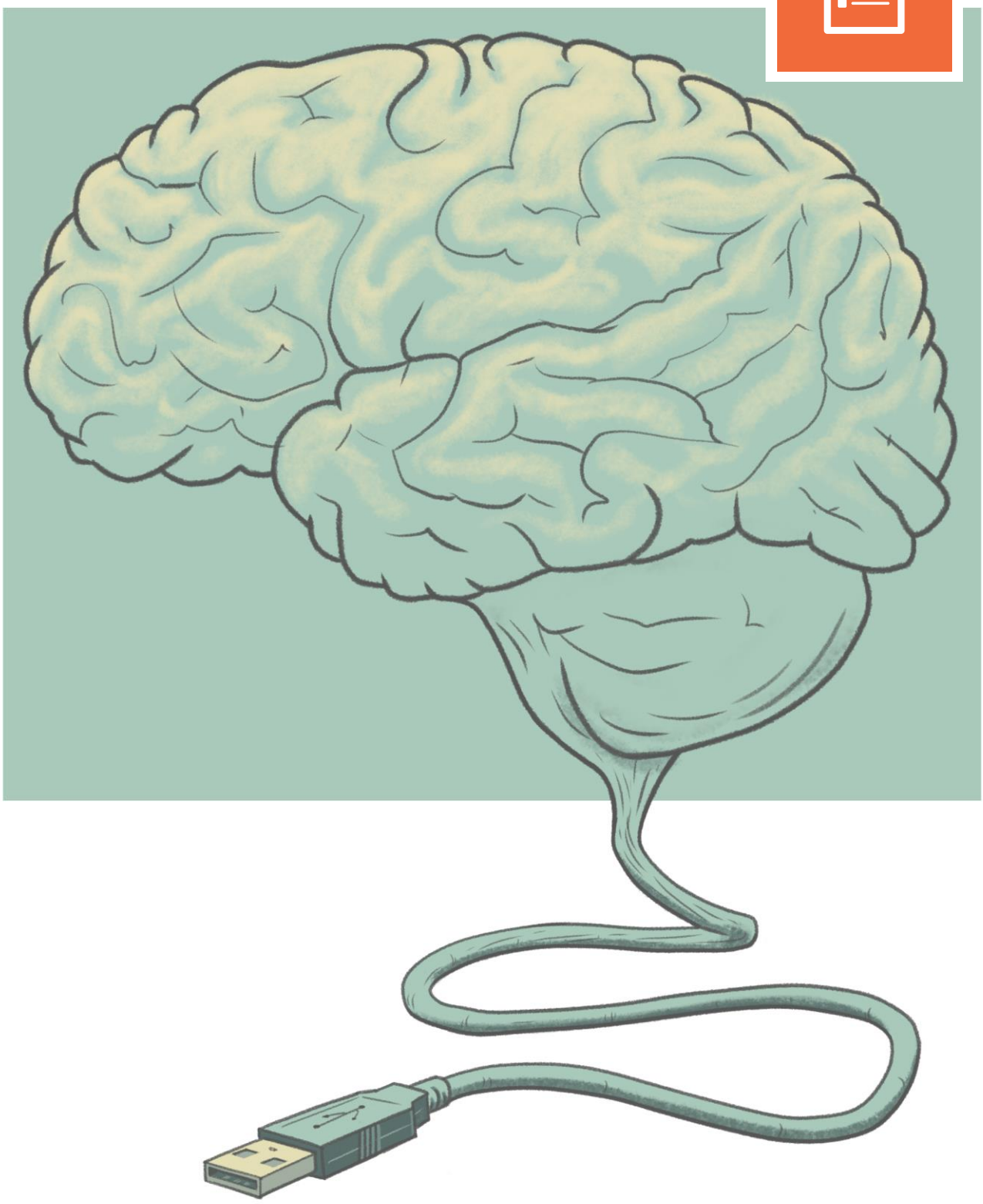


Kunstig intelligens og personvern

Rapport, januar 2018

Datatilsynet

Innhold

Kort om rapporten

Nesten all bruk av kunstig intelligens (KI) forutsetter store mengder data for å kunne lære og ta intelligente avgjørelser. Potensialet for radikalt bedre tjenester, forskningsmessige gjennombrudd og økonomisk gevinst, setter kunstig intelligens høyt på agendaen i de fleste sektorer.

Vi kommer til å se stadig flere juridiske og etiske dilemmaer hvor potensialet for betydelig samfunnsgevinst må veies opp mot grunnleggende personvernhensyn. Denne rapporten har som mål å beskrive og hjelpe til med å forstå hvordan personvernet blir berørt av utvikling og bruk av kunstig intelligens.

Datatilsynet mener debatt og kunnskap om personvernimplikasjonene ved kunstig intelligens er nødvendig – både for å ivareta enkeltmenneskers personvernrettigheter, *og* for å ivareta samfunnsbehov utover personvernet.

Hvis folk ikke har tillit til at opplysninger om dem behandles på en god måte, kan det begrense villigheten til å dele opplysninger, slik som hos legen eller på sosiale medier I en slik situasjon står vi ovenfor betydelige utfordringer for ytringsfriheten og tilliten til myndighetene. At folk vegrer seg for å dele opplysninger om seg selv, vil også være en betydelig utfordring for kommersiell bruk av personopplysninger i sektorer som media, handel og finans.

Denne rapporten bygger videre på de juridiske vurderingene og de tekniske beskrivelsene i rapporten «Big Data – personvernprinsipper under press»[1] fra 2014. Vi supplerer her ved å gå dypere inn i den tekniske beskrivelsen av kunstig intelligens, samt ved å se nærmere på fire KI-relevante utfordringer knyttet opp mot personvernprinsippene i forordningen:

- Rettferdighet og diskriminering
- Formålsbestemthet
- Dataminimalisering
- Gjennomsiktighet og retten til informasjon

Listen er ikke fullstendig, men det er et utvalg av personvernproblemstillingene som etter vår vurdering er mest relevante i forbindelse med bruk av kunstig intelligens akkurat nå. Rapporten drøfter dessuten Datatilsynets rolle som tilsynsetat, og til slutt gir vi en rekke eksempler på og anbefalinger om metoder og verktøy for å ivareta personvernet i utvikling og bruk av kunstig intelligens.

Målgruppen for denne rapporten er alle som jobber med, eller av andre grunner er interessert i, kunstig intelligens. Vi håper teknologer, samfunnsvitere, jurister og andre faggrupper kan ha nytte av rapporten.

Utarbeidelsen av rapporten har vært en læringsprosess for oss i Datatilsynet. Vi har fått mye igjen fra å lytte til ulike aktører sine erfaringer og vurderinger knyttet til kunstig intelligens og personvern. En stor takk til Inmeta, Privacy International, Finanstilsynet, Google, Sintef, NTNU, Big Insight ved Universitetet i Oslo og Norsk Regnesentral, Sparebank 1 Stavanger, Information Commissioner's Office i England, Office of the Privacy Comissioner i Canada, Riksrevisjonen og Center for Artificial Intelligence Research ved Universitetet i Agder.

Rettskilder og begrepsbruk

I denne rapporten bruker vi kunstig intelligens som et samlebegrep som betegner ulike former for KI, inkludert maskinlæring og dyp læring.

Rapporten tar utgangspunkt i **EUs nye personvernforordning** (også kalt GDPR).[2] Forordningen skal gjøres til norsk rett gjennom en ny personopplysningslov som trer i kraft 25. mai 2018.

Vi har også brukt **forordningens fortale** for å tolke artiklenes innhold. Fortalen er ikke juridisk bindende, men hjelper med å forklare artiklenes innhold.

Vi har dessuten brukt **uttalelser fra Artikkel 29-gruppen** (Article 29 Working Party) og deres retningslinjer for individuelle automatiserte avgjørelser og profilering.[3] Artikkel 29-gruppen er den øverste rådgivende forsamlingen for EU-kommisjonen i spørsmål om personvern og informasjonssikkerhet.

[1] https://www.datatilsynet.no/om-personvern/rapporter-og-utredninger/temarapporter/big-data/

[2] Den europeiske personvernforordningen på engelsk: http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2016:119:FULL

[3] Foreløpig siste versjon av arbeidsdokumentet: http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=50083

Kunstig intelligens og personvern

Kunstig intelligens er et begrep som beskriver datasystemer som kan lære av egne erfaringer og løse komplekse problemstillinger i ulike situasjoner – egenskaper vi tidligere har tenkt er unike for mennesker. Data, i mange tilfeller personopplysninger, er drivstoffet som gjør at systemet kan lære og bli intelligente.

Utviklingen av kunstig intelligens har hatt store fremskritt de siste årene og mulighetene virker lovende: en bedre og mer effektiv offentlig sektor, nye metoder for klima- og miljøvern, et sikrere samfunn og kanskje også løsningen på kreftgåten?

Vi er med andre ord i startgropen på noe som uomtvistelig vil ha en betydelig effekt på samfunnet. Derfor er det viktig at vi tar diskusjonen nå – hvilke rammer trenger vi for at vi kan realisere mulighetene som kunstig intelligens innebærer på en sikker og rettferdig måte? For det er ikke til å komme bort ifra at bruken av kunstig intelligens reiser mange problemstillinger blant annet innen etikk, sikkerhet, rettslig ansvar og så videre. Denne rapporten tar for seg en slik problemstilling, nemlig bruken av personopplysninger og personvern innen kunstig intelligens.

Fra vinter til vår – hvorfor nå?

Kunstig intelligens ble kjent allerede på 1950-tallet som et begrep og en teknikk det ble knyttet store forhåpninger til. De første fremskrittene ble fulgt av mange tiår som ofte blir kalt KI-vinteren fordi utviklingen man håpet på ikke innfridde forventningene. De siste årene har vi imidlertid sett en vårløsning.

I dag ser vi at det brukes KI til å løse spesifikke oppgaver slik som for eksempel bilde- og talegjenkjenning. Dette kalles ofte *spesialisert* KI. *Generell* KI viser til systemer som har samme fleksibilitet i læring og problemløsning som mennesker. Dette ligger sannsynligvis fremdeles flere tiår frem i tid.

I tillegg til en kraftig økning i prosesseringskraft, samt billigere og større lagringskapasitet, har tilgangen på store mengder data bidratt til vårløsningen for KI. Stordata refererer ofte til data i stort volum, fra mange kilder og gjerne i sanntid.[4] De enorme datastrømmene kan utnyttes for å oppnå samfunnsgevinst ved å analysere dataene og finne mønstre og sammenhenger.

Og det er her KI kan gjøre en forskjell. Mens tradisjonelle analysemetoder er avhengige av å bli programmert til å finne sammenhenger og koblinger, lærer KI av alle dataene den ser. Datasystemene kan derfor hele tiden respondere på nye data og tilpasse analysen sin, uten menneskelig innblanding. KI hjelper derfor til med å bryte de tekniske begrensningene som tradisjonelle metoder møter når stordata skal analyseres.

Større etterspørsel etter data, strengere regler

De nye personvernreglene vi får i mai 2018, styrker rettighetene våre når det gjelder egne personopplysninger, og kravene til de som behandler data skjerpes. Virksomhetene får større ansvar for at de behandler personopplysningene i tråd med regelverket, og kravene til åpenhet blir strengere.

Hvis bruk av KI innebærer en behandling av personopplysninger, utløses det både plikter og rettigheter. Forordningen krever at de som behandler data vurderer personvernkonsekvensene og bygger inn personvern allerede i utviklingsfasen av et system. Dette er forutsetninger for å sørge for at folks opplysninger behandles på en sikker og god måte.

På samme tid som kravene skjerpes, vokser etterspørselen etter data. Systemer basert på KI blir kun intelligente hvis de har nok relevante data de kan lære av.

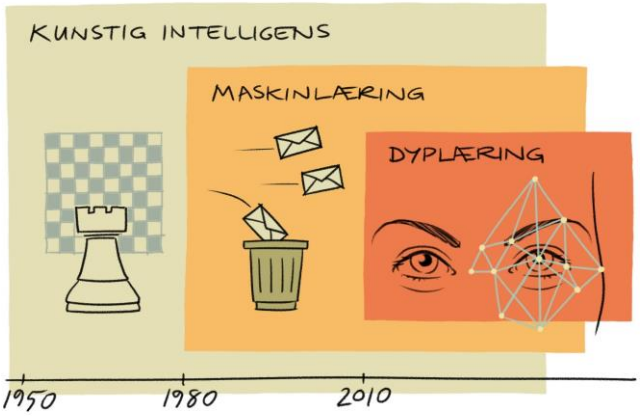
En intelligent chatbot (et dataprogram som mennesker kan samhandle med enten skriftlig eller muntlig) analyserer all informasjon den blir matet med – en kombinasjon av spørsmål fra kunder og svar fra kundeservice. Ut i fra analysen «forstår» chatboten hva en kunde lurer på og kan gi et meningsfullt svar. Jo mer informasjon chatboten har å basere sin analyse på, jo bedre og mer presist blir svaret den gir.

---

[4] https://ico.org.uk/for-organisations/guide-to-data-protection/big-data/

Kunstig intelligens, maskinlæring og dyp læring

Kunstig intelligens, maskinlæring og dyp læring er begreper som ofte brukes som synonymer selv om det egentlig ikke er en presis begrepsbruk. Illustrasjonen under viser relasjonen mellom begrepene og utvikling i tid.



Kunstig intelligens er et paraplybegrep som omfatter mange forskjellige typer maskinlæring. Maskinlæring kan beskrives som «et sett teknikker og verktøy som lar maskiner «tenke» ved å lage matematiske algoritmer basert på akkumulert data».[5] Systemet kan tenke uavhengig av menneskelig input, og selv bygge nye algoritmer. På denne måten kan vi få ut kunnskap fortløpende, etter hvert som systemet blir eksponert for stadig nye og ulike typer av data.

Dyp læring er en form for maskinlæring, og noen former for dyp læring er bygd opp rundt samme prinsippene som det nevrale nettverket i hjernen. Slike systemer tar ofte utgangspunkt i et kjent treningsdatasett som hjelper de selvlærende algoritmene å få nettverket til å utføre en oppgave. Dette forutsetter at nettverket selv kan avgjøre hva som er riktig respons for å løse oppgaven.[6] Metoden var blant annet avgjørende for at dataprogrammet AlphaGo kunne slå en av verdens beste spillere i det kinesiske brettspillet Go (se faktaboks). Dette ble regnet som en viktig milepæl for kunstig intelligens.

Er det mulig å kombinere kunstig intelligens og godt personvern?

I arbeidet med denne rapporten har vi snakket med ulike aktører som utvikler eller bruker KI. Vi sitter igjen med et inntrykk av at de fleste sektorene har tatt i bruk KI i relativt begrenset omfang, og at teknologien som brukes ofte er begrenset. Det samsvarer også ganske godt med Datatilsynets begrensede saksporteføljе og veiledningshenvendelser når det kommer til KI og personvern.

Vi er altså fremdeles er i en tidlig fase av utviklingen av KI, og dette er et godt tidspunkt å sørge for at teknologien følger spillereglene i samfunnet. Svaret på spørsmålet om det er mulig å benytte seg av kunstig intelligens og ivareta personvernet samtidig, er ja. Det er både mulig og nødvendig for å ivareta grunnleggende personvernrettigheter.

AlphaGo

AlphaGo er et dataprogram som vant over en av verdens beste spillere i det kinesiske brettspillet Go.

Go er et spill med så mange mulige kombinasjoner at det er per i dag er umulig å regne ut alle, derfor trengte man en mer intelligent tilnærming til spillet enn å benytte rå regnekraft. AlphaGo er utviklet av Deepmind som er eksperter på dyp læring, noe som ble benyttet som en del av programmet.

Programmet ble opplært ved å gå igjennom historikken fra en mengde kamper mellom mennesker. Videre spilte programmet mot seg selv for å lære mer om hvilke trekk og strategier som ga best resultater.

Et av de mest interessante resultatene utover at AlphaGo vant, var at programmet tok i bruk nye strategier som hittil ikke var normalt kjent. Disse har blitt publisert og brukes nå av Go-spillere.

(Kilde: https://www.blog.google/topics/machine-learning/alphago-machine-learning-game-go/)

[5] https://iq.intel.com/artificial-intelligence-and-machine-learning/

[6] https://no.wikipedia.org/wiki/Nevralt_nettverk, https://en.wikipedia.org/wiki/Deep_learning

Hvordan fungerer kunstig intelligens?

Det er to hovedaspekter ved kunstig intelligens som er spesielt relevante når det kommer til personvern. Det ene er at programvaren selv kan ta avgjørelser, og det andre er at systemet utvikler seg selv ved å lære av erfaring.

For at et datasystem skal lære må vi gi det erfaring, og denne erfaringen kommer ved at vi tilfører datasystemet informasjon. Informasjonen kan da komme i forskjellige formater. Hvis man ønsker et system som kun skal utføre bildegjenkjenning/-analyse, vil erfarings-grunnlaget naturlig nok bestå av bilder. For andre oppgaver kan datagrunnlaget bestå av tekst, tale eller tall. Noen systemer vil benytte seg av person-opplysninger, mens andre systemer bruker data som ikke kan knyttes til enkeltpersoner.

Maskinlæring

For å forstå hvorfor kunstig intelligens trenger store mengder data, er det nødvendig å forstå hvordan selve læringen av systemet foregår.

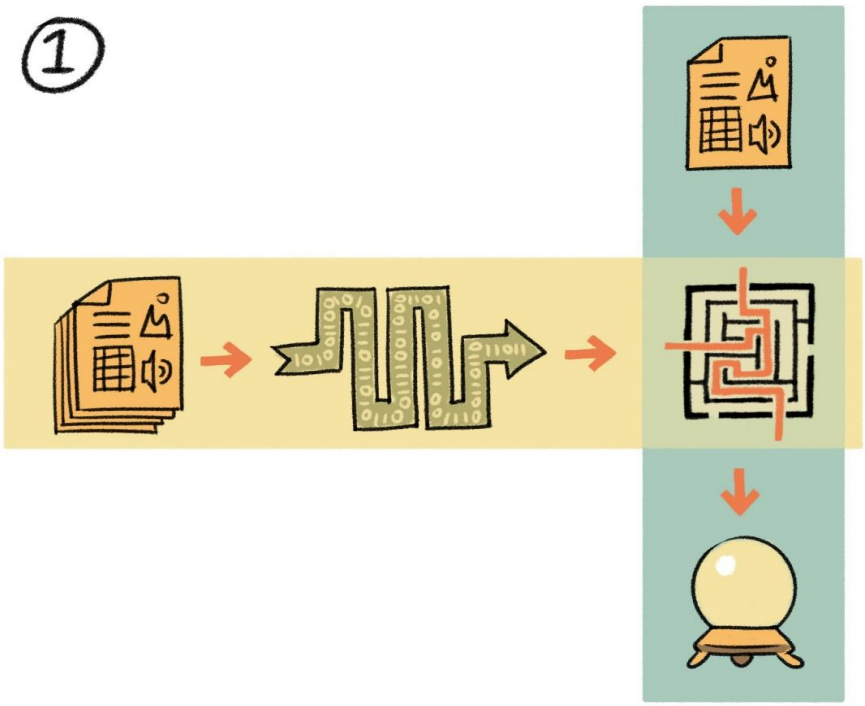
Opplæring av kunstig intelligens krever altså erfaring i form av data. På generell basis vil maskinlæring foregå slik (illustrert med figur 1, *fra venstre til høyre*):

1. Læringen starter med et utvalg av opplysninger som inneholder et mønster/likheter.
2. Ved hjelp av maskinlæring avdekkes mønstrene som finnes i opplysningene.
3. Det blir generert en modell som kan gjenkjenne mønstrene som er avdekket når nye data blir prosessert av modellen.

Modell er et samlebegrep på sluttresultatet av læring. Det finnes mange ulike typer modeller og det er disse som blir brukt i kommersielle applikasjoner — slik som å predikere hva slags type TV-serie en strømmings-bruker liker. Det som er felles for modellene er at de inneholder essensen av treningsdataene. Fordi fremtidige data som modellen skal behandle, sjelden vil være helt identiske med treningsdataene, ønsker man en generalisering. Enkelte data som skiller seg fra hovedmengden i treningsdataene, vil derfor normalt være fjernet fra modellen.

Bruk av modellen vil foregå slik: (illustrert med figur 1, *fra topp til bunn*)

1. Modellen mottar data av samme type som ble brukt til læringen.
2. Modellen avgjør hvilket mønster de nye dataene ligner mest på.
3. Modellen kommer med et estimert resultat.



Det finnes flere former for læring som kan benyttes, avhengig av om man har kategorisert informasjon eller ikke. Med kategorisert informasjon mener vi det som på engelsk refereres til som «labeled data». Det betyr at informasjonen kommer med merkelapper. Hvis vi tenker oss at datagrunnlaget er bilder, kan kategoriene eller merkelappene være for eksempel kjønn, etnisitet, hund eller katt.

Under har vi listet opp de vanligste hovedkategoriene av læring, og beskriver hvordan data blir benyttet i disse.

Veiledet læring (Supervised learning)

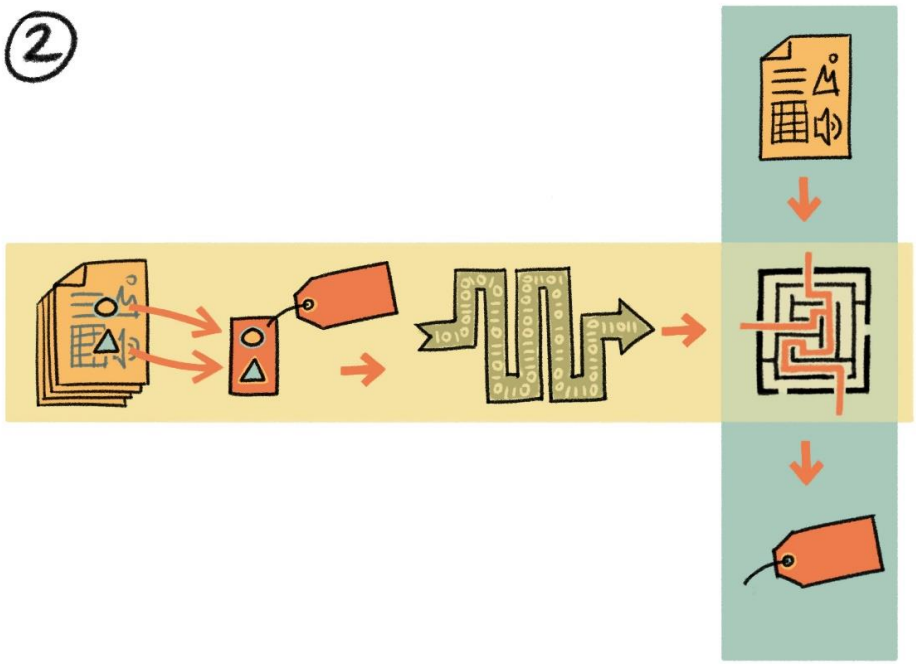
Veiledet læring betyr at det benyttes kategoriserte data. Veiledningen skjer da i form av de merkelappene som følger med dataene. Datasettet vil skilles i to, gjerne en 80/20 splitt. Deretter benyttes 80 prosent av dataene til å trene opp modellen. De siste 20 prosentene brukes til å verifisere hvor nøyaktig modellen er på ukjente data. Det hjelper ikke om modellen er helt nøyaktig på treningsdataene dersom den er lite nøyaktig på nye og ukjente data. Hvis modellen er for godt tilpasset treningsdataene, vil den normalt ikke kunne gi gode resultater på nye data. Dette kalles over-tilpassing («overfitting»). Derfor ønsker man en viss grad av generalisering i modellen.

Treningsdataene kan for eksempel være bilder, hvor det følger med informasjon om hva de enkelte bildene inneholder. Dette fungerer litt på samme måte som man lærer opp et barn. Vi peker ut en mengde objekter og setter navn på dem. Hvis det for eksempel pekes ut et antall katter for et barn, vil barnet etter hvert klare å gjenkjenne også andre katter enn de man tidligere har vist fram for barnet. På samme måte vil en maskinlæringsmodell opparbeide seg den samme evnen til å gjenkjenne objekter basert på kategoriserte bilder.

Dersom man jobber med et datasett og ønsker å skille menn og kvinner, kan forskjellig egenskaper være relevante. Hva man kan benytte kommer an på hvilket datagrunnlag man har. For eksempel lever kvinner i snitt lenger enn menn, derfor kan levealder være en relevant egenskap for å skille kjønn. Denne egenskapen vil nok vise seg litt snever i de fleste tilfellene, men er ment som et eksempel. Har man et datagrunnlag som består av bilder, kan hårlengde, bruk av sminke eller smykker være relevante egenskaper. I eksempelet under vises det hvordan to egenskaper ved data benyttes ved læring.

Læringen skjer slik (illustrert med figur 2, *venstre til høyre*):

1. Det benyttes et datasett med kategoriserte data.
2. Avhengig av datatype og hva som ansees som relevant kan det velges ut hvilke egenskaper ved dataene som skal benyttes (runding og trekant) til læringen. Dataene kommer også med en merkelapp som forteller hva som er riktig svar.
3. Det bygges en modell som basert på samme egenskaper vil gi en merkelapp.



For kategoriserte data vi ofte også vite hvilke egenskaper som er mest avgjørende for å sette dem i riktig kategori eller produsere rett svar. Ofte vil det være viktig å ha personer med god kjennskap til det aktuelle fagfeltet for å vite hvilke egenskaper som er mest relevante. Riktig utvalg av relevante egenskaper kan være langt viktigere enn mengden av data, noe vi kommer tilbake til senere. En fordel med kategoriserte data er at det blir lett å sjekke nøyaktigheten av modellen.

Når man tar i bruk modellen, skjer dette (figur 2*, topp til bunn*):

1. Nye data av samme type som treningsdataene legges inn i systemet.
2. De relevante egenskapene blir matet inn i modellen og bearbeidet.
3. Modellen produserer et resultat som er i samsvar med merkelappene i opplæringen.

Uveiledet læring (Unsupervised learning)

Ved uveiledet læring benyttes det data som ikke er forhåndskategoriserte. Ved denne typen læring ønsker man at systemet skal klare å gruppere data som ligner på hverandre. Hvis vi for enkelthets skyld igjen tar som eksempel at datamaterialet består av katte- og hundebilder, så ønsker man at de i størst mulig grad skal sorteres i to grupper – den ene med hundebilder og den andre med kattebilder.

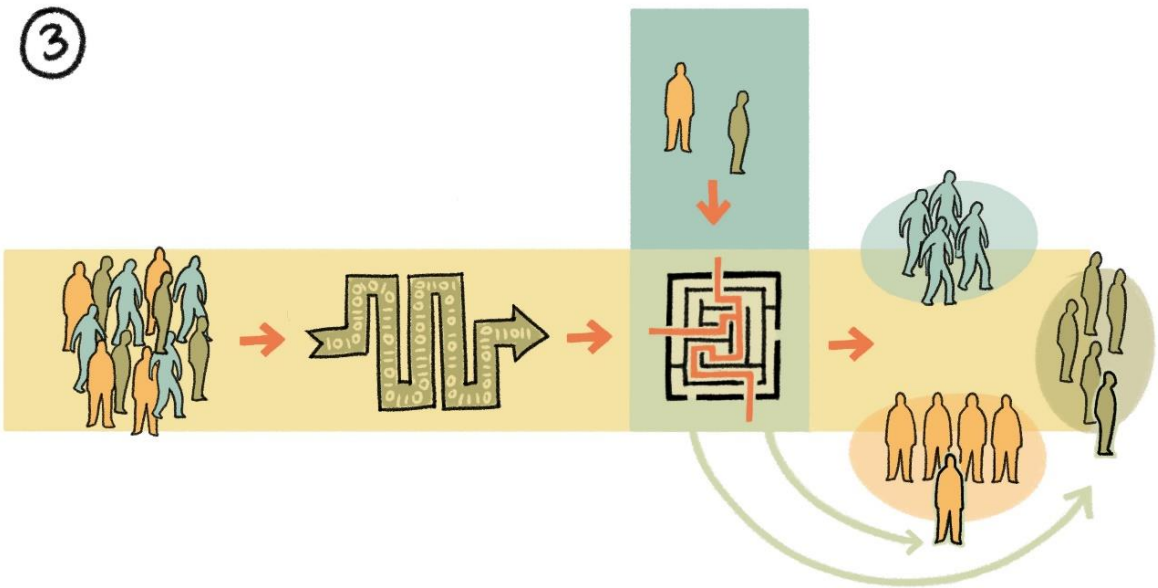
Læringen skjer på denne måten (figur 3, *venstre til høyre*):

1. Det hele starter med et datasett. For at datasettet skal gi mening, må det være en viss grad av likheter eller mønster i datagrunnlaget.
2. Mønstrene avdekkes.
3. En modell som kan gjenkjenne og skille mønstrene blir laget.

Ved bruk av modellen skjer dette (figur 3, *topp til bunn*):

1. Nye ikke-kategoriserte data av samme type som treningsdataene legges inn i systemet.
2. Modellen identifiserer mønsteret til dataene.
3. Modellen forteller hvilken gruppe de nye dataene tilhører.

En ulempe med dette er at modellen ikke kan plassere data i andre grupper enn de som ble oppdaget under læringsprosessen. Det er derfor veldig viktig at treningsgrunnlaget er representativt.



Forsterkende læring (Reinforcement learning)

Dette er en læringstype som baserer seg på prøving og feiling, samt at modellen optimaliseres ved å lære hvilke handlinger som leder mot målet. Det betyr at det er behov for mindre eller ingen data i læringen av systemet.

AlphaGo Zero

Vi har tidligere nevnt AlphaGo som et eksempel på maskinlæring. AlphaGo ble først trent opp med et datasett som besto av 30 000 Go-spill. For ytterligere å forbedre AlphaGo sin evne til å spille Go, ble den satt til å spille mot seg selv. På denne måten kunne erfaringsgrunnlaget økes betraktelig via prøving og feiling, uten at det var nødvendig med data fra flere kamper. Det ga også AlphaGo muligheten til å oppdage trekk og strategier som ikke var med i det opprinnelige treningssettet.

Den nyeste versjonen av AlphaGo – AlphaGo Zero – ble laget for å starte helt uten å benytte seg av treningsdata. Den ble kun programmert med reglene for Go, og fikk ikke starthjelp med informasjon om tidligere spilte kamper. Deretter lærte den ved å spille mot seg selv. Etter 40 dager var den dyktig nok til å slå den tidligere AlphaGo utgaven 100-0. Det er også interessant at Zero-versjonen av AlphaGo krever langt mindre regnekraft for å oppnå disse resultatene.

(Kilde: https://deepmind.com/blog/alphago-zero-learning-scratch/)

Resultatene av læring

Uavhengig av hvilke algoritmer eller metoder som benyttes for maskinlæringen, vil resultatet bli en «modell» som altså er et samlebegrep for all maskinlæring. Denne modellen kan så benyttes sammen med nye data og produsere et resultat av ønsket type. Dette kan for eksempel være en kategorisering, en grad av sannsynlighet eller lignende.

Det er verdt å merke seg at modellen normalt ikke inneholder kildedataene direkte. Den inneholder en aggregert representasjon av alle dataene som har blitt brukt til å lære opp systemet. Et unntak er beslutningstrær, der man vil kunne finne en varierende grad av datagrunnlaget i modellen. Begrensningene her ligger i hvordan man beskjærer treet etter læring, eller om man setter en nivåbegrensning før læring. Normalt vil man velge å gjøre en av disse tingene fordi man ønsker at modellen skal generalisere og ikke over-tilpasse. I en dyplæringsmodell vil datagrunnlaget bli representert som vektinger (tallverdier) i det nevrale nettverket. Det skal derfor ikke være mulig å hente ut igjen data, slik som for eksempel personopplysninger som har blitt brukt til å lære opp modellen. Vi ser nærmere på disse modellene litt senere, i avsnittet Den svarte boksen.

Modellbruk – statisk og dynamisk (offline/online)

Når en modell skal tas i bruk, kan det gjøres på to måter. Den første måten er at man tar i bruk en **statisk modell** (offline) som aldri endrer seg ved bruk. Den statiske modellen vil, som navnet tilsier, alltid oppføre seg likt og gi samme resultater gjennom hele livsløpet. All ny opplæring av modellen vil skje i et testmiljø, og alle endringer forutsetter at man erstatter modellen med en ny utgave. Dette betyr at man har full kontroll på modellen som er i bruk.

Den andre muligheten er en **dynamisk modell** (online). Modellen tas i bruk på samme måte som den statiske modellen. Forskjellen er imidlertid at den dynamiske modellen har mulighet til å benytte seg av inputdata for å forbedre og tilpasse seg endringer. Dette kan for eksempel være nødvendig i forbindelse med overvåking av kredittkorttransaksjoner for å avsløre svindel. Transaksjonene kan endre seg ut fra brukerens livssituasjon eller relatert til jobb, for eksempel ved at transaksjonene blir gjort på helt nye steder. Disse nye bruksmønstrene vil da med en statisk modell kunne bli kategorisert som mistenkelige og potensielt resultere i et sperret kredittkort. En modell vil derfor kunne bli mindre nøyaktig over tid om den ikke kan oppdateres kontinuerlig.

Filter for søppelpost er også et typisk bruksområde for en dynamisk modell, der brukeren selv kan forbedre modellen ved å angi feilkategoriserte eposter. Ulempen med dynamiske modeller, er en mindre grad av kontroll på utviklingen og at endringene straks får effekt. Et godt eksempel her er Microsoft sin chatbot Tay som lærte av samtaler «hun» hadde med internettbrukere. Etter kort tid på Twitter ble chatboten omtalt som en Hitler-

elskende sexrobot av media. Microsoft besluttet å fjerne Tay kun 24 timer etter at den ble lansert.[7]

Jo mer treningsdata, jo bedre?

Desto mer treningsdata vi kan fôre modellen med, jo bedre, er et typisk mantra vi ofte hører i forbindelse med maskinlæring. I de fleste sammenhenger krever datamaskinene langt mer data for opplæring enn det mennesker trenger for å lære det samme. Dette er per i dag en begrensing ved maskinlæring, noe som kompenseres ved å benytte betydelige datamengder – ofte større enn et menneske ville klart å håndtere.

Det er viktig å merke seg at kvaliteten på trenings-dataene, samt hvilke egenskaper man benytter, i mange sammenhenger kan være vesentlig viktigere enn kvantitet. Når en modell skal trenes opp, er det viktig at utvalget av data som modellen trenes med er representativt for oppgaven som senere skal løses. Det hjelper lite med enorme mengder data dersom de kun dekker en liten del av det modellen senere skal jobbe med.

I forbindelse med veiledet læring er riktig kategorisering veldig viktig. Hvis man har data som er feil klassifisert, vil det naturlig nok påvirke treningsresultatet negativt. Som det så klassisk heter; søppel inn gir søppel ut.

Datamengde i bredde og dybde

Effektiviteten av maskinlæringen kan bli sterkt påvirket av hvordan datagrunnlaget blir presentert for algoritmene som utarbeider modellene, og av hvilke egenskaper man velger å benytte.

På samme måte som i et regneark kan et datasett til maskinlæring bestå av rader og kolonner. Hvis man har noe som for eksempel relaterer seg til personer, kan kolonnene gjerne være alder, kjønn, bosted, sivilstand, høyde, vekt, nasjonalitet og så videre. Radene vil representere hvert enkelt individ. Her må det vurderes hvor mye informasjon man faktisk trenger om enkeltpersonene for å trene opp de modellene man ønsker, samt hvilken informasjon som er relevant for å oppnå formålet.

Når det skal velges ut relevante egenskaper (feature selection), vil det ofte være behov for personer med ekspertise på området. Det er ikke alltid at datagrunnlaget forteller alt.

Det er viktig å gjøre et godt utvalg, ellers risikerer man å ende opp med for mange egenskaper. Dette betegnes i fagmiljøene gjerne som «the Curse of Dimensionality». Enkelt sagt leder for mange egenskaper til at likhetene drukner i ulikheter. Dette vil lede til at det trengs enorme mengder data for å kompensere for dette.

En ulempe med å redusere utvalget av egenskaper er derimot at man kan gå glipp av sammenhenger som man ikke på forhånd vet om eller kan tenke seg til. Dette er noe av grunnen til at det er viktig å ha med personer med domenekunnskap inn i denne fasen. Det bør også gjøres en vurdering på hva som er et godt nok resultat.

Eksempel

På et amerikansk sykehus ble det gjort et forsøk på å kategorisere risiko for komplikasjoner for pasienter med lungebetennelse. Resultatet ble at pasienter som hadde *både* astma og lungebetennelse, ble kategorisert som lavrisikopasienter – til legenes store overraskelse.

Disse pasientene hadde høyere risiko, men bedre overlevelsesrate. Det modellen ikke klarte å fange opp, var at den tilsynelatende lave risikoen var et resultat av at disse pasientene fikk bedre oppfølging og mer intensiv behandling.

Dette viser noe av risikoen ved bruk av data uten domenekunnskap, og at datagrunnlaget ikke alltid forteller alt.

(Kilde: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf)

[7] http://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/

Skatteetaten

Skatteetaten har utviklet en prediktiv analyse for å hjelpe seg med å velge ut hvilke selvangivelser som bør kontrolleres for å oppdage feil eller juks. De testet om lag 500 ulike variabler som sa noe om skattyternes demografi, hendelser i livet og øvrige opplysninger i selvangivelsen. I den endelige modellen er det bare med 30 variabler. Det er blant annet opplysninger om bruk av fradragsposter i år og i fjor, alder, økonomiske forhold som inntekt og formue, samt og opplysninger knyttet til enkelte poster.

Dette er et godt eksempel på at ikke alt av tilgjengelig data alltid trenger å benyttes for å oppnå et formål. Uten at vi kjenner til hvordan Skatteetaten har jobbet med utvelgelse av kategorier for sitt prosjekt, ser vi at de valgte å bruke at begrenset utvalg, og de påpeker selv at dette var nok til å oppnå formålet med prosjektet.

(Kilde: Skatteetatens Analysenytt 1-2016, http://www.skatteetaten.no/globalassets/pdfer/skatteetatens_analysenytt/analysenytt-1_2016_web_hele.pdf)

Det er verd å nevne her at dyp læring skiller seg litt ut. Utvalget og tilpassing av egenskaper er ikke like viktig som ved andre læringsmetoder. Utvalget av egenskaper skjer eksempelvis via vektingen i et nevrale nettet. Ulempen ved å ikke gjøre tilpassinger er at man trenger en enorm økning i treningsdata.

Tilpasning av egenskaper (feature engineering)

Hvordan egenskapene i et datasett presenteres for maskinlæringen er en viktig faktor for å oppnå gode resultater. Relevante sammenhenger kan skjule seg hvis ikke dataene benyttes riktig. I mange tilfeller er det langt mer å hente på smart bruk av data, enn å øke mengden.

Et eksempel er datoer. La oss ta utgangspunkt i datoen 1.10.2017. Dette forteller oss at det er første dag i måneden og at det er måned nummer 10. Det kan godt tenkes at informasjonen er mer nyttig om det gjøres om til hvilken dag i uken det er, i dette tilfellet en søndag.

For Norge, som har fire ganske distinkte årstider, kan det tenkes at å gruppere månedene vil kunne representere dataene på en bedre måte. Måned 10 kan da representeres som høst. Høst kan igjen representeres som tallverdien 3, med utgangpunkt i vår=1, sommer=2, høst=3 og vinter=4. På denne måten kan man også avlede flere egenskaper fra et datapunkt, eller redusere antall forskjellige verdier. Hvis man henter datoer fra flere kilder må man forsikre seg om at de er i samme format. I data fra USA kan for eksempel måned være 1 og dag 10 i datoen 1.10.2017.

En normalisering av egenskaper kan også være nødvendig for at enkelte egenskaper ikke skal skape ubalanse i treningsdataene, eller for at et fåtall ekstrem-verdier ikke skal påvirke resten på en uønsket måte. Vi kan enkelt sett se på dette som at man sørger for at alt er i samme målestokk. Hvis man har egenskaper hvor en endring på 0.1 betyr like mye som en endring på 1000 for en annen egenskap, er det viktig å sørge for at de er tilpasset en felles skala.

Nok er nok?

Når det gjelder mengden data man trenger til læring så kan det være vanskelig å estimere før man starter. Dette påvirkes av hvilken type maskinlæring man benytter seg av, utvelgelse og antall egenskaper, samt kvalitet på datagrunnlaget. Det vil selvsagt også være relevant hvor nøyaktig en modell trenger å være for at formålet blir oppnådd. Dersom en person som gjør jobben er 75 prosent nøyaktig, vil så det samme være godt nok for modellen? Hvis man sikter mot en 100 prosent nøyaktighet vil man i mange sammenhenger trenge betydelige mengder data.

Bruksområdet kan her være førende for hva som er rimelig sett i forhold til bruk av personopplysninger til treningsdata. Hvis formålet er å diagnostisere dødelige sykdommer, vil det skille seg fra for eksempel å profilere noen for å servere dem en best mulig rettet annonse.

Skal man holde seg til dataminimeringsprinsippet, vil det være naturlig å starte med en begrenset mengde treningsdata, og så følge med på hvordan nøyaktigheten av modellen utvikler seg etterhvert som man mater inn nye data. Et verktøy som benyttes til denne vurderingen er læringskurver.[8] Ved bruk av læringskurver kan man

[8] https://www.coursera.org/learn/machine-learning/lecture/Kont7/learning-curves http://www.ritchieng.com/machinelearning-learning-curve/

starte med et mindre datasett og så se når kurven flater ut og nye data ikke gir noen merverdi i opplæringen.

Den svarte boksen

En av bekymringene i forbindelse med maskinlæring er at man ikke alltid vet hvordan resultatet blir produsert. Hvilke egenskaper, eller hvilke kombinasjoner av disse, var viktigst? Ofte vil modellen kun produsere et svar uten noen forklaring. Dette blir ofte kalt «den svarte boksen» – du vet ikke hva som skjer inne i boksen. Det interessante spørsmålet er om det er mulig å studere modellen, og på den måten finne ut hvordan den kom frem til det konkrete resultatet.

Skatteetaten har som nevnt bygd en prediktiv modell som hjelper dem med å velge ut hvilke selvangivelser de skal se nærmere på. De uttaler følgende: «Når vi bygger modellen på denne måten, vet vi ikke nødvendigvis hva det er som gjør at en bestemt skatteyter blir rangert til å ha stor risiko for feil. Rangeringen er et resultat av komplekse sammensetninger av dataene i modellen.»

Uttalelsen fra Skatteetaten underbygger at svart boks-problematikken er relevant. I dette tilfellet benyttes det kun 30 forskjellige egenskaper, men et system kan tenkes å benytte ekstremt mye mer enn det. Da vil det bli ytterligere vanskelig å finne hva som var relevant for utfallet.

Hvordan forstå og forklare hva som ligger bak?

Når det har blitt benyttet maskinlæring, er slutt-produktet en modell. Når det kommer til maskin-læringsmodeller, er det varierende hvor lett det er å ettergå resultatet, selv om man har benyttet samme treningsdata.

Dyp læring og nevrale nettverk er ofte det første som blir nevnt når det prates om svart boks-problematikk, uten at det setter avgrensningen for problemstillingen.

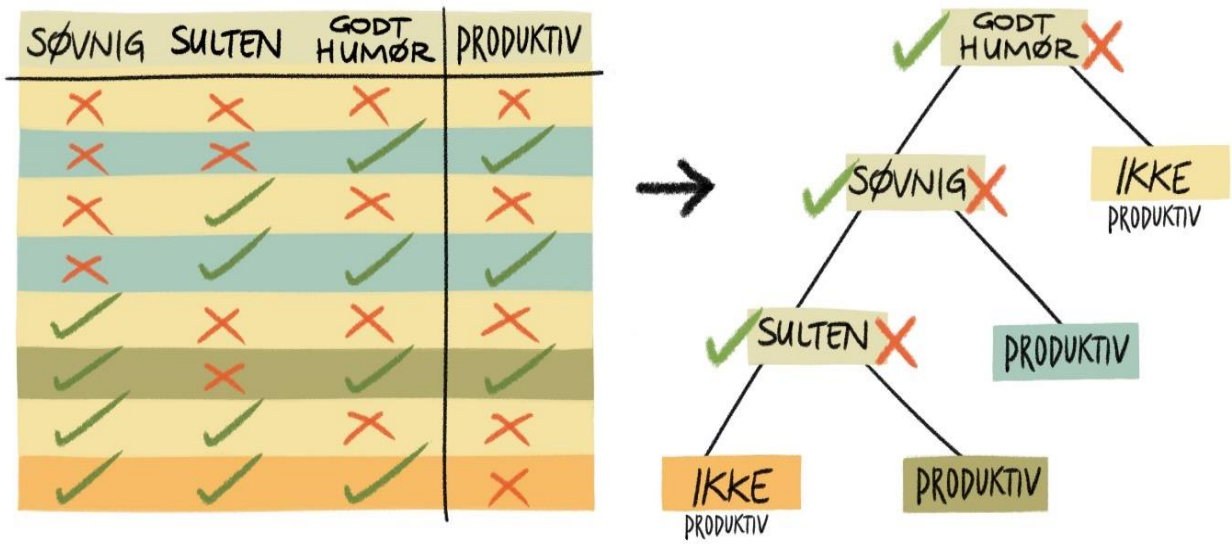
Vi vil her se på to eksempler som representerer to ytterpunkter for hvor enkle eller vanskelige modellene er å forstå og ettergå, nemlig såkalte beslutningstrær og dype nevrale nettverk:

Beslutningstrær er ofte enkle å forklare

Et beslutningstre er en av de enkleste modellene. I den enkleste formen for et beslutningstre, brytes alle data ned på en slik måte at de plasseres inn i et tre. Man starter på toppen, og basert på egenskapens verdi vil man på hvert nivå velge en gren. Dette gjøres helt til bunnen av beslutningstreet. Her finnes det endelige utfallet, altså beslutningen (se figur under).

Gjennomsiktigheten er stor for denne typen modell, i det minste når treet er basert på en overkommelig mengde data. Det vil være mulig å bevege seg oppover i treet for å se hvilke kriterier som ligger til grunn for resultatet.

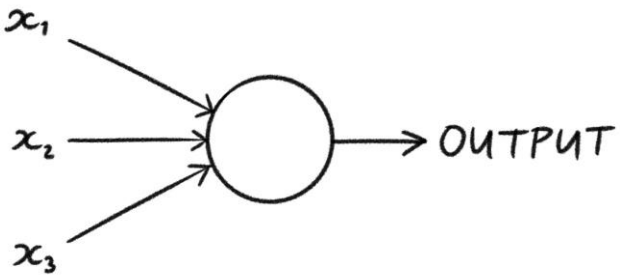
Ved økende datamengde vil man imidlertid komme til et punkt hvor det vil være vanskelig for et menneske å få oversikt og forståelse.



Nevrale nettverk er vanskelig å forklare

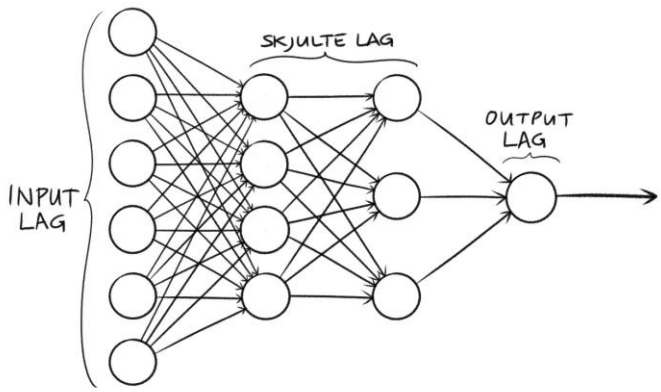
Bruk av såkalte nevrale nettverk, er en metode som i stor grad er inspirert av hvordan vi ser for oss at den menneskelige hjernen fungerer. De nevrale nettverkene bygges opp av en i utgangspunktet svært enkel komponent (perceptron), men til gjengjeld kan det benyttes svært mange komponenter. Dette kan lede til store og komplekse nettverk.

Perceptronet ser ut som illustrert nedenfor, med en varierende mengde innganger og én utgang:



Hvert «ben» på perceptronet vil ha en vekting. Denne vektingen avgjør hvor stor påvirkningskraft input-egenskapen som benyttes har på sluttverdien. Disse verdiene er det som justeres når nettverket trenes for å gi ønskede resultater. Ofte gjøres dette ved å jobbe seg bakover i nettverket for å justere vektingen på de nødvendige preceptronene slik at sluttresultatet blir riktig (backpropagation). Dette er en automatisert prosess som er en del av læringen.

Et nevralt nettverk vil bestå av tre deler; et input lag, et eller flere skjulte lag, og et output lag:



Hvis det er mer enn ett skjult lag, regnes det som dyp læring. I figuren over har vi et enkelt nevralt nettverk hvor alle inndata beveger seg kun fra venstre til høyre, og kommer ut som et resultat. Det finnes flere varianter av disse nevrale nettverkene. Noen vil danne løkker og sende data også fra høyre til venstre inne i nettverket, før det endelig resultatet produseres.

En av utfordringene her er at inndataene sees på isolert. I mange sammenhenger jobbes det med informasjon som har en kontekst. For eksempel kan enkelte ord ha forskjellig mening basert på kontekst. Den konteksten trenger heller ikke være i samme setning. Dette er noe av grunnen til at enkelte nevrale nettverk har en form for korttidshukommelse. Det gir dem mulighet til å produsere forskjellige utfall basert på hvilke data som ble behandlet rett før, noe som selvsagt også gjør det vanskeligere å finne ut hvordan et resultat ble utledet. Dette tilsier også at det å kun undersøke algoritmene for å finne ut hvordan de jobber og hvilke avgjørelser de produserer, kan være svært vanskelig.

Antall lag i ett nevralt nettverk kan variere. Som et eksempel kan det nevnes at Microsoft i 2016 vant en konkurranse for bildegjenkjenning med et nettverk som hadde 152 lag.[9] Størrelsen på nettverket og antall koblinger vil da bli et resultat av hvor mange input-verdier som eksisterer og hvordan lagene ble koblet sammen. Uansett kan man se for seg at størrelsen på det nevnte nevrale nettverket er langt utenfor hva man kan forstå eller undersøke uten egnede verktøy. Vi kommer tilbake til denne typen verktøy i siste kapittel.

[9] https://blogs.microsoft.com/ai/2015/12/10/microsoft-researchers-win-imagenet-computer-vision-challenge/

Kunstig intelligens møter personvernforordningen

Personvernforordningen har bestemmelser om den behandlingsansvarlige sine plikter og de registrerte sine rettigheter når det behandles personopplysninger. Forordningen vil derfor gjelde både når kunstig intelligens *utvikles* ved hjelp av personopplysninger, og når kunstig intelligens brukes for å *analysere* eller *ta avgjørelser* om enkeltpersoner.

I dette kapittelet vil vi gå gjennom de personvern-prinsippene og artiklene i personvernforordningen som er spesielt relevante ved utvikling og bruk av kunstig intelligens.

De grunnleggende personvernprinsippene

Reglene for behandling av personopplysninger bygger på noen grunnleggende prinsipper. I artikkel 5 i forordningen står prinsippene som gjelder for all behandling av personopplysninger. Kjernen i prinsippene er at opplysningene skal benyttes på en mest mulig personvernvennlig måte og at alle har rett til å bestemme over opplysninger om seg selv. Bruk av personopplysninger ved utvikling og bruk av kunstig intelligens, utfordrer flere av disse prinsippene.

Oppsummert betyr prinsippene at personopplysninger skal:

- behandles på en lovlig, rettferdig og gjennomsiktig måte (prinsippet om lovlighet, rettferdighet og gjennomsiktighet)
- samles inn for spesifikke, uttrykkelig angitte og berettigede formål og ikke behandles på en ny måte som er uforenlig med disse formålene (prinsippet om formålsbegrensning)
- være adekvate, relevante og begrenset til det som er nødvendig for formålene de behandles for (prinsippet om dataminimering)
- være korrekte og om nødvendig oppdaterte (prinsippet om riktighet)
- ikke lagres i identifiserbar form i lengre perioder enn det som er nødvendig for formålene (prinsippet om lagringsbegrensning)
- behandles på en måte som sikrer tilstrekkelig sikkerhet for personopplysningene (prinsippet om integritet og fortrolighet)

§ Personopplysning

Personopplysninger er alle opplysninger som kan knyttes til en enkeltperson. (Forordningen artikkel 4 nr. 1).

Opplysningene kan være *direkte* knyttet til enkeltpersonen, slik som for eksempel navn, fødselsnummer eller lokaliseringsopplysninger.

Opplysningene kan også være *indirekte* knyttet til en enkeltperson. Det betyr at personen kan identifiseres på bakgrunn av en kombinasjon av ett eller flere elementer som er spesifikke for personens fysiske, fysiologiske, genetiske, psykiske, økonomiske, kulturelle eller sosiale identitet.

§ Behandling

En behandling omfatter enhver *bruk* av person-opplysninger, for eksempel innsamling, registrering, lagring, organisering, strukturering, endring, bruk, utlevering, spredning eller andre former for tilgjengeliggjøring, sletting og så videre.

(Forordningen artikkel 4 nr. 2)

§ Behandlingsansvarlig

En behandlingsansvarlig er en fysisk eller juridisk person, offentlig myndighet, institusjon eller annet organ som alene eller sammen med andre bestemmer formålet med behandlingen og hvilke midler som skal benyttes.

(Forordningen artikkel 4 nr. 7).

I tillegg er det et prinsipp at den behandlingsansvarlige har ansvar for, og skal kunne påvise, at prinsippene overholdes (prinsippet om ansvarlighet).

Vi vil her gå gjennom de viktigste personvern-utfordringene ved utvikling og bruk av kunstig intelligens. Vi knytter disse utfordringene til de personvernprinsippene som er mest relevante for kunstig intelligens – nemlig prinsippene om rettferdighet, formålsbegrensning, dataminimering og gjennomsiktighet.

Skjeve algoritmer møter prinsippet om rettferdighet

Det er lett å tenke at kunstig intelligens vil kunne gjøre mer objektive analyser og dermed ta bedre avgjørelser enn et menneske. Kunstig intelligens blir tross alt ikke påvirket av lavt blodsukker, en dårlig dag eller et ønske om å tilgodese en venn.

Algoritmer og modeller er imidlertid ikke mer objektive enn menneskene som lager dem eller personopp-lysningene som benyttes i opplæringen. Modellens resultat kan bli uriktig eller diskriminerende dersom treningsdataene gir et skjevt bilde av virkeligheten, eller dersom de ikke er relevante for området de skal virke på. En slik behandling av personopplysninger vil være i strid med prinsippet om rettferdighet.

Prinsippet innebærer nemlig at all behandling av personopplysninger skal gjøres med respekt for de registrertes interesser, samt innenfor den registrertes rimelige forventninger om hva opplysningene skal brukes til. Prinsippet krever også at den behandlings-ansvarlige iverksetter tiltak for å hindre usaklig forskjellsbehandling av enkeltpersoner. I fortalen til forordningen beskrives det bruk av egnede matematiske eller statistiske fremgangsmåter som mulige tiltak.

Dette er imidlertid ikke tilstrekkelig for å ivareta prinsippet. Modellen må også trenes på relevante og riktige opplysninger og lære hvilke opplysninger som kan vektlegges og ikke. Modellen kan ikke legge vekt på opplysninger om rasemessig eller etnisk opprinnelse, politisk oppfatning, religion eller overbevisning, fagforeningsmedlemskap, genetisk status, helsetilstand eller seksuell orientering hvis det vil lede til usaklig forskjellsbehandling og diskriminering.

Ved mistanke eller påstand om at en modell leder til urettferdige eller diskriminerende resultater, kan Datatilsynet undersøke om prinsippet om rettferdig behandling av personopplysning er ivaretatt. Undersøkelsene kan for eksempel omfatte en gjennomgang av dokumentasjonen som begrunner valget av opplysninger, hvordan algoritmen er trent opp og om den er tilstrekkelig testet før den ble tatt i bruk.

Eksempel

I et system for utmåling av straff og kausjonsbetingelser i USA, kom det påstand om diskriminering ved hjelp av kunstig intelligens. Systemet blir brukt til å forutsi risikoen for at domfelte vil begå ny kriminalitet.

Tidsskriftet ProPublica studerte avgjørelser systemet hadde fattet, og konkluderte med at det diskriminerte afroamerikanere. Antallet afroamerikanere som feilaktig ble flagget med høyrisiko for å begå nye lovbrudd, var dobbelt så høyt som for hvite.

Selskapet som utviklet systemet var uenig i ProPublicas konklusjon, men ville ikke gi innsyn i hvilke kriterier og beregninger som inngikk i algoritmen. Det er derfor ikke mulig hverken for de domfelte eller offentligheten å få klarhet i hvorfor avgjørelsene blir som de blir.

(Kilde: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing)

Kunstig intelligens møter prinsippet om formålsbegrensning

Mange av modellene som utvikles med kunstig intelligens skal brukes til gode formål, slik som for eksempel diagnostisering av kreft. Er det fritt fram for gjenbruk av personopplysninger så lenge det gjøres for et godt formål?

Prinsippet om formålsbegrensning innebærer at formålet for behandlingen av personopplysninger må være tydelig angitt og fastsatt når personopplysningene

samles inn. Dette er helt grunnleggende for at de registrerte skal kunne ha kontroll over opplysningene sine. Beskrivelse av formålet med behandlingen er også en forutsetning for at den registrerte skal kunne ta et informert valg om å samtykke til behandlingen.

Utvikling og bruk av kunstig intelligens krever imidlertid ofte mange forskjellige typer personopplysninger – opplysninger som i noen tilfeller egentlig er samlet inn for andre formål. For eksempel kan det tenkes at en persons handlinger på Facebook inngår i en algoritme som avgjør om hun får boliglån av en bank. En slik gjenbruk av opplysninger kan være nyttig og gi mer presise analyser enn hva som har vært teknisk mulig tidligere, men det kan også være i strid med prinsippet om formålsbegrensning.

Når allerede innsamlede personopplysninger skal brukes på nytt, må den behandlingsansvarlige vurdere om det nye formålet er forenlig med det opprinnelige formålet. Hvis det ikke er tilfelle, kreves det et nytt samtykke eller et annet behandlingsgrunnlag. I det nevnte eksempelet om Facebook, bør den registrerte ha forstått og samtykket til at opplysninger fra handlinger på Facebook kan brukes av banken i vurderingen om boliglån for at viderebehandlingen skal være i samsvar med prinsippet om formålsbegrensning.

Ny teknologi – ny viten?

Prinsippet om formålsbegrensning er svært viktig for å ivareta de registrertes kontroll over egne opplysninger. Det finnes imidlertid unntak fra prinsippet. En viderebehandling anses for eksempel som forenlig med det opprinnelige formålet dersom det skjer for vitenskapelig eller historisk forskning, statistiske formål eller arkivformål i allmennhetens interesse. Dette reiser spørsmål om hva vitenskapelig forskning er, og om hvorvidt bruk og utvikling av kunstig intelligens er vitenskapelig forskning.

Stadig flere forskningsmiljøer ved universiteter og sykehus jobber med å utvikle verktøy som bruker kunstig intelligens. Det kan være modeller som finner risikomomenter for skatte- eller trygdesvindel, eller bildegjenkjenning som diagnostiserer kreft i svulster. Men hvor går grensen for hva som er vitenskapelig forskning?

Personvernforordningen definerer ikke hva som er vitenskapelig forskning. En alminnelig forståelse av begrepet tilsier imidlertid at det må dreie seg om et arbeid som har som mål å finne ny kunnskap eller ny viten.[10] Videre gir forordningens fortale (punkt 159) en anvisning om at vitenskapelig forskning bør tolkes vidt og for eksempel omfatte teknologisk utvikling og demonstrasjon, grunnleggende forskning, anvendt forskning og privatfinansiert forskning. Disse momentene tilsier at *utvikling* av kunstig intelligens – i noen tilfeller – kan anses som vitenskapelig forskning.

Bruk av kunstig intelligens til å for eksempel vurdere om en person er kredittverdig, kan derimot ikke sies å ha som formål å finne ny kunnskap. I dette tilfellet vil nok bruken av kunstig intelligens ikke kunne defineres som vitenskapelig forskning. Men kan man alltid skille mellom utvikling og bruk av kunstig intelligens?

Når den ferdige modellen er statisk (offline), kan utvikling og bruk skilles klart fra hverandre. En slik

§ Forenlig med opprinnelig formål?

I forordningens fortale (punkt 50) står det at følgende momenter bør inngå i vurderingen av om ny bruk av personopplysninger er forenelig med det opprinnelige formålet:

- enhver forbindelse mellom det opprinnelige formålet og formålene med den tiltenkte behandlingen
- i hvilken sammenheng personopplysningene er blitt samlet inn
- de registrertes forhold til den behandlingsansvarlige og hvordan det påvirker de registrertes rimelige forventningene om videre bruk
- personopplysningenes art
- konsekvensene av den tiltenkte viderebehandlingen for de registrerte
- om både de opprinnelige behandlingsaktivitetene og de tiltenkte viderebehandlingsaktivitetene omfattes av nødvendige garantier

Listen er ikke uttømmende og alle momenter som er relevante i hver enkelt sak, må inngå i vurderingen.

[10] Store Norske Leksikon

modell utvikles med treningsdata og testes med tilsvarende data før den tas i bruk. Når modellen blir tatt i bruk, er trenings- og testdata fjernet fra algoritmen og modellen behandler kun de personopplysningene som den anvendes på, for eksempel opplysninger om den som søker om lån. Fordi algoritmen er statisk, vil den ikke lære mer av de personopplysningene den anvendes på. Den vil heller ikke utvikle intelligensen etter at den er tatt i bruk.

Andre modeller utvikler og forbedrer seg kontinuerlig etter hvert som de tilføres flere personopplysninger. Dette kan for eksempel være modeller som gir beslutningsstøtte til leger. Modellen lærer noe nytt for hver pasient den mottar opplysninger om, eller hver vitenskapelige artikkel den leser. Den nye kunnskapen kan så brukes på neste pasient.

Når modellen utvikler seg kontinuerlig, er det vanskelig å skille mellom utvikling og bruk, og dermed hvor forskningen slutter og bruken starter. På det nåværende tidspunkt er det er derfor vanskelig å konkludere i spørsmålet om hvorvidt utvikling og bruk av disse modellene er vitenskapelig forskning eller ikke. Grensene for hva som er vitenskapelig forskning må gås opp i praksis etter at den nye personvernforordningen får virkning.

Vi understreker at bruk av personopplysninger for vitenskapelig forskning er underlagt egne regler i personvernforordningen (artikkel 89). Bruken skal være omfattet av nødvendige garantier for å sikre de registrertes rettigheter og friheter. Garantiene skal sikre tekniske og organisatoriske tiltak for særlig å sikre prinsippet om dataminimering.

Artikkel 89 åpner for at det i nasjonal rett kan gjøres unntak fra de registrertes rett til innsyn i, korrigering av og begrensning i bruken av opplysninger, samt til å fremme innsigelser. Det er ikke mulig å si hvordan løsningen blir i Norge før den nye loven er vedtatt.

Kunstig intelligens møter dataminimering

Utvikling av kunstig intelligens er ofte avhengig av store mengder personopplysninger.

Prinsippet om dataminimering stiller imidlertid krav om at opplysningene som brukes skal være adekvate, relevante og begrenset til det som er nødvendig for å oppnå formålet de behandles for. Det betyr at en behandlingsansvarlig ikke kan bruke flere personopplysninger enn det som faktisk er nødvendig, og at det må velges opplysninger som er relevante for formålet.

En utfordring ved utvikling av kunstig intelligens, er at det kan være utfordrende å definere formålet med bruken fordi man ikke kan forutse hva algoritmen vil lære. Formålet kan også endres etter hvert som maskinen lærer mer. Dette utfordrer prinsippet om dataminimering fordi det kan være vanskelig å definere akkurat hvilke opplysninger som er nødvendige.

Men prinsippet om dataminimering er mer enn et prinsipp om å begrense antall opplysninger som inngår i opplæringen eller bruken av en modell. Prinsippet inneholder også et krav om proporsjonalitet, altså at bruken av personopplysninger ikke må utgjøre et større inngrep overfor de registrerte enn det som er nødvendig. Dette kan for eksempel ivaretas ved å begrense graden av identifisering av individene som inngår i datagrunnlaget. Graden av identifisering begrenses både av *mengden* opplysninger og av *hvilke* opplysninger som brukes siden noen opplysninger sier mer om en person enn andre. Bruk av pseudonymiserings- eller krypteringsteknikker skjuler de registrertes identitet og bidrar til å begrense inngrepet.

Prinsippet tvinger også utviklere til å sette seg inn i området modellen skal fungere på og hvilke opplysninger som faktisk er nødvendige og relevante for å nå formålet. Videre må utviklere vurdere hvordan formålet kan oppnås på minst mulig inngripende måte overfor de registrerte. Vurderingene som gjøres må dokumenteres, slik at de kan legges frem for Datatilsynet ved et eventuelt tilsyn eller i forbindelse med forhåndsdrøftelse.

Selv om det er vanskelig å vite på forhånd hvilke opplysninger som er nødvendige og relevante for å utvikle og lære opp en algoritme – og dette kan endre seg underveis – er det svært viktig at prinsippet om dataminimering ivaretas gjennom fortløpende vurderinger av hvilke opplysninger som trengs. Dette gjelder både av hensyn til de registrerte, men også fordi irrelevante opplysninger utgjør en risiko for at sammenhenger som algoritmen finner ikke er reelle, men heller tilfeldige sammenhenger som ikke bør vektlegges.

Presset på bruken av personopplysninger øker i takt med at analyser basert på kunstig intelligens kan bidra til effektivisering og bedre tjenester. Datatilsynet mener at prinsippet om dataminimering bør ha en sentral rolle i utviklingen av kunstig intelligens slik at rettighetene til de registrerte og tilliten til modellene ivaretas.

§ Vurdering av personvernkonsekvenser

Alle som behandler personopplysninger må vurdere personvernkonsekvensene dersom det er sannsynlig at behandlingen medfører høy risiko for enkeltpersoners rettigheter og friheter. Dette gjelder særlig ved bruk av ny teknologi, og det skal tas hensyn til behandlingens art, omfang, formål og sammenhengen den utføres i.

Dersom risikoen er høy og den behandlingsansvarlige ikke kan begrense den, har han plikt til å starte forhåndsdrøftelser med Datatilsynet.

(Forordningen artikkel 35 og 36)

Den svarte boksen møter prinsippet om gjennomsiktig behandling

Personvern handler i stor grad om å ivareta den enkeltes rett til å bestemme over opplysninger om seg selv. Dette krever at de behandlingsansvarlige er åpne om at de bruker personopplysninger – at bruken er gjennomsiktig.

Gjennomsiktighet oppnås blant annet ved å gi informasjon om behandlingen til de registrerte. De registrerte må få informasjon om hvordan opplysningene brukes, enten opplysningene hentes inn fra den registrerte selv eller fra andre (forordningen artikkel 13 og 14). Informasjonen må dessuten være lett tilgjengelig, for eksempel på en hjemmeside, og være skrevet i et klart og forståelig språk (forordningen artikkel 12). Informasjonen skal gjøre de registrerte i stand til å bruke sine rettigheter etter personvern-forordningen.

Det kan være utfordrende å oppfylle prinsippet om gjennomsiktighet når kunstig intelligens utvikles og brukes. Avanserte former for kunstig intelligens er vanskelig både å forstå og forklare, og kan gjøre det tilnærmet umulig å forklare hvordan opplysninger blir koblet og vektlagt i en spesifikk behandling.

Det er også en utfordring at informasjon om modellen kan røpe forretningshemmeligheter og immaterielle rettigheter, noe forordningens fortale (punkt 63) presiserer at innsynsretten ikke skal gjøre. Hensynet til andres rettigheter, slik som en virksomhets forretnings-hemmeligheter, kan likevel ikke brukes til å nekte en registrert innsyn i alle opplysningene om henne. Her gjelder det å finne en pragmatisk løsning. I de fleste tilfeller vil det være uproblematisk å gi den registrerte informasjonen hun trenger for å ivareta sine interesser uten å samtidig avsløre forretningshemmeligheter.

Til tross for at kunstig intelligens er komplisert å forstå og forklare, gjelder prinsippet om gjennomsiktig behandling av personopplysninger fullt ut ved utvikling og bruk av kunstig intelligens.

Vi vil nedenfor vil vi se nærmere på informasjonsplikten og den registrertes rettigheter.

Generell informasjon

Når det samles inn personopplysninger må den behandlingsansvarlige alltid gi en del *generell informasjon* slik som

- identiteten til den behandlingsansvarlige
- hvordan den behandlingsansvarlige kan kontaktes
- formålet med behandlingen
- det rettslige grunnlaget for behandlingen
- hvilke kategorier av personopplysninger som blir behandlet
- samt de registrertes rett til innsyn i opplysningene

Det må også gis informasjon om risikoer, regler, garantier, de registrertes rettigheter i forbindelse med behandlingen, og hvordan disse rettighetene kan utøves.

Det utløses i tillegg en *utvidet informasjonsplikt* når det samles inn personopplysninger for automatiserte avgjørelser. Bruk av kunstig intelligens er en form for automatisert behandling, og i noen tilfeller er det også modellen som tar avgjørelsen. Det er viktig å avklare hva som kreves for å anse en avgjørelse som automatisert, før vi ser nærmere på den utvidede informasjonsplikten.

Individuelle automatiserte avgjørelser

Individuelle automatiserte avgjørelser er beslutninger om enkeltpersoner som baseres på maskinell behandling. Et eksempel kan være når det gis fartsbot kun på bakgrunn av kamerabevis fra automatisk fartskontroll. I personvernforordningen artikkel 22 er automatiserte avgjørelser definert og regulert.

Utgangspunktet er at det ikke er tillatt å bruke individuelle automatiserte avgjørelser. Det er imidlertid unntak fra forbudet dersom den automatiserte avgjørelsen er nødvendig for å inngå en avtale, er tillatt i lovgivning eller er basert på et uttrykkelig samtykke fra den registrerte. Forordningen definerer ikke hva som utgjør et uttrykkelig samtykke i motsetning til et vanlig samtykke, men ordlyden tilsier at det må dreie seg om en eksplisitt handling fra den registrerte.

For å oppfylle personvernforordningens krav, må avgjørelsen være **basert utelukkende på automatisert behandling,** og den må ha **rettsvirkning** eller på tilsvarende **måte i betydelig grad påvirke en person**.

At den automatiserte avgjørelsen må være basert utelukkende på automatisert behandling, betyr at det ikke kan være noen form for menneskelig inngripen i avgjørelsesprosessen. For at noe skal falle inn under begrepet «menneskelig inngripen», må et menneske ha gjort en selvstendig vurdering av opplysningene som ligger til grunn, og ha myndighet til å overprøve eventuelle anbefalinger som modellen har kommet med. Reglene om automatiserte avgjørelser kan ikke omgås ved å fabrikkere menneskelig inngripen.

Hva som menes med rettsvirkning er ikke definert i forordningen. En naturlig forståelse av ordlyden vil være at den automatiserte avgjørelsen må påvirke den registrertes rettigheter eller plikter, for eksempel rettigheter som er fastsatt i lov eller i kontrakt. Se eksempler i faktaboksen.

Alternativet om at den automatiserte avgjørelsen på tilsvarende måte i betydelig grad påvirker en person, er heller ikke nærmere definert. Vi antar at avgjørelsen må ha potensiale til å påvirke omstendighetene, oppførselen eller valgene til personen som er gjenstand for den automatiserte avgjørelsen. Det er imidlertid vanskelig å si akkurat hvor terskelen går, blant annet fordi vurderingen har betydelige subjektive elementer.

Når det benyttes automatiserte avgjørelser må det iverksettes tiltak for å verne den registrertes rettigheter, friheter og berettigede interesser. Den registrerte må kunne kreve at et menneske tar den endelige avgjørelsen, og hun må kunne klage på den.

Automatiserte avgjørelser som **involverer særskilte kategorier av personopplysninger (sensitive personopplysninger)** er bare tillatt dersom den registrerte samtykker til det, eller hvis det finnes hjemmel i lov.

Det er viktig å være oppmerksom på at en sammenstilling av ulike personopplysninger kan gi sensitiv informasjon om enkeltpersoner, noe som vil være en behandling av særskilte personopplysninger.

§ Forordningens artikkel 22

Vår tolkning av artikkel 22 er basert på det siste utkastet til Artikkel 29-gruppens retningslinjer for automatiserte avgjørelser.

Utkastet er basert på innspill fra 64 berørte virksomheter, og skal etter planen publiseres i begynnelsen av februar 2018.

I Artikkel 29-gruppen møter representanter fra EU-landenes datatilsynsmyndigheter. Norge har observatørstatus som EØS-land. Gruppens uttalelser tillegges normalt stor vekt.

(Article 29 Data Protection Working Party: xx/2017 on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679)

Eksempler

Rettsvirkning:

- om du får innreiseforbud til et land
- om du oppfyller vilkårene for å motta dagpenger eller sosialstønad, eller
- om strømtilførselen kuttes fordi regningene ikke er betalt

Avgjørelser som på tilsvarende måte i betydelig grad påvirker en enkeltperson:

- automatisk avslag på søknad om kreditt på internett
- e-rekruttering uten menneskelig inngripen

En studie kombinerte «likes» på Facebook med informasjon fra en enkel spørreundersøkelse og predikerte med 88 prosent sikkerhet mannlige brukeres seksuelle orientering. Videre predikerte de etnisitet med 95 prosent nøyaktighet og om en bruker var kristen eller muslim med 82 prosent nøyaktighet.[11] Dette vil utløse de samme pliktene etter personvernforordningen som om det fra starten av ble behandlet sensitive person-opplysninger.

§ Særskilte kategorier personopplysninger

Særskilte kategorier av personopplysninger er opplysninger om rasemessig eller etnisk opprinnelse, politisk oppfatning, religion, overbevisning eller fagforeningsmedlemskap, samt behandling av genetiske opplysninger og biometriske opplysninger med det formål å entydig identifisere en person, helse-opplysninger eller opplysninger om en persons seksuelle forhold eller seksuelle orientering.

(Forordningen artikkel 9)

Rett til informasjon ved individuelle automatiserte avgjørelser

I tillegg til å gi den generelle informasjonen som er nevnt over, må det gis informasjon om at personopplysningene samles inn for å brukes i en automatisert avgjørelse. Det må også gis relevant informasjon om modellens underliggende logikk, samt betydningen og forventede konsekvenser av den automatiserte behandlingen.

At det må gis informasjon modellens logikk, betyr at det må gis informasjon om for eksempel beslutningstrær som brukes, hvordan opplysningene vektlegges og hvordan opplysninger kobles. Fordi informasjonen skal være forståelig for den registrerte, er det ikke alltid nødvendig å gi en omfattende forklaring av algoritmen, eller å legge frem selve algoritmen.

Det må også gis informasjon om hvordan den automatiserte avgjørelsen kan påvirke den registrerte. Et forsikringsselskap som bruker automatiserte avgjørelser for å fastsette forsikringspremien for bil basert på kundens kjøremønster, bør derfor informere om konsekvensene av behandlingen og at farlig kjøring kan føre til høyere forsikringspremie.

Denne informasjonen skal gis før den automatiserte behandlingen starter og er avgjørende for at den registrerte skal kunne fremme innsigelse mot behandlingen eller gi sitt samtykke til den.

Rett til forklaring på den automatiserte avgjørelsen?

Men har den registrerte krav på å få forklaring på avgjørelsens innhold etter at den er truffet, altså hvordan modellen kom til resultatet?

I fortalen står det at nødvendige garantier ved automatisert behandling skal omfatte «spesifikk informasjon … og rett til … å få en forklaring på avgjørelsen som er truffet etter en slik [automatisert] vurdering» (punkt 71). Fortalen sier dermed at den registrerte har krav på en forklaring på hvordan modellen kom frem til resultatet, det vil si hvordan opplysningene er vektet og vurdert i de konkrete tilfellene.

En slik rett til forklaring kommer derimot ikke frem av selve forordningen. Det er ikke klart hva de språklige forskjellene mellom fortale og artiklene innebærer,[12] men fortalen i seg selv er ikke juridisk bindende og gir ikke alene en rett til forklaring.

Uavhengig av hva de språklige forskjellene innebærer, må den behandlingsansvarlige gi så mye informasjon som er nødvendig for at den registrerte skal kunne bruke sine rettigheter. Dette betyr at avgjørelsen må forklares slik at den registrerte kan forstå resultatet.

Retten til forklaring betyr ikke nødvendigvis at den svarte boksen må åpnes, men forklaringen må gjøre den registrerte i stand til å forstå hvorfor en avgjørelse ble som den ble, eller hva som må endres for at avgjørelsen

[11] Michael Kosinski, David Stilwell and Thore Graepel. «Private traits and attributes are predictable from digital records of human behaviour. Proceedings of the National Academy of Sciences of the United States of America»: http://www.pnas.org/content/110/15/5802.full.pdf

[12] Se for eksempel Andre Burt, «Is there a right to explanation for machine learning in the GDPR?»: https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/ cf. Sandra Wachter, Brent Mittelstadt, Luciano Floridi, International Data Privacy law, forthcoming, «Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation», tilgjengelig på https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2903469

skal bli annerledes.[13] Det må gis informasjon om hvordan den registrerte kan angripe avgjørelsen, enten ved å klage eller ved å kreve menneskelig inngripen.

Rett til forklaring når det ikke er en automatisert avgjørelse?

Noen ganger skjer det en automatisert *behandling* som ikke leder til en automatisert *avgjørelse*. I stedet bruker et menneske informasjon fra den automatiserte behandlingen for å ta en avgjørelse, for eksempel ved bruk av et beslutningsstøtteverktøy. Vilkårene for å være en automatisert avgjørelse er dermed ikke oppfylt. Spørsmålet blir da om den registrerte likevel har krav på den samme forklaringen som når avgjørelsen er automatisert.

Det finnes ingen egne artikler i forordningen eller uttalelser i fortalen om retten til forklaring av en spesifikk avgjørelse når vilkårene for automatiserte avgjørelser ikke er oppfylt.

Den registrerte har uansett krav på å få den informasjonen som er nødvendig for at hun skal kunne ivareta sine rettigheter. Prinsippet om gjennomsiktighet stiller også krav til informasjon.

Innsynsretten gir også den registrerte rett å få informasjon om personopplysningene som er brukt i avgjørelsen. Den gir imidlertid ingen rett til en forklaring av avgjørelsen.

Selv om det ikke finnes en rett til forklaring når avgjørelsen ikke er automatisert, tilsier prinsippet om gjennomsiktighet at den behandlingsansvarlige bør gi en tilsvarende forklaring som ved automatiserte avgjørelser.

§ Annet relevant regelverk

I tillegg til personvernforordningen er det også andre regelverk som krever at avgjørelsen må begrunnes.

Offentlig sektor er for eksempel underlagt forvaltningsloven som blant annet krever at enkeltvedtak skal begrunnes. Den det gjelder har rett på informasjon om hvilke regler og faktiske forhold vedtaket bygger på, samt hvilke hovedhensyn som har vært avgjørende (forvaltningsloven § 24 og 25).

[13] Se for eksempel Sandra Wachter, Brent Mittelstadt og Chris Russel, «Counterfactual explanations without opening the black box: automated decisions and the GDPR».

Tilsyn med algoritmene

I tiden fremover vil vi oppleve at flere og flere av beslutningene om oss blir tatt ved hjelp av kunstig intelligens. Det kan være avgjørelser som gjelder om du kan få ta opp lån, hva prisen på bilforsikringen skal være eller hvilket innhold nettavisen viser deg. Samtidig blir det stadig vanskeligere å ha oversikt over og innsikt i de komplekse systemene som tar avgjørelsene om oss. For at folk skal ha tillit til at tjenestene behandler opplysningene deres på en god måte, er vi avhengige av at de som tilbyr tjenestene følger personvernreglene.

Datatilsynet har i oppgave å føre tilsyn med at virksomheter både i privat og offentlig sektor følger personvernregelverket. Men hvordan fører man egentlig tilsyn med en algoritme som kanskje er skjult i en svart boks?

Datatilsynets tilsynskompetanse

Personvernforordningen bestemmer hvilken undersøkelsesmyndighet Datatilsynet har i forbindelse med tilsyn. For å undersøke om personopplysninger behandles i samsvar med regelverket, kan Datatilsynet blant annet gjennomføre personvernrevisjon, også kalt tilsyn. Et tilsyn skal klarlegge om den behandlingsansvarlige har rutiner og retningslinjer som skal sikre at regelverket etterleves, og om rutiner og retningslinjer følges.

I forbindelse med et tilsyn kan representanter fra Datatilsynet kreve å få all informasjon som trengs for å utføre sine oppgaver. Det kan være dokumentasjon på organisatoriske og tekniske tiltak, risikovurderinger, personvernkonsekvensvurderinger, hvordan henvendelser fra de registrerte skal følges opp og opplæring av de ansatte.

Representantene kan også kreve å få adgang til lokaler, databehandlingsutstyr og -midler, samt til personopplysningene som behandles. Adgang til lokaler og utstyr skal skje i samsvar med nasjonale prosessregler. Datatilsynet har i sitt høringssvar til forslaget til ny personopplysningslov ytret at det bør vurderes å gi tilsynet samme bevissikringsmuligheter som for eksempel Konkurransetilsynet har i dag.

Hva skal Datatilsynet kontrollere hos en aktør som bruker kunstig intelligens?

En aktør som utvikler eller bruker kunstig intelligens, står overfor akkurat de samme kravene i loven som en virksomhet som bruker et system som ikke benytter kunstig intelligens. På et gjennomsnittlig tilsyn vil Datatilsynet blant annet sjekke om virksomheten har behandlingsgrunnlag, om den har tilfredsstillende internkontroll og rutiner, at det er gjennomført risikovurderinger og at det er iverksatt tekniske og organisatoriske tiltak for å sikre opplysningene.

Det er noen tema som kan være spesielt aktuelle å kontrollere hos virksomheter som benytter kunstig intelligens, slik som overholdelse av prinsippene som er beskrevet tidligere i denne rapporten; at data ikke benyttes til nye formål uten ha å tilstrekkelig behandlingsgrunnlag, at en virksomheter ikke behandler flere personopplysninger enn de trenger, at det er iverksatt tiltak for å sikre rettferdig behandling og at de registrerte informeres slik loven krever.

Hvis en virksomhet utvikler kunstig intelligens, kan det også være aktuelt å kontrollere hvor mye og hvilke treningsdata som brukes, samt hvordan disse benyttes i treningsprosessen. Hvis en virksomhet bruker et system basert på kunstig intelligens, kan det være relevant å sjekke om virksomheten tester resultatene og reviderer systemet for å sikre at det ikke bruker personopplysninger på en ulovlig eller diskriminerende måte. Det vil også være relevant å undersøke om systemet er utviklet basert på prinsippene for innebygd personvern.

Hvor dypt går et tilsyn?

I de fleste tilsynssituasjoner vil det være tilstrekkelig for Datatilsynet å innhente dokumentasjon på at virksomheten følger regelverket. En virksomhet må kunne forklare og dokumentere, og i noen tilfeller demonstrere, at de behandler personopplysninger i henhold til reglene. Det betyr at en virksomhet må vite hvordan et system behandler personopplysninger og kunne redegjøre for dette. Hvis en virksomhet ikke kan redegjøre for bruken av personopplysninger, har Datatilsynet blant annet myndighet til å gi bøter og stoppe behandlingen.

Hvis Datatilsynet mistenker at en redegjørelse inneholder feil eller uriktig informasjon, kan vi be virksomheten verifisere opplysningene i rutiner og vurderinger, for eksempel ved at virksomheten må demonstrere hvordan et system behandler person-opplysninger. Dette kan for eksempel være aktuelt der det er mistanke om at en algoritme bruker opplysninger virksomheten ikke har behandlingsgrunnlag for, eller ved mistanke om at algoritmen kobler sammen opplysninger som fører til en beslutning som er diskriminerende.

Per i dag gjennomfører Datatilsynet få kontroller av IT-systemer når vi er på tilsyn. I noen tilfeller der det er behov sjekker vi hva som skjer inne i et system, for eksempel for å se hvor lenge kameraopptak lagres. Vi forventer at behovet for å kontrollere IT-systemer vil øke i årene fremover i takt med økt bruk av automatiserte analyser og beslutninger i alle sektorer. Personvernforordningen legger dessuten større vekt på ansvarlighet og internkontroll hos behandlings-ansvarlige og mindre vekt på forhåndskontroll fra Datatilsynet.[14]

Hvordan føre tilsyn med en svart boks?

«Vanlige» algoritmer er i stor grad relativt enkle å forholde seg til. De er programmert til å utføre noen spesifikke handlinger. Hvis for eksempel inntekten din er x og din eksisterende gjeld er y, kan du ta opp et lån på beløp z. Dette er et forenklet eksempel, men det viser at det er mulig å se hva input er og hvordan dataene kobles sammen for å få et gitt resultat.

Modeller som er basert på for eksempel dyp læring og nevrale nettverk, er derimot komplekse og lite gjennomsiktige, noe som gjør det utfordrende å kontrollere hva som faktisk skjer inne i systemet. Dette krever tilstrekkelig kunnskap om kunstig intelligente systemer for å kunne vite hva man skal se etter, samt hvilke spørsmål som er relevante å stille. I en tilsynssituasjon der det er behov for å ta en dypere titt inn i systemet, trengs det avansert teknisk kompetanse.

Av ressursmessige hensyn kan det være en løsning å innhente ekstern kompetanse i de tilfellene der tilsynet trenger å gjøre en «dyp» kontroll av et system som er basert på kunstig intelligens. Det er viktig at tilsynet har både kunnskap og ressurser for å kunne avdekke brudd på regelverket, slik som å unngå at algoritmer forsterker sosiale forskjeller eller fører til utilsiktet diskriminering, samt ulovlig gjenbruk av data.

[14] Se veileder om virksomhetens ansvar etter personvernforordningen på Datatilsynets nettsider, https://www.datatilsynet.no/regelverk-og-skjema/veiledere/virksomhetens-ansvar-etter-nytt-regelverk

Løsninger og anbefalinger

Et personvernprinsipp som ligger til grunn for all utvikling og bruk av kunstig intelligens, er ansvarlighet. Dette prinsippet står sentralt i personvernforordningen og legger et større ansvar på den behandlingsansvarlige når det gjelder å sørge for at all behandling skjer i samsvar med regelverket, samt dokumentasjon av dette. Kravet om ansvarlighet gjelder også databehandlere.

Vi vil i dette kapittelet komme med eksempler på verktøy og løsninger som kan hjelpe den behandlingsansvarlige å etterleve regelverket. Først vil vi omtale to av kravene i det nye personvernregelverket som er spesielt viktige i forbindelse med bruk og utvikling av kunstig intelligens; vurdering av personvernkonsekvenser og innebygd personvern. Deretter følger eksempler på verktøy og metoder som kan hjelpe til med å ivareta personvernet i løsninger som bruker kunstig intelligens. Til slutt kommer anbefalinger til utviklere, løsningsleverandører, virksomheter som kjøper og bruker kunstig intelligens, sluttbrukere og myndigheter.

Vurder personvern-konsekvensene – og bygg personvernet inn i løsningene!

De nye personvernreglene styrker enkeltpersoners rettigheter. Samtidig blir virksomhetenes plikter skjerpet. To nye krav som er spesielt relevante for virksomheter som tar i bruk kunstig intelligens, er kravet om innebygd personvern og vurdering av personvernkonsekvenser.

Innebygd personvern

Den behandlingsansvarlige skal bygge personvern inn i løsningene sine og sørge for at personvernet er ivaretatt i standardinnstillingene. Disse kravene er beskrevet i artikkel 25 i forordningen og gjelder ved utvikling av programvare, bestilling av nye systemer, løsninger og tjenester, samt videreutvikling av disse.

Regelverket krever at det tas hensyn til personvern i alle utviklingsfaser av et system, i rutiner og i den daglige bruken. Standardinnstillinger skal settes mest mulig personvernvennlige, og man skal bygge inn personvern-hensyn inn allerede i utviklingsfasen av løsningen.[15] Prinsippet om dataminimering nevnes uttrykkelig i bestemmelsen om innebygd personvern.

Vurdering av personvernkonsekvenser

Alle som behandler personopplysninger har en plikt til å vurdere risikoen knyttet til behandlingen. Dersom en virksomhet mener at en planlagt behandling sannsynligvis vil utgjøre en høy risiko for enkeltpersoners rettigheter og friheter, har den plikt til å gjennomføre en vurdering av personvernkonsekvenser (DPIA). Dette er beskrevet i forordningens artikkel 35.

Når risikoen vurderes, skal det tas hensyn til arten, omfanget, sammenhengen og formålet med behandlingen. Det må også tas hensyn til om det benyttes ny teknologi. Det er dessuten et krav om å utrede personvernkonsekvenser når det gjøres en systematisk og omfattende vurdering av personlige forhold i de tilfellene opplysningene brukes til automatiserte avgjørelser, eller når det behandles særskilte kategorier av personopplysninger (sensitive personopplysninger) i stort omfang. Også systematisk overvåking av offentlig område i stort omfang forutsetter dokumentasjon for at personvernkonsekvenser er blitt gjennomført.

Konsekvensanalysen skal som et minimum inneholde:

- en systematisk beskrivelse av behandlingen, dens formål og eventuelt hvilken berettiget interesse den ivaretar
- en vurdering av om behandlingen er nødvendig og forholdsmessig, sett opp mot formålet
- en vurdering av risikoen behandlingen har for personers rettigheter, herunder retten til personvern
- hvilke tiltak som skal settes i verk mot risikoen som er identifisert

Datatilsynet skal involveres i forhåndsdrøftelser dersom konsekvensanalysen viser at den planlagte behandlingen kan utgjøre en høy risiko for de registrerte, og at

[15] Les Datatilsynets veileder om programvareutvikling med innebygd personvern: https://www.datatilsynet.no/regelverk-og-skjema/veiledere/programvareutvikling-med-innebygd-personvern/

risikoen ikke kan reduseres ved at den behandlingsansvarlige iverksetter tiltak (forordningen artikkel 36).

Verktøy og metoder for godt personvern i kunstig intelligens

Kunstig intelligens er en teknologi i rask utvikling. Det samme gjelder verktøy og metoder som kan hjelpe til med å løse personvernutfordringene ved bruk av kunstig intelligens. Vi har samlet et utvalg eksempler for å illustrere noen mulighetene som finnes. Metodene er ikke evaluert etter praktisk bruk, men vurdert ut ifra et mulig potensiale. Det betyr at teknikken kanskje kan være uegnet i dag, men at konseptet er spennende og har et potensiale i seg for videre forskning og fremtidig bruk.

Vi har kategorisert metodene i tre kategorier:

- Metoder for å redusere behovet for treningsdata.
- Metoder som ivaretar personvernet uten at datagrunnlaget reduseres.
- Metoder for å unngå svart boks-problematikken.

1. Metoder for å redusere behovet for treningsdata

En av utfordringene vi har pekt på i denne rapporten, er at det ofte er behov for store mengder med data når man skal benytte seg av maskinlæring. Ved å foreta et riktig utvalg av egenskaper, og gjøre en god tilpassing av disse, kan behovet for data reduseres. I tillegg er dette et utvalg andre metoder:

Generative Adversarial Networks[16]
Generative Adversarial Networks (GAN) er en metode for å generere syntetiske data. Per i dag har dette i all hovedsak blitt benyttet til generering av bilder. GAN har imidlertid potensiale til å være en metode for å generere store mengder syntetiske treningsdata av høy kvalitet også på andre områder. Dette vil dermed kunne imøtekomme behovet for både kategoriserte data og store mengder med data, uten at det er nødvendig å benytte store mengder med reelle personopplysninger.

Federated learning[17]
Dette er en form for distribuert læring. Federated learning fungerer ved at den siste versjonen av en hovedmodell lastes ned til en klientenhet, for eksempel en mobiltelefon. Deretter forbedres modellen lokalt på klientenheten basert på lokale data. Endringene på modellen sendes tilbake til serveren hvor den blir slått sammen med endringsinformasjon om modeller fra andre enheter. Det blir så trukket ut et gjennomsnitt av endringsinformasjonen som benyttes til å forbedre hovedmodellen. Den nye forbedrede hovedmodellen kan nå lastes ned av alle klientene. Dette gir en mulighet til å forbedre en eksisterende modell basert på et stort antall brukere, men uten at brukernes data trenger å deles.

Kapselnettverk[18]
Kapselnettverk er en nyere variant av nevrale nettverk, og krever blant annet mindre data for å lære enn det som er vanlig for dyp læring i dag. Dette gir en stor fordel ved at man vil trenge langt mindre data til maskinlæringen.

2. Metoder som ivaretar personvernet uten at datagrunnlaget reduseres

Det optimale ville være hvis man kunne benytte så mye data som man ønsket til maskinlæring, uten at det gikk på bekostning av personvernet. Innen kryptologifeltet finnes det noen lovende muligheter på dette området:

Differential privacy[19]
La oss for eksempel ta utgangspunkt i en database med personer og egenskaper knyttet til disse. Ved uthenting av informasjon fra databasen så vil svaret inneholde bevist tilført støy. Det vil derfor være mulig å hente ut informasjon om personene i databasen, men ikke nøyaktig informasjon om enkeltpersoner. En database skal ikke kunne gi et merkbart forskjellig resultat på en spørring om en enkelt person blir fjernet fra databasen, eller ikke. De overordnede trendene eller trekkene ved datasettet vil ikke endre seg.

Homomorfisk kryptering
Homomorfisk kryptering er en krypteringsmetode som gjør det mulig å behandle data mens de fortsatt er krypterte. Dette gjør at konfidensialiteten kan ivaretas uten å begrense muligheten til å bruke datagrunnlaget. Per i dag har homomorfisk kryptering begrensninger som gjør at løsninger som benytter dette vil få en stor ytelsesreduksjon, men teknologien er lovende.

Microsoft har for eksempel publisert et white paper på en løsning som benyttet homomorfisk kryptering i

[16] https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
[17] https://research.googleblog.com/2017/04/federated-learning-collaborative.html
[18] https://openreview.net/pdf?id=HJWLfGWRb
[19] https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf, https://arxiv.org/abs/1412.7584

forbindelse med bildegjenkjenning.[20] Det foregår også et aktivt arbeid for å standardisere løsninger for homomorfisk kryptering.[21]

Transfer learning[22]
Det er ikke slik at det alltid er nødvendig å utvikle modellene fra bunnen av. En annen mulighet kan være å basere seg på eksisterende modeller som løser lignende oppgaver. Ved å bruke disse som et utgangspunkt, kan man ofte oppnå samme resultat med færre data og kortere prosesseringstid. Det finnes biblioteker med ferdigtrente modeller som man kan benytte seg av.

RAIRD
Statistisk sentralbyrå (SSB) og Norsk senter for forskningsdata (NSD) har utviklet en løsning med betegnelsen RAIRD[23] som tillater at man kan forske på deres data uten å ha direkte tilgang til det fullstendige datagrunnlaget.

I korte trekk fungerer løsningen ved at forskerne har et grensesnitt som kun gir tilgang til metadata til det underliggende datagrunnlaget. Datagrunnlaget kan for eksempel være et kreftdiagnoseregister som inneholder felter for alder, kjønn, fødselsdato og fødested. Forskeren kan så gjøre spørringer basert på metadataene og få ut en rapport som kun inneholder aggregerte data.

Løsningen er lagt opp for å forhindre at man kan hente ut data om veldig små grupper og enkeltpersoner. Denne typen løsning kan dermed benyttes også når man trenger data til maskinlæring. Istedenfor at man får en rapport som sluttresultat, kunne man fått en modell ut av systemet.

3. Metoder for å unngå svart boks-problematikken

En av problemstillingene som har blitt nevnt er manglende gjennomsiktighet i forbindelse med maskinlæring og automatiserte avgjørelser. Dette er en utfordring for både de som benytter et slikt system og menneskene som blir behandlet i det. Utviklere av løsninger som baserer seg på maskinlæring kunne ha en stor fordel av å vite hva som skjer under panseret for å kvalitetssikre og forbedre utviklingen.

Explainable AI (XAI)[24]
XAI er en tanke om at alle automatiserte avgjørelser som blir tatt skal være mulig å forklare. Når det er mennesker med i en prosess, vil det som oftest være ønskelig at det følger med en forklaring på utfallet. Her vil det være interessante muligheter. Hvordan kan man bygge nye løsninger som i tillegg til å være nøyaktige også gir gode forklaringer? Det vil også være et behov for å kunne ettergå løsninger som ikke har dette innebygd. For utviklere som benytter seg av «Transfer learning» vil det sannsynligvis også være attraktivt.

På dette området er det blant annet et prosjekt i regi av Defense Advanced Research Projects Agency (DARPA) hvor de ønsker mer forskning på forståelige forklaringer på automatiserte avgjørelser. Blant annet har de sponset Oregon State University med 6,5 millioner dollar over fire år for å forske på temaet. Målet er å kunne lage kunstig intelligens som kan forklare avgjørelsene sine slik at man kan forstå og ha tillitt til systemet. Uansett er det grunn til å tro at denne forskningen vil hjelpe hele feltet fremover.

LIME[25]
En tilnærming til XAI er LIME. Lime er en modell-agnostisk løsning som lager forklaringer som vanlige mennesker kan forstå. Hvis man for eksempel har bildegjenkjenning, vil den kunne vise hvilke deler av bildet som er relevant for hva den tror bildet er. Dette gjør det enkelt for hvem som helst å forstå grunnlaget for en avgjørelse.

---

[20] https://www.microsoft.com/en-us/research/publication/cryptonets-applying-neural-networks-to-encrypted-data-with-high-throughput-and-accuracy/
[21] http://homomorphicencryption.org/
[22] http://www.cs.utexas.edu/~ml/publications/area/125/transfer_learning
[23] http://raird.no/
[24] https://www.darpa.mil/program/explainable-artificial-intelligence
[25] https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

Anbefalinger for godt personvern i utvikling og bruk av kunstig intelligens

Som en oppsummering følger her et knippe anbefalinger vi har laget for personvernvennlig utvikling og bruk av kunstig intelligens.

Anbefalinger til aktører som driver med forskning og utvikling av kunstig intelligens

Disse anbefalingene er rettet mot aktører som driver med grunnutvikling av kunstig intelligens. Det vil i hovedsak være forskningsmiljøer på universiteter og store kommersielle virksomheter. Dette er en viktig målgruppe fordi det er aktørene som utvikler grunnteknologien som er forutsetningen for videre bruk av kunstig intelligens.

- Forsk på hvordan man kan gjøre teknologien mer personvernvennlig, slik som hvordan man kan utvikle teknologi som gjør det enkelt for brukere av teknologien å etterleve regelverket. Det kan for eksempel forskes på løsninger som bruker mindre treningsdata, anonymiseringsteknikker og løsninger som forklarer hvordan systemene behandler data og hvordan de konkluderer. Andre interessante forskningsområder er hvordan legge til rette for revisjon av systemene, gjerne i regi av tredjeparter.
- Tenk tverrfaglig. Kunstig intelligens er mer enn bare teknologi. Det er viktig å sette sammen tverrfaglige team som kan vurdere de etiske og samfunnsmessige konsekvensene av systemene som utvikles. Forskning kan også belyse hvor bruk av kunstig intelligens kan ha stor samfunns-messig verdi og på hvilke områder det kan være mer problematisk.

Anbefalinger til løsningsleverandører

Disse anbefalingene er ment for virksomheter som benytter seg av grunnteknologi utviklet av andre – virksomheter som benytter kunstig intelligens i egne prosjekter eller i løsninger som tilbys andre. De kan være databehandlere eller kun selge selve systemet. Anbefalingene er også relevante for forskningsmiljøer som bruker grunnteknologi utviklet av andre.

- Sett dere inn i personvernregelverket – både hvilke plikter dere har, og hvilke rettigheter og plikter bestilleren eller brukeren av løsningen har.
- Velg modeller som er tilpasset personvernbehovene til bestilleren, for eksempel kan ikke alle typer modeller forklare hvordan den kom fram til et spesifikt resultat.
- Begrens mengden personopplysninger i treningsdata til det som er relevant og nødvendig for formålet.
- Sørg for, og dokumenter at, løsningen dere utvikler oppfyller kravene om innebygd personvern.
- Dokumenter hvordan kravene i personvern-regelverket oppfylles. Det er et krav i regelverket, og kunder eller brukere av teknologien vil spørre etter dette.
- Gi kundene veiledning i hvordan ulike løsninger ivaretar personvernet, for eksempel om løsningen kan bidra til å oppfylle informasjonsplikten og hvordan kunden kan teste eller revidere løsningen for å sikre at den etterlever regelverket og interne krav.

Anbefalinger til virksomheter som kjøper og bruker systemer basert på KI

Disse anbefalingene er rettet mot virksomheter som kjøper inn og bruker IT-løsninger som er basert på kunstig intelligens-teknologi. Denne målgruppen må stille krav.

- Gjør en risikovurdering, og ved behov utred personvernkonsekvensene før du kjøper et system, før du setter det i bruk og etter at systemet er tatt i bruk.
- Still krav om at løsningen dere bestiller oppfyller kravene til innebygd personvern.
- Gjør regelmessige tester av systemet for sikre at løsningen etterlever kravene i regelverket, for eksempel for å unngå skjult forskjellsbehandling.
- Sørg for at løsningen ivaretar rettighetene til brukerne dine, for eksempel retten til å begrense behandlingen.
- Sørg for å ha gode systemer for å ivareta de registrertes rettigheter, slik som retten til informasjon, innsyn og sletting. Systemet må også legge til rette for samtykkehåndtering, inkludert tilbaketrekking av samtykke.
- Vurder å etablere bransjenormer, etiske retningslinjer eller et «personvernpanel» bestående av eksterne eksperter på teknologi, samfunn og personvern. Disse kan gi innspill på juridiske, etiske og teknologiske utfordringer – og muligheter – forbundet med bruk av kunstig intelligens.

Anbefalinger til sluttbrukere

En sluttbruker vil si den registrerte som bruker en tjeneste eller som er gjenstand for behandling av personopplysninger ved bruk av kunstig intelligens. Her er en oppsummering av dine rettigheter:

- **Rett til informasjon.** Du har rett på forståelig og lett tilgjengelig informasjon om at personopplysninger om deg blir behandlet. Denne retten gjelder både når virksomheter innhenter opplysninger fra deg direkte, og når de hentes fra andre kilder. Du skal da blant annet få vite hva opplysningene skal brukes til (*formål*) og hvilket *behandlingsgrunnlag* virksomheten baserer behandlingen på. Det kan for eksempel være hjemmel i lov, en avtale eller ditt uttrykkelige samtykke.
- **Samtykke**. I mange sammenhenger må den som skal behandle personopplysninger om deg få ditt samtykke før behandlingen kan settes i gang. Den som behandler opplysningene er selv ansvarlig for å dokumentere at et gyldig samtykke er blitt gitt, det vil si at du har gitt en frivillig, spesifikk, informert og utvetydig erklæring om at du godtar at dine personopplysninger blir behandlet. Du har også rett til å trekke samtykker du har gitt tidligere.
- **Rett til innsyn.** Du har rett til å kontakte virksomhetene og be om å få vite *om* de behandler opplysninger om deg, og i så fall *hva* som da er registrert om deg. Du har som regel rett til å få en kopi av de registrerte opplysningene. Det er imidlertid noen unntak fra innsynsretten, for eksempel innen justissektoren.
- **Retten til å korrigere og slette opplysninger.** Du har rett til å kreve at uriktige eller unødvendige opplysninger om deg blir korrigert eller slettet.
- **Retten til å komme med innsigelse** mot at opplysninger om deg blir behandlet. Du kan ha rett til å protestere mot at personopplysninger om deg blir behandlet. Dersom du protesterer mot direkte markedsføring, skal den stoppes uten at du trenger å gi noen nærmere begrunnelse. I andre sammenhenger kan det være at du må begrunne innsigelsesretten din med forhold som er knyttet til din situasjon. Da må virksomheten stoppe behandlingen, med mindre de kan påvise at de har tvingende berettigede grunner til å behandle opplysningene, og at disse grunnene veier tyngre enn dine interesser, rettigheter og friheter.
- **Retten til å kreve begrenset behandling**. Dersom du mener at noen opplysninger er feil, behandles ulovlig, eller du har benyttet retten til å protestere mot behandlingen, kan virksomheten bli nødt til å stoppe bruken av opplysningene, men fortsatt lagre disse fram til uenigheten er avklart.
- **Dataportabilitet**. Dersom du etter avtale eller samtykke har gitt fra deg opplysninger om deg selv, kan du kreve disse opplysningene utlevert fra virksomheten i en strukturert, alminnelig anvendt og maskinlesbar form.

Anbefalinger til myndigheter

Disse anbefalingene er til lovgivere og politiske beslutningstakere da de som er premissgivere for utvikling og bruk av kunstig intelligens.

- Sørg for at offentlig sektor går foran som et godt eksempel ved bruk av kunstig intelligens. Det krever at offentlig sektor har høy bevissthet rundt etikk og personvernkonsekvensene ved løsningene de bruker, samt at de utvikler en bestillerkompetanse som gjør at løsningene har innebygd personvern og oppfyller lovkravene.
- Bevilg midler til forskning som sikrer at teknologien håndterer personopplysningene i samsvar med personvernregelverket. Dette er ikke bare et lovkrav, men kan også være en konkurransefordel for norsk næringsliv.
- Sørg for at tilsynsmyndigheter har relevant kompetanse, og legg til rette for erfarings-utveksling og kunnskapsdeling på tvers av sektorene.
- Sørg for at lovgivningen holder følge med den teknologiske utviklingen. Dette gjelder all lovgivning som er relevant for bruk av personopplysninger.

Besøksadresse:
Tollbugata 3, 0152 Oslo

Postadresse:
Postboks 8177 Dep.,
0034 Oslo

postkasse@datatilsynet.no
Telefon: +47 22 39 69 00

datatilsynet.no
personvernbloggen.no
twitter.com/datatilsynet