ledelsen kommer til at lægge for

stor vægt på systemets vurderinger – en leder kan tænke, at en medarbejder godt nok virker presset, men at systemet jo har vurderet risikoen for stress og mistrivsel som lav.

Arbejdspladsen i *Trivsel* indsamler store mængder detaljerede data om medarbejderne, herunder hvilke opgaver de udfører, og hvordan. Medarbejdere kan berettiget have en oplevelse af, at hver eneste arbejdsopgave udføres under detaljeret opsyn af ledelsen. Medarbejdere kan også have en bekymring – begrundet eller ej – for at de indsamlede data vil blive anvendt til andre formål end trivselsmålinger, for eksempel vurdering af medarbejderes præstation. Ligesom i *Opgaveløsning* kan denne omfattende dataindsamling skabe risiko for, at medarbejdere føler sig overvåget, på en måde der i sig selv skaber observationsstress. Paradoksalt risikerer arbejdspladsens indsamling og anvendelse af data, at skabe netop den mistrivsel, som det var intentionen, at den skulle hjælpe med at opdage og forebygge.

Indsamling og anvendelse af medarbejderdata risikerer også at flytte medarbejderes motivation og fokus fra varetagelse af arbejdsopgaver til optimering af den måde, medarbejderen fremstår i de indsamlede data. Hvis medarbejdere oplever indsamlingen som stressende overvågning, eller er bekymrede for, at en klassificering som høj risiko for stress og mistrivsel kan have negative konsekvenser, vil de have et incitament til, at ændre adfærd på en måde, som reducerer stress eller sandsynligheden for at blive klassificeret som høj risiko. Kommunen skaber altså, ligesom i *Opgaveløsning*, risiko for gaming-effekter. Medarbejdere kan eksempelvis forsøge at tilrettelægge mest muligt af deres arbejde på en måde, hvor der ikke kan samles data, for eksempel ved at afholde uformelle fysiske møder, som hverken registreres som e-mail, telefonsamtaler eller kalenderaftaler. Denne uformelle mødeform kan på engang være mindre effektiv end de typer kommunikation den erstatter, og underminere kommunens muligheder for, at måle og forebygge mistrivsel.

Endelig kan beslutningsstøttesystemet, ligesom i *Rekruttering*, have algoritmisk bias. Det kan opstå som en konsekvens af bias i de data, systemet er trænet på, eller forskelle i hvordan data hænger sammen med trivsel på tværs af relevante grupper. Hvis eksempelvis mandlige ansatte historisk har været mere tilbøjelige end kvindelige ansatte, til at dække over stress og mistrivsel, så kan denne forskel afspejle sig i hvordan systemet vurderer stress og mistrivsel. Systemet kan blive mere tilbøjeligt til at klassificere kvindelige medarbejdere som havende høj risiko, eller få lavere præcision, når det vurderer mandlige medarbejdere. Og hvis der er forskel på hvilke stillinger mænd og kvinder typisk varetager, og hvordan de indsamlede medarbejderdata hænger sammen med mistrivsel på tværs af disse stillinger, så kan systemet virke forskelligt for de to grupper.

7.3.2. Frit eller ufrit samtykke

Hvis kommunen i *Trivsel* indsamler medarbejderdata, og derved skaber observationsstress, eller gør medarbejderne sårbare for fejl og algoritmiske bias, som skaber ulighed i muligheder eller skader medarbejdere, så er der dataetiske grunde til at beskytte medarbejdernes privatliv. Man kan tale om, at medarbejderne i disse tilfælde har en moralsk ret til privatliv, og at kommunens indsamling og anvendelse af medarbejderdata er moralsk problematisk. I den situation kan arbejdspladsen ønske at legitimere indsamling og anvendelse af medarbejderdata, ved at indhente samtykke fra medarbejderne.

Hvis et samtykke skal gøre en moralsk forskel, så skal det være autonomt, informeret og frit. Kommunens medarbejdere kan, hvis arbejdspladsen informerer grundigt, indfri de første to betingelser. Det kan være mere vanskeligt, at indfri den tredje betingelse, hvis kommunen eksempelvis gør samtykke til indsamling og anvendelse af medarbejderdata til en forudsætning for ansættelse. I den situation tvinges medarbejdere til at vælge mellem samtykke til indsamling af data, og det økonomiske, sociale og identitetsmæssige tab, som det normalt udgør, at miste sit arbejde. Hvis arbejdspladsen lægger så massivt pres på medarbejderes beslutning, kan samtykket dårligt betegnes som frit. Men selv hvis kommunen ikke eksplicit knytter sanktioner til afvisning af samtykke, kan medarbejdere nære mistanke om, at det alligevel kan have negative konsekvenser. Disse bekymringer kan i sig selv lægge pres på medarbejdernes beslutning i en grad, så samtykket bliver ufrit og derfor moralsk irrelevant. Hvis den konkrete eller hypotetiske trussel om negative konsekvenser gør samtykke ufrit, så kan kommunen ikke retfærdiggøre indsamling og anvendelse af medarbejderdata ved at henvise til medarbejdernes samtykke. I den situation må de dataetiske udfordringer, som kommunens system rejser, håndteres ved at sikre en tilstrækkelig balance mellem de dataetiske grunde som taler for og imod indsamling og anvendelse af medarbejderdata.

Litteratur

Adams-Prassl, Jeremias. "Regulating Algorithms at Work: Lessons for a 'European Approach to Artificial Intelligence'." *European Labour Law Journal* 13, no. 1 (2022): 30-50.

Adams-Prassl, Jeremias, Reuben Binns, and Aislinn Kelly-Lyth. "Directly Discriminatory Algorithms." *The Modern Law Review* 86, no. 1 (2023): 144-75. https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-2230.12759.

Adamson, Adewole S., and Avery Smith. "Machine Learning and Health Care Disparities in Dermatology." *JAMA Dermatology* 154, no. 11 (2018): 1247-48. https://doi.org/10.1001/jamadermatol.2018.2348.

AI Now Institute. *Algorithmic Management: Restraining Workplace Surveillance.* (11 April 2023). https://ainowinstitute.org/publication/algorithmic-management.

Aiello, John R., and Kathryn J. Kolb. "Electronic Performance Monitoring and Social Context: Impact on Productivity and Stress." *Journal of Applied Psychology* 80 (1995): 339-53. https://doi.org/10.1037/0021-9010.80.3.339.

Ajunwa, Ifeoma, Kate Crawford, and Jason Schultz. "Limitless Worker Surveillance." *California Law Review* (2017): 735-76.

Akhtar, Marya, Frej Klem Thomsen, Rikke Frank Jørgensen, and Pernille Boye Koch. Når Algoritmer Sagsbehandler. Institut for Menneskerettigheder (2021). https://menneskeret.dk/files/media/document/Algoritmer_8.K.pdf.

Alexandrie, Gustav. "Surveillance Cameras and Crime: A Review of Randomized and Natural Experiments." *Journal of Scandinavian Studies in Criminology and Crime Prevention* 18, no. 2 (2017): 210-22. https://doi.org/10.1080/14043858.2017.1387410.

Algoritmer Data og Demokrati-projektet. *Ansvarlig Og Værdiskabende Anvendelse Af Medarbejderdata - Anbefalinger Til Den Digitaliserede Arbejdsplads.* (2023). https://taenketanken.mm.dk/wp-content/uploads/2023/11/Anbefalingskatalog_Ansvarlig-og-vaerdiskabende-anvendelse-af-medarbejderdata.pdf.

Altman, M., A. Wood, and E. Vayena. "A Harm-Reduction Framework for Algorithmic Fairness." *IEEE Security & Privacy* 16, no. 3 (2018): 34-45. https://doi.org/10.1109/MSP.2018.2701149.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *ProPublica.* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arneson, Richard J. "Equality and Equal Opportunity for Welfare." *Philosophical Studies* 56, no. 1 (1989): 77 - 93.

Article 29 Data Protection Working Party. *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679.* (2017). https://ec.europa.eu/newsroom/article29/redirection/document/49826.

————.*Opinion 2/2017 on Data Processing at Work.* (2017). https://ec.europa.eu/newsroom/article29/redirection/document/45631.

————.*Opinion 06/2014 on the Notion of Legitimate Interests of the Data Controller under Article 7 of Directive 95/46/Ec.* (2014). https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf.

Asgarinia, Haleh. "Convergence of the Source Control and Actual Access Accounts of Privacy." *AI and Ethics* (2023). https://doi.org/10.1007/s43681-023-00270-z.

Bacchini, Fabio, and Ludovica Lorusso. "Race, Again: How Face Recognition Technology Reinforces Racial Discrimination." *Journal of Information, Communication and Ethics in Society* 17, no. 3 (2019): 321-35. https://doi.org/10.1108/JICES-05-2018-0050.

Backhaus, N. "Context Sensitive Technologies and Electronic Employee Monitoring: A Meta-Analytic Review." Paper presented at the 2019 IEEE/SICE International Symposium on System Integration (SII), 14-16 Jan. 2019 2019.

Baiocco, Sara, Enrique Fernandez-Maciás, Uma Rani, and Annarosa Pesole. *The Algorithmic Management of Work and Its Implications in Different Contexts*. International Labour Organisation & European Commission (2022). https://www.ilo.org/wcmsp5/groups/public/---ed_emp/documents/publication/wcms_849220.pdf.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine-Learning*. 2019. https://fairmlbook.org/.

Barocas, Solon, and Andrew D. Selbst. "*Big Data*'s Disparate Impact." [In en]. *California Law Review* 104, no. 3 (2016): 671-732. https://doi.org/10.2139/ssrn.2477899. https://www.ssrn.com/abstract=2477899.

Bech-Nielsen, P.C. "Her Er Skatterådets Wolt-Afgørelse I Fuld Længde." *Radar*, January 27 2022. https://radar.dk/artikel/her-er-skatteraadets-wolt-afgoerelse-i-fuld-laengde.

Bell, Samuel James, and Levent Sagun. "Simplicity Bias Leads to Amplified Performance Disparities." Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago, IL, USA, Association for Computing Machinery, 2023.

Bevan, Gwyn, and Christopher Hood. "What's Measured Is What Matters: Targets and Gaming in the English Public Health Care System." *Public Administration* 84, no. 3 (2006): 517-38. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9299.2006.00600.x.

Bhave, Devasheesh P. "The Invisible Eye? Electronic Performance Monitoring and Employee Job Performance." *Personnel Psychology* 67, no. 3 (2014): 605-35. https://onlinelibrary.wiley.com/doi/abs/10.1111/peps.12046.

Bostrup, Jens. "Det Ligner Et Almindeligt Storrumskontor, Men Nyskabelsen Bliver Synlig, Når Man Vender Sig Om." *Politiken* (København), October 18, 2022 2022. https://politiken.dk/viden/art9018913/Det-ligner-et-almindeligt-storrumskontor-men-nyskabelsen-bliver-synlig-n%C3%A5r-man-vender-sig-om.

————."Tellis Oplevelse På Jobcenteret Fører Nu Til En Optrapning I Kampen Mod Diskriminerende Algoritmer." *Politiken*, 2021. https://politiken.dk/viden/Tech/art8140892/Tellis-oplevelse-p%C3%A5-jobcenteret-f%C3%B8rer-nu-til-en-optrapning-i-kampen-mod-diskriminerende-algoritmer.

Bringsjord, Selmer, and Naveen Sundar Govindarajulu. "Artificial Intelligence." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, 2018. https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=artificial-intelligence.

Brownstein, Michael. "Implicit Bias." In *The Stanford Encyclopedia of Philosophy,* edited by Edward Zalta, 2015. http://plato.stanford.edu/archives/spr2015/entries/implicit-bias/.

Burton, Jason W., Mari-Klara Stein, and Tina Blegind Jensen. "A Systematic Review of Algorithm Aversion in Augmented Decision Making." 33, no. 2 (2020): 220-39. https://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.2155.

Carey, Alycia N., and Xintao Wu. "The Statistical Fairness Field Guide: Perspectives from Social and Formal Sciences." *AI and Ethics* 3, no. 1 (2023): 1-23. https://doi.org/10.1007/s43681-022-00183-3.

Castro, Clinton, and Michele Loi. "The Fair Chances in Algorithmic Fairness: A Response to Holm." *Res Publica* 29, no. 2 (2023): 331-37. https://doi.org/10.1007/s11158-022-09570-3.

"Celonis." https://www.celonis.com/.

Chen, Jiawei, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. "Bias and Debias in Recommender System: A Survey and Future Directions." *ACM Transactions on Information Systems* 41, no. 3 (2023): 1-39.

Chouldechova, Alexandra. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data* 5, no. 2 (2017): 153-63. https://ui.adsabs.harvard.edu/\#abs/2016arXiv161007524C.

Chouldechova, Alexandra, and Aaron Roth. "The Frontiers of Fairness in Machine Learning." *arXiv e-prints*. (2018). Accessed October 01, 2018. https://ui.adsabs.harvard.edu/\#abs/2018arXiv181008810C.

Cohen, Gerald Allan. "On the Currency of Egalitarian Justice." In *On the Currency of Egalitarian Justice and Other Essays in Political Philosophy*, edited by Michael Otsuka, 3-43. Princeton: Princeton University Press, 2011.

"Comfy App." https://comfyapp.com/.

Corbett-Davies, Sam, and Sharad Goel. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." *arXiv preprint arXiv:1808.00023* (2018).

Dastin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women." In Ethics of Data and Analytics, 296-99: Auerbach Publications, 2022.

Dataetisk Råd, and Analyse & Tal. En Hverdag Af Data. (2023). https://www.ogtal.dk/cases/en-hverdag-af-data.

Datatilsynet. Udtalelse Fra Datatilsynet: Kommuners Hjemmel Til AI-Profileringsværktøjet Asta. (2022). https://www.datatilsynet.dk/afgoerelser/afgoerelser/2022/maj/udtalelse-vedroerende-kommuners-hjemmel.

Davis, Steven. "Is There a Right to Privacy?". *Pacific Philosophical Quarterly* 90, no. 4 (2009): 450-75.

De Stefano, Valerio. "'Negotiating the Algorithm': Automation, Artificial Intelligence and Labour Protection." *Artificial Intelligence and Labour Protection (May 16, 2018). Comparative Labor Law & Policy Journal* 41, no. 1 (2019).

Den uafhængige ekspertgruppe på højt niveau om kunstig intelligens. Etiske Retningslinjer for Pålidelig Kunstig Intelligens. Europa-Kommissionen (2018). https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60420.

Dieterich, William, Christina Mendoza, and Tim Brennan. Compas Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. NorthPointe (2016).

Dietvorst, Berkeley J. , Joseph P. Simmons, and Cade Massey. "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them." *Management Science* 64, no. 3 (2018): 1155-70. https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2016.2643.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err." *Journal of Experimental Psychology: General* 144, no. 1 (2015): 114-26.

Doyle, Tony. "Privacy and Perfect Voyeurism." *Ethics and Information Technology* 11, no. 3 (2009): 181-89. http://dx.doi.org/10.1007/s10676-009-9195-9.

Dunlop, Connor. An Eu AI Act That Works for People and Society. Ada Lovelace Institute (2023). https://www.adalovelaceinstitute.org/policy-briefing/eu-ai-act-trilogues/.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. "Fairness through Awareness." arXiv:1104.3913 [cs] (2011). http://arxiv.org/abs/1104.3913.

Dworkin, Ronald. Sovereign Virtue – the Theory and Practice of Equality. Cambridge: Harvard University Press, 2000.

————.*Taking Rights Seriously*. London: Gerald Duckworth & Co. Ltd., 2005. [1977].

Englich, Birte, and Thomas Mussweiler. "Sentencing under Uncertainty: Anchoring Effects in the Courtroom." *Journal of Applied Social Psychology* 31 (2001): 1535-51.

Ensign, Danielle, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. "Runaway Feedback Loops in Predictive Policing." 1st Conference on Fairness, Accountability and Transparency, arXiv, 2017.

Eurofound. *European Company Survey 2019 - Workplace Practices Unlocking Employee Potential*. (2019). https://www.eurofound.europa.eu/publications/flagship-report/2020/european-company-survey-2019-workplace-practices-unlocking-employee-potential.

Eyal, Nir. "Informed Consent." In *The Stanford Encyclopedia of Philosophy,* edited by Edward N. Zalta, 2019. https://plato.stanford.edu/archives/spr2019/entries/informed-consent/.

Fazelpour, Sina, and David Danks. "Algorithmic Bias: Senses, Sources, Solutions." *Philosophy Compass* 16, no. 8 (2021): 1-16. https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12760.

Ferguson, A.G. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. NYU Press, 2017.

"Forcepoint." https://www.forcepoint.com/.

Franco-Santos, Monica, and David Otley. "Reviewing and Theorizing the Unintended Consequences of Performance Management Systems." *International Journal of Management Reviews* 20, no. 3 (2018): 696-730. https://onlinelibrary.wiley.com/doi/abs/10.1111/ijmr.12183.

Friedler, Sorelle A., Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning." Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, Association for Computing Machinery, 2019.

Fritts, Megan, and Frank Cabrera. "AI Recruitment Algorithms and the Dehumanization Problem." *Ethics and Information Technology* 23, no. 4 (2021): 791-801. https://doi.org/10.1007/s10676-021-09615-w.

Gavison, Ruth. "Privacy and the Limits of Law." *The Yale Law Journal* 89, no. 3 (1980): 421-71. https://doi.org/10.2307/795891. http://www.jstor.org/stable/795891.

Giacosa, Elisa, Gazi Mahabubul Alam, Francesca Culasso, and Edoardo Crocco. "Stress-Inducing or Performance-Enhancing? Safety Measure or Cause of Mistrust? The Paradox of Digital Surveillance in the Workplace." *Journal of Innovation & Knowledge* 8, no. 2 (2023). https://www.elsevier.es/en-revista-journal-innovation-knowledge-376-articulo-stress-inducing-or-performance-enhancing-safety-measure-S2444569X23000537.

Goddard, Kate, Abdul Roudsari, and Jeremy C Wyatt. "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators." *Journal of the American Medical Informatics Association* 19, no. 1 (2011): 121-27. https://doi.org/10.1136/amiajnl-2011-000089.

Goldberg, Lewis. "Man Versus Model of Man: A Rationale, Plus Som Evidence, for a Method of Improving on Clinical Inferences." *Psychological Bulletin* 73, no. 6 (1970): 422-32.

Goodhart, C. A. E. "Problems of Monetary Management: The Uk Experience." In Monetary Theory and Practice: The Uk Experience, 91-121. London: Macmillan Education UK, 1984.

Grgic-Hlaca, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. "Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning." Association for the Advancement of Artificial Intelligence, 2018.

Grothar, Patrick, Mei Ngan, and Kayee Hanaoka. Face Recognition Vendor Test (Frvt) - Part 3: Demographic Effects. National Institute of Standards and Technology (U.S. Department of Commerce: 2019).

Hald, Casper Waldemar, Julie Karnøe Tranholm-Mikkelsen, and Katrine Lindtner Andersen. Digital Dataindsamling På Arbejdspladsen - En Undersøgelse Af Lederes Holdninger Til Og Oplevelser Med Indsamling Af Digitale Medarbejderdata På Arbejdspladsen. Mandag Morgen, Algoritmer, Data og Demokrati-projektet, Dansk Erhverv, DI, Djøf, DM, FH, Finansforbundet, Forsikringsforbundet, og IDA (2023). https://taenketanken.mm.dk/wp-content/uploads/2023/09/Minirapport_Digital-dataindsamling-Lederanalyse08.pdf.

Hardt, Moritz, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning." *arXiv:1610.02413 [cs]* (2016). http://arxiv.org/abs/1610.02413.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2017. https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12_toc.pdf.

Havstein, Eriksen A. "Wolt Skal Betale Erstatning Til Alle Bude, Der Kommer Til Skade." *Fagbladet 3F*, July 5 2023.https://fagbladet3f.dk/artikel/wolt-skal-betale-erstatning-til-alle-bude-der-kommer-til-skade.

Heidari, Hoda, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. "Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making." *Advances in neural information processing systems* 31 (2018).

Heidari, Hoda, Michele Loi, Krishna P Gummadi, and Andreas Krause. "A Moral Framework for Understanding Fair Ml through Economic Models of Equality of Opportunity." Paper presented at the Proceedings of the conference on fairness, accountability, and transparency, 2019.

Hellman, Deborah. "Measuring Algorithmic Fairness." *Virginia Law Review* 106, no. 4 (2020): 811-66.

Hickok, Merve, and Nestor Maslej. "A Policy Primer and Roadmap on AI Worker Surveillance and Productivity Scoring Tools." *AI and Ethics* 3, no. 3 (2023): 673-87. https://doi.org/10.1007/s43681-023-00275-8.

Hill, Robin K. "What an Algorithm Is." *Journal of Philosophy & Technology* 29, no. 1 (2016): 35-59. https://doi.org/10.1007/s13347-014-0184-5.

Hills, P. J., D. Dickinson, L. M. Daniels, C. A. Boobyer, and R. Burton. "Being Observed Caused Physiological Stress Leading to Poorer Face Recognition." [In eng]. *Acta Psychol (Amst)* 196 (2019): 118-28. https://doi.org/10.1016/j.actpsy.2019.04.012.

Himma, Kenneth Einar. "Positivism, Naturalism, and the Obligation to Obey Law." *The Southern Journal of Philosophy* 36, no. 2 (1998): 145-61. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-6962.1998.tb01749.x.

Holm, Sofie Caroline Falkenberg, and Sofie Dalum. Fire Ud Af Fem Får Indsamlet Medarbejderdata. DM Akademikerforeningen (2023). https://dm.dk/media/1can022l/fire-ud-af-fem-faar-indsamlet-medarbejderdata.pdf.

IDA. Guide Til Ledere Om Overvågning Og Monitorering På Arbejdspladser. (2023).

————.Guide Til Tillidsvalgte Om Overvågning Og Monitorering På Arbejdspladser. (2023). https://ida.dk/media/13402/overvaagning-paa-arbejdspladsen-2023.pdf.

Inness, Julie. Privacy, Intimacy, and Isolation. Oxford: Oxford University Press, 1992.

InterGuard. "Employee Productivity Tracking Software." https://www.interguardsoftware.com/employee-productivity-tracking/.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. New York: Springer, 2013.

Jobin, Anna, Marcello Ienca, and Effy Vayena. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1, no. 9 (2019): 389-99. https://doi.org/10.1038/s42256-019-0088-2.

Kahneman, Daniel, Andrew M. Rosenfield, Linnea Gandhi, and Tom Blaser. "Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making." Harvard Business Review October (2016). https://hbr.org/2016/10/noise.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. *Noise: A Flaw in Human Judgment*. London: William Collins, 2021.

Kahneman, Daniel, and Amos Tversky, eds. *Choices, Values, and Frames*. New York: Cambridge University Press, 2009.

Kalischko, Thomas, and René Riedl. "Electronic Performance Monitoring in the Digital Workplace: Conceptualization, Review of Effects and Moderators, and Future Research Opportunities." *Frontiers in psychology* 12 (2021): 633031.

Kelly, Thomas. *Bias: A Philosophical Study*. Oxford University Press, 2022.

Keren, Arnon, and Ori Lev. "Informed Consent, Error and Suspending Ignorance: Providing Knowledge or Preventing Error?". *Ethical Theory and Moral Practice* 25, no. 2 (2022): 351-68.

Khan, Arif Ali, Sher Badshah, Peng Liang, Muhammad Waseem, Bilal Khan, Aakash Ahmad, Mahdi Fahmideh, Mahmood Niazi, and Muhammad Azeem Akbar. "Ethics of AI: A Systematic Literature Review of Principles and Challenges." Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering, Gothenburg, Sweden, Association for Computing Machinery, 2022.

Khan, Falaah Arif, Eleni Manis, and Julia Stoyanovich. "Fairness as Equality of Opportunity: Normative Guidance from Political Philosophy." *arXiv preprint arXiv:2106.08259* (2021).

Kilbertus, Niki, Adrià Gascón, Matt J. Kusner, Michael Veale, Krishna P. Gummadi, and Adrian Weller. "Blind Justice: Fairness with Encrypted Sensitive Attributes." *arXiv:1806.03281 [cs, stat]* (2018). http://arxiv.org/abs/1806.03281.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "Human Decisions and Machine Predictions." *NBER Working paper series* (2017). http://www.nber.org/papers/w23180.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. "Discrimination in the Age of Algorithms." *arXiv e-prints*. (2019). Accessed February 01, 2019. https://ui.adsabs.harvard.edu/\#abs/2019arXiv190203731K.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *arXiv e-prints*. (2016). Accessed September 01, 2016. https://ui.adsabs.harvard.edu/\#abs/2016arXiv160905807K.

Kroll, Joshua, Joanna Huey, Solon Barocas, Edward Felten, Joel Reidenberg, David Robinson, and Harlan Yu. "Accountable Algorithms." *University of Pennsylvania Law Review* 165, no. 3 (2017): 633. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3.

Kusner, Matt J., Joshua R. Loftus, Chris Russell, and Ricardo Silva. "Counterfactual Fairness." *arXiv e-prints*. (2017). Accessed March 01, 2017. https://ui.adsabs.harvard.edu/\#abs/2017arXiv170306856K.

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. "How We Analyzed the Compas Recidivism Algorithm." *ProPublica*. (2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Leicht-Deobald, Ulrich, Thorsten Busch, Christoph Schank, Antoinette Weibel, Simon Schafheitle, Isabelle Wildhaber, and Gabriel Kasper. "The Challenges of Algorithm-Based Hr Decision-Making for Personal Integrity." *Journal of Business Ethics* 160, no. 2 (2019): 377-92. https://doi.org/10.1007/s10551-019-04204-w.

Lidén, Moa, Minna Gräns, and Peter Juslin. "'Guilty, No Doubt': Detention Provoking Confirmation Bias in Judges' Guilt Assessments and Debiasing Techniques." *Psychology, Crime & Law* 25, no. 3 (2019): 219-47. https://doi.org/10.1080/1068316X.2018.1511790.

Lippert-Rasmussen, Kasper. "Algorithm-Based Sentencing and Discrimination." (2022): 74-96.

———.Born Free and Equal? A Philosophical Inquiry into the Nature of Discrimination. Oxford: Oxford University Press, 2013.

———."Discrimination and the Aim of Proportional Representation." Politics, Philosophy & Economics 7 (2008): 159-82.

Lipton, Zachary C. The Mythos of Model Interpretability. 2017. arXiv.

Lipton, Zachary C., Alexandra Chouldechova, and Julian McAuley. "Does Mitigating Ml's Impact Disparity Require Treatment Disparity?" 32nd Conference on Neural Information Processing Systems, 2018.

List, Christian. "Group Agency and Artificial Intelligence." *Philosophy & Technology* 34, no. 4 (2021): 1213-42. https://doi.org/10.1007/s13347-021-00454-7.

Lomborg, Stine. "Everyday AI at Work - Self-Tracking and Automated Communication for Smart Work." In *Everyday Automation*, edited by Sarah Pink, Martin Berg, Deborah Lupton and Minna Ruckenstein, 126-39: Routledge, 2022.

Lundgren, Björn. "A Dilemma for Privacy as Control." *The Journal of Ethics* 24, no. 2 (2020): 165-75. https://doi.org/10.1007/s10892-019-09316-z.

Lyell, D., and E. Coiera. "Automation Bias and Verification Complexity: A Systematic Review." [In eng]. *Journal of the American Medical Informatics Association* 24, no. 2 (2017): 423-31. https://doi.org/10.1093/jamia/ocw105.

Macnish, Kevin. "An Eye for an Eye: Proportionality and Surveillance." *Ethical Theory and Moral Practice* 18, no. 3 (2015): 529-48.

———."Government Surveillance and Why Defining Privacy Matters in a Post-Snowden World." *Journal of Applied Philosophy* 35, no. 2 (2018): 417-32.

Mainz, Jakob. "An Indirect Argument for the Access Theory of Privacy." *Res Publica* 27, no. 3 (2021): 309-28. https://doi.org/10.1007/s11158-021-09521-4.

Marmor, Andrei. "What Is the Right to Privacy?". *Philosophy and Public Affairs* 43, no. 1 (2015): 3-26.

Martin, Angela J., Jackie M. Wellen, and Martin R. Grimmer. "An Eye on Your Work: How Empowerment Affects the Relationship between Electronic Surveillance and Counterproductive Work Behaviours." *The International Journal of Human Resource Management* 27, no. 21 (2016): 2635-51. https://doi.org/10.1080/09585192.2016.1225313.

Work Behaviours." *The International Journal of Human Resource Management* 27, no. 21 (2016): 2635-51. https://doi.org/10.1080/09585192.2016.1225313.

Mateescu, Alexandra, and Aiha Nguyen. *Workplace Monitoring & Surveillance*. Data & Society (2019). https://datasociety.net/wp-content/uploads/2019/02/DS_Workplace_Monitoring_Surveillance_Explainer.pdf.

McCloskey, H. J. "Privacy and the Right to Privacy." Philosophy 55, no. 211 (1980): 17 - 38.

Menges, Leonhard. "A Defense of Privacy as Control." *The Journal of Ethics* 25, no. 3 (2021): 385-402. https://doi.org/10.1007/s10892-020-09351-1.

Migliano, Simon. *Employee Monitoring Software Demand Trends 2020-23*. Top10VPN (2023). https://www.top10vpn.com/research/covid-employee-surveillance/.

Mill, John Stuart. "On Liberty." In On Liberty and Other Writings, edited by Stefan Collini, 1-116. Cambridge: Cambridge University Press, 2009.

Millum, Joseph, and Danielle Bromwich. "Informed Consent: What Must Be Disclosed and What Must Be Understood?". *American Journal of Bioethics* 21, no. 5 (2021): 46-58.

Mims, Christopher. "More Bosses Are Spying on Quiet Quitters. It Could Backfire." *The Wall Street Journal*, September 17 2022. https://www.wsj.com/articles/more-bosses-are-spying-on-quiet-quitters-it-could-backfire-11663387216.

Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. "Algorithmic Fairness: Choices, Assumptions, and Definitions." *Annual Review of Statistics and Its Application 8*, no. 1 (2021): 141-63. https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-042720-125902.

Mittelstadt, Brent. "Principles Alone Cannot Guarantee Ethical AI." *Nature Machine Intelligence* 1, no. 11 (2019): 501-07. https://doi.org/10.1038/s42256-019-0114-4.

Molnar, Christoph. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. 2019. https://christophm.github.io/interpretable-ml-book/.

Moltke, Henrik, and Marcel Mirzaei-Fard. "Wolt-Budet Laura Vil Have En Overenskomst: 'Det Burde Ikke Være Anderledes, Fordi Vi Arbejder for En App'." *DR*, June 11 2020. https://www.dr.dk/nyheder/penge/wolt-budet-laura-vil-have-en-overenskomst-det-burde-ikke-vaere-anderledes-fordi-vi

Moore, Adam D. "Privacy: Its Meaning and Value." *American Philosophical Quarterly* 40, no. 3 (2003): 215-27. http://www.jstor.org/stable/20010117.

Moreau, Therese. "Ida Overvåger Medarbejdere Med Uigennemskuelig AI: "Vi Ved, at Systemet Ikke Forstår Alt"." *ING/Datatech*, October 25, 2022. https://pro.ing.dk/datatech/artikel/ida-overvaager-medarbejdere-med-uigennemskuelig-ai-vi-ved-systemet-ikke-forstaar.

Moreau, Therese, and Frederik Kulager. "Vi Har Skilt Jobcentrenes Algoritme Ad." *Zetland*, June 10, 2021. https://www.zetland.dk/historie/sOMVZ7qG-aOz9m93B-a30b8.

MPLOY. Evaluering Af Projekt "Samtaler Og Indsats Der Modvirker Langtidsledighed". Styrelsen for Arbejdsmarked og Rekruttering, (2018). https://star.dk/media/8004/evaluering-af-projekt-samtaler-og-indsats-der-modvirker-langtidsledighed.pdf.

MSI-AUT. A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility within a Human Rights Framework. Council of Europe (2018). https://rm.coe.int/draft-study-of-the-implications-of-advanced-digital-technologies-inclu/16808ef255.

Muller, Jerry. The Tyranny of Metrics. Princeton University Press, 2018.

Munch, Lauritz, and Jakob Mainz. "To Believe, or Not to Believe – That Is Not the (Only) Question: The Hybrid View of Privacy." *The Journal of Ethics* (2023). https://doi.org/10.1007/s10892-023-09419-8.

Munk, Grit, Marie Langmach, Julie Karnøe Tranholm-Mikkelsen, and Zenia Søjberg. Danskernes Holdninger Til Og Oplevelser Med Indsamling Af Digitale Medarbejderdata På Arbejdspladsen. Mandag Morgen, Algoritmer, Data og Demokratiprojektet, IDA – Ingeniørforeningen, HK Danmark, og Dataetisk Råd (2023). https://algoritmer.org/medarbejderdata/.

Nagel, Thomas. "Concealment and Exposure." *Philosophy and Public Affairs* 27, no. 1 (1998): 3-30.

Negrón, Wilneida. Little Tech Is Coming for Workers. CoWorker (2021). https://home.coworker.org/wp-content/uploads/2021/11/Little-Tech-Is-Coming-for-Workers.pdf.

Nickerson, Raymond S. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of general psychology* 2, no. 2 (1998): 175-220.

Nissenbaum, Helen. "Privacy as Contextual Integrity." *Washington Law Review* 79, no. 1 (2004): 101-39.

Novelli, Claudio, Federico Casolari, Rotolo Antonino, Taddeo Mariarosaria, and Luciano Floridi. "Taking AI Risks Seriously: A New Assessment Model for the AI Act." *AI & Society* 38, no. 3 (2023). https://doi.org/10.1007/s00146-023-01723-z.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown/Archetype, 2016.

Parent, William A. "Privacy, Morality, and the Law." *Philosophy and Public Affairs* 12, no. 4 (1983): 269-88.

Penney, Jonathon W. "Chilling Effects: Online Surveillance and Wikipedia Use." *Berkeley Technology & Law Journal* 31, no. 1 (2016): 118-82.

————."Internet Surveillance, Regulation, and Chilling Effects Online: A Comparative Case Study." *Internet Policy Review* 6, no. 2 (2017): 1-39.

Perry, James L., Trent A. Engbers, and So Yun Jun. "Back to the Future? Performance-Related Pay, Empirical Research, and the Perils of Persistence." *Public Administration Review* 69, no. 1 (2009): 39-51. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6210.2008.01939_2.x.

Powers, Madison. "A Cognitive Access Definition of Privacy." *Law and Philosophy* 15, no. 4 (1996): 369 - 86.

Prahl, Andrew, and Lyn Van Swol. "Understanding Algorithm Aversion: When Is Advice from Automation Discounted?". *Journal of Forecasting* 36, no. 6 (2017): 691-702. https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2464.

Rachels, James. "Why Privacy Is Important." *Philosophy & Public Affairs* 4, no. 4 (1975): 323-33.

Rassin, Eric. "Context Effect and Confirmation Bias in Criminal Fact Finding." 25, no. 2 (2020): 80-89. https://onlinelibrary.wiley.com/doi/abs/10.1111/lcrp.12172.

Ravid, Daniel M., Jerod C. White, David L. Tomczak, Ahleah F. Miles, and Tara S. Behrend. "A Meta-Analysis of the Effects of Electronic Performance Monitoring on Work Outcomes." *Personnel Psychology* 76, no. 1 (2023): 5-40. https://onlinelibrary.wiley.com/doi/abs/10.1111/peps.12514.

Rawls, John. *A Theory of Justice*. Oxford: Oxford University Press, 1999. [1971].

Robinson, Joseph P, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. "Face Recognition: Too Bias, or Not Too Bias?" Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020.

Roemer, John E. "A Pragmatic Theory of Responsibility for the Egalitarian Planner." *Philosophy & Public Affairs* 22, no. 2 (1993): 146-66. http://www.jstor.org/stable/2265444.

Roessler, Beate. *The Value of Privacy*. Polity Press, 2005.

Rubel, Alan. "Claims to Privacy and the Distributed Value View." Article. *San Diego Law Review* 44, no. 4 (2007): 921-56. http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=31197946&site=ehost-live.

————."The Particularized Judgment Account of Privacy." *Res Publica* 17, no. 3 (2011): 275-90.

Ryberg, Jesper. "Privacy Rights, Crime Prevention, Cctv, and the Life of Mrs. Aremac." *Res Publica* 13, no. 2 (2007): 127-43.

Samuelson, William, and Richard Zeckhauser. "Status Quo Bias in Decision Making." journal article. *Journal of Risk and Uncertainty* 1, no. 1 (1988): 7-59. https://doi.org/10.1007/bf00055564. http://dx.doi.org/10.1007/BF00055564.

Segall, Shlomi. "Should the Best Qualified Be Appointed?*." Journal of Moral *Philosophy* 9, no. 1 (2012): 31-54. https://brill.com/view/journals/jmp/9/1/article-p31_4.xml.

————."What's So Bad About Discrimination?". *Utilitas* 24, no. 1 (2012): 82-100.

Selbst, Andrew D, and Julia Powles. "Meaningful Information and the Right to Explanation." *International Data Privacy Law* 7, no. 4 (2017): 233-42. https://doi.org/10.1093/idpl/ipx022.

Smith, M. J., P. Carayon, K. J. Sanders, S. Y. Lim, and D. LeGrande. "Employee Stress and Health Complaints in Jobs with and without Electronic Performance Monitoring." *Applied Ergonomics* 23, no. 1 (1992): 17-27. https://www.sciencedirect.com/science/article/pii/000368709290006H.

Solove, Daniel J. "'I've Got Nothing to Hide' and Other Misunderstandings of Privacy." *San Diego Law Review* 44 (2007): 745-72.

————. "A Taxonomy of Privacy." *University of Pennsylvania Law Review* 154, no. 3 (2006): 477-560.

Sommer, Mathias. "Overblik: Uber Får Sparket I Flere Europæiske Lande." *DR*, March 28 2017. https://www.dr.dk/nyheder/penge/overblik-uber-faar-sparket-i-flere-europaeiske-lande.

Stoycheff, Elizabeth, Juan Liu, Kai Xu, and Kunto Wibowo. "Privacy and the Panopticon: Online Mass Surveillance's Deterrence and Chilling Effects." *New Media & Society* 21, no. 3 (2019): 602-19. https://journals.sagepub.com/doi/abs/10.1177/1461444818801317.

Tavani, Herman T. "Philosophical Theories of Privacy: Implications for an Adequate Online Privacy Policy." *Metaphilosophy* 38, no. 1 (2007): 1–22.

Teebken, Mena Angela. "What Makes Workplace Privacy Special? An Investigation of Determinants of Privacy Concerns in the Digital Workplace." Paper presented at the AMCIS, 2021.

Thomsen, Frej Klem. "Algorithmic Indirect Discrimination, Fairness and Harm." *AI and Ethics* (2023). https://doi.org/10.1007/s43681-023-00326-0.

Thomson, Judith Jarvis. "The Right to Privacy." *Philosophy & Public Affairs* 4, no. 4 (1975): 295-314. http://www.jstor.org/stable/2265075.

Trade Union Congress. *Technology Managing People*. (London: 29 November 2020). https://www.tuc.org.uk/sites/default/files/2020-11/Technology_Managing_People_Report_2020_AW_Optimised.pdf.

UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. (2021). https://unesdoc.unesco.org/ark:/48223/pf0000381137/PDF/381137eng.pdf.multi.

UNI Global Union. *Top 10 Principles for Workers' Data Privacy and Protection*. (Nyon: 2017). https://uniglobalunion.org/wp-content/uploads/uni_workers_data_protection-1.pdf.

Veale, Michael, and Frederik Zuiderveen Borgesius. "Demystifying the Draft Artificial Intelligence Act." *Computer Law Review International* 4 (2021).

Velíz, Carrisa. "The Internet and Privacy." In *Ethics and the Contemporary World*, edited by David Edmonds, 149-59. Abingdon: Routledge, 2019.

VMware. *The New Remote Work Era: Trends in the Distributed Workforce*. (2023). https://www.vmware.com/content/microsites/learn/en/655785_REG.html

Volante, Louis. "Teaching to the Test: What Every Educator and Policy-Maker Should Know." *Canadian Journal of Educational Administration and Policy* 35 (2004).

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law* 7, no. 2 (2017): 76-99. https://papers.ssrn.com/abstract=2903469.

Walker, Tom. "Informed Consent and the Requirement to Ensure Understanding." *Journal of Applied Philosophy* 29, no. 1 (2011): 50-62.

Weingart, Peter. "Impact of Bibliometrics Upon the Science System: Inadvertent Consequences?". Scientometrics 62 (2005): 117-31.

Welsh, Brandon C., and David P. Farrington. "Public Area Cctv and Crime Prevention: An Updated Systematic Review and Meta-Analysis." *Justice Quarterly* 26, no. 4 (2009): 716-45.

Wertheimer, A. Coercion. Princeton University Press, 2014.

Westin, A.F. *Privacy and Freedom*. New York: Ig Publishing, 1967.

Yost, Allison Brown, Tara S. Behrend, Garett Howardson, Jessica Badger Darrow, and Jaclyn M. Jensen. "Reactance to Electronic Surveillance: A Test of Antecedents and Outcomes." *Journal of Business and Psychology* 34, no. 1 (2019): 71-86. https://doi.org/10.1007/s10869-018-9532-2.

Yu, Martin C., and Nathan R. Kuncel. "Pushing the Limits for Judgemental Consistency: Comparing Random Weighting Schemes with Expert Judgments." Personal Assessments and Decisions 6, no. 2 (2020): 1-10.