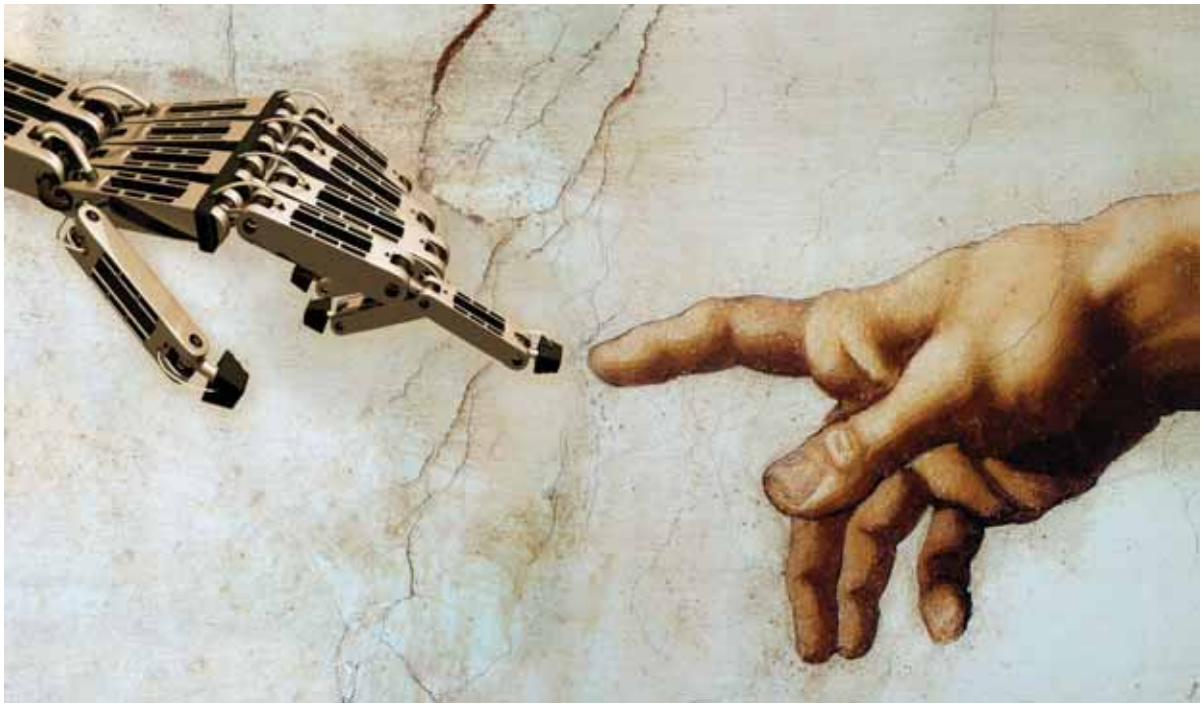


CHRISTIAN BENDIKSEN | EIRIK NORMAN HANSEN

Når juss møter AI

Rettslig regulering av kunstig intelligens

GYLDENDAL

CHRISTIAN BENDIKSEN OG EIRIK NORMAN HANSEN

NÅR JUSS MØTER AI

RETTSLIG REGULERING AV KUNSTIG INTELLIGENS

GYLDENDAL

1. utgave, 1. opplag 2019

ISBN 978-82-05-52178-0
Omslagsdesign: Kristin Berg Johnsen
Omslagsillustrasjon: © Mike Agliolo / Getty Images
Layout: Bøk Oslo AS
Sats: Bøk Oslo AS
Brødtekst: Minion 10,5/15

Alle henvendelser om boken kan rettes til
Gyldendal Juridisk
Postboks 6730 St. Olavs plass
0130 Oslo

www.gyldendal.no/juridisk
juridisk@gyldendal.no

Alle Gyldendals bøker er produsert i miljøsertifiserte trykkerier.
Se www.gyldendal.no/miljo

Forord

Dette er en bok om juss, men det er ikke en bok som primært er skrevet for jurister. Ettersom kunstig intelligens tas i bruk på stadig flere områder, vil flere og flere måtte forholde seg til juridiske konsekvenser av at AI tar beslutninger. Dette er en bok for dem. Alt fra bedriftsledere som vurderer å anskaffe AI til å løse konkrete problemer, til kunder og leverandører som må forholde seg til kontrakter inngått ved AI. Derfor er språket (forhåpentligvis) enklere og eksemplene annerledes enn i en juridisk lærebok.

Svært mange mennesker har hjulpet oss med å få dette til å bli en bok. Vi sendte forespørsler om intervjuer til over et dusin IT-selskaper, og de aller fleste tok seg tid til å snakke med oss om praktisk implementering av AI, databehov og dataanalyse. Vi retter en stor takk til samtlige.

Videre har flere fagpersoner tatt seg tid til å lese manus og gi oss innspill og korreksjoner. Morten Goodwin ved Universitetet i Agder tok seg tid til å gi oss mye verdifull informasjon om hvordan man kan formidle maskinlæring på en måte som – forhåpentligvis – er forståelig for vanlige mennesker, og hvordan data brukes til trening. Mye av innholdet i de to første kapitlene er basert på hans innspill, men misforståelsene er selvsagt bare våre. Advokat Alexander Mollan ga oss kommentarer til kapittelet om personvern, og advokat Torbjørn Evjent ga verdifulle innspill til kapittelet om strafferett. Og vi kommer heller ikke unna en stor takk til familiene våre, som ikke har sett så mye til far som de har ønsket i løpet av det året som har gått fra tentativ disposisjon til ferdig manus.

Første disposisjon skrev vi for ganske nøyaktig et år siden, i mars 2018, og denne boken er oppdatert per 15. mars 2019. Men mye kommer til å skje fremover. Avi-

sene rapporterer daglig om nye AI-prosjekter på helt nye områder. Forskningen står heller ikke stille. Det juridiske fakultet ved Universitetet i Oslo har varslet at de vil starte større forskningsprosjekter både på AI og på store datastrømmer, og EU har varslet at deres gjennomgang av produktansvar og AI vil være ferdig i løpet av 2019. I den grad disse initiativene er offentlige før bokens trykketidspunkt, har vi forsøkt å innarbeide dem etter beste evne.

Innhold

KAPITTEL 1

Om kunstig intelligens

«Så fort det virker, kaller ingen det AI lenger …»[1]

1.1 Hva er kunstig intelligens?

1.1.1 Innledning

I samtlige offentlige rapporter, og i juridisk og teknisk litteratur, om kunstig intelligens er det kun én ting alle kan være enige om: Det finnes ingen generelt akseptert definisjon av «kunstig intelligens».[2] Det britiske House of Lords kommenterte tørt at det ikke er særlig overraskende, all den tid det heller ikke finnes noen definisjon av organisk (eller menneskelig) intelligens, som jo gjerne brukes som sammenligningsgrunnlag.[3]

Definisjonene spenner fra det altomfattende til det spesifikke og fra en sterk betoning av menneskelig intelligens som sammenligningsgrunnlag til en understrekning av at kunstig intelligens vil skape tankeprosesser og -resultater som mennesker aldri vil kunne replisere, langt mindre forstå.

På den ene siden finner vi utgangspunktet for rapporten levert av den franske matematikeren Cédric Villiani til president Emmanuel Macron, hvor det fremheves at «kunstig intelligens definerer i praksis i mindre grad et klart definert sett av forskningsoppgaver enn et program basert på en ekstremt ambisiøs målsetting: å forstå hvorledes menneskelig tankevirksomhet fungerer, og å reprodusere dette – å skape kognitive prosesser som er sammenlignbare med de menneskelige».[4]

Det britiske House of Lords baserer seg på en definisjon som også legger vekt på forholdet til menneskelig intelligens: «teknologier med evne til å gjennomføre

oppgaver som ellers ville kreve menneskelig intelligens, så som tolkning av synsinntrykk, talegjenkjenning og oversettelse av språk … og som i dag vanligvis har evne til å lære av eller tilpasse seg nye erfaringer eller stimuli».[5]

En tilsvarende formulering har allerede dukket opp i lovgivning. Nevadas lov om autonome kjøretøy mv. har følgende definisjon av kunstig intelligens i avsnittet om autonome biler:

> «Kunstig intelligens» betyr bruken av datamaskiner og tilknyttet utstyr til å gjøre en maskin i stand til å duplisere eller imitere etter menneskelig adferd.[6]

Samtidig har en ekspertgruppe nedsatt av EU[7] nylig kommet med en definisjon som i større grad fokuserer på det funksjonelle i systemet og i mindre grad på det rent spesifikt menneskelige i prosessene:

> Kunstig intelligens (AI) refererer til systemer designet av mennesker som, gitt et komplekst mål, handler i den fysiske eller digitale verdenen gjennom å oppfatte sine omgivelser, tolke innsamlede strukturerte eller ustrukturerte data, resonnere i henhold til kunnskap utledet fra disse dataene og beslutte de(n) beste handlingen(e) (i henhold til predefinerte parametre) for å nå det gitte målet.[8]

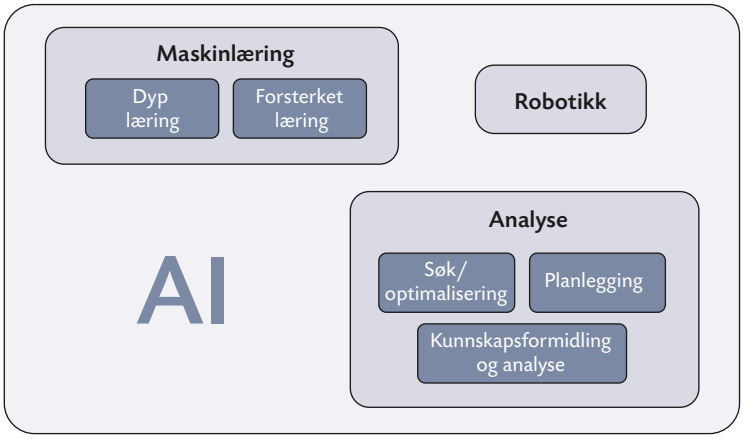
Dermed ser det ut som at EU i mindre grad enn de øvrige kildene vektlegger systemenes evne til å etterligne menneskelige tankeprosesser. Dette kan også muligens la seg utlede av den finske studien om AI, hvor man også har understreket «systemets evne til å operere på en målrettet måte og forutse sine omgivelser»,[9] men uten å koble dette mot menneskelig adferd. Det er dermed mulig at man kan spore en vridning fra et fokus på at disse systemene etterligner mennesket, til et fokus på hvilke egenskaper systemene må kunne mestre for å nå dette målet.

Etter vår oppfatning har House of Lords uansett et godt poeng når de understreker at selv om ulike former for AI ikke nødvendigvis handler om å kopiere menneskelig adferd, og selv om de systemene som kan grupperes under «AI», kanskje mer presist lar seg beskrive under andre definisjoner, så er «AI» eller «kunstig intelligens» et nyttig begrep for et bredt spekter av ulike systemer. Disse påvirker samfunnet på ulike vis, og deres innvirkning gjør forholdet til menneskelig adferd om ikke sentralt, så i hvert fall til et meget vesentlig poeng.[10]

Men hva består egentlig «AI» av? De seneste arbeidene på området[11] legger følgende hovedbestanddeler til grunn:

- *Natural language processing*: Systemet må kunne kommunisere på et naturlig språk, som engelsk. Eventuelt, som EU legger opp til, må systemet kunne agere i den fysiske verdenen gjennom en kobling til robotikk eller lignende.
- Bearbeidelse av kunnskap: Systemet må kunne erverve kunnskap, forstå at denne datamengden *er* kunnskap, og lagre den et sted.
- Automatisert analyse/refleksjon (*reasoning*): Systemet må være i stand til å analysere ervervet/lagret kunnskap.
- Maskinlæring: Systemet må kunne lære fra sine omgivelser.

Eller som AI HLEG fremstiller det i sitt kapittel om AI som vitenskapelig disiplin, hvor man vektlegger forholdet mellom persepsjon, analyse og læring på den ene siden beslutningstagning og handling på den andre siden:



Figur 1.1

Tanken er her at robotikk uttrykker egenskapene som gjør AI i stand til å handle i den fysiske verdenen. Vi er vel ikke umiddelbart enige i at robotikk er en nødvendig del av en autonom AI, og hoveddelen av denne boken dreier seg om ulike former for AI som ikke kontrollerer noe fysisk objekt, men som likevel skaper ulike former for juridiske utfordringer. Samtidig er oppdelingen av temaene maskinlæring og resonnering i underkategorier et nyttig utgangspunkt for den videre analysen av hva AI egentlig er og gjør.

1.1.2 Ulike former for kunstig intelligente systemer

I denne boken vil vi legge den vide beskrivelsen av AI til grunn: at disse systemene inkluderer rasjonelle og/eller lærende systemer som ikke nødvendigvis repliserer menneskelig adferd, samt systemer som repliserer menneskelig adferd uten nødvendigvis å være rasjonelle eller lærende.

Dette medfører at noen ganske primitive IT-systemer vil falle innenfor kategorien «AI» fordi de tilsynelatende opptrer på en menneskelig måte. Disse skal vi beskrive i punktet nedenfor.

Samtidig er det slik at utvikling, leveranse og drift av ulike former for ekspertsystemer i liten grad skiller seg fra normale IT-kontrakter. Det er de selvlærende og resonnerende systemene som skaper helt nye problemstillinger. Derfor er det disse systemene som danner basisen for den overveiende delen av analysene i denne boken.

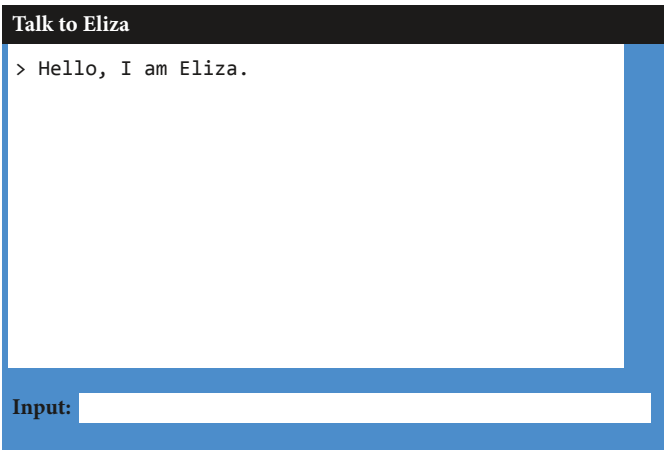
Men for å forstå de rettslige problemene slike systemer skaper, bør man ha en viss oversikt over hvordan disse smarte systemene er bygget opp, og hvordan de fungerer.

1.1.2.1 Nærmere om systemer som ikke er intelligente, men som virker som om de er det

Det finnes mange ulike former for IT-systemer som etterligner menneskelig adferd på ulike vis, eller som gir resultater som tilsynelatende er gjennomtenkte. Men det er ikke gitt at maskinlæring, forsterkende læring eller nevrale nettverk faktisk er nødvendig for å skape slike resultater. En undersøkelse fra mars 2019 av 2830 europeiske firmaer som ble markedsført som «AI Startups», viste at 41 % av dem faktisk ikke brukte AI i betydningen maskinlæring overhodet.[12] Dette behøver ikke å bety at disse bedriftene svindler hverken kunder eller investorer. Det kan bare tenkes at maskinlæring ennå ikke er nødvendig for å skape de resultatene kundene trenger. I så fall kan det tenkes at andre former for systemer er like nyttige og leverer sikrere – men også langt mindre fleksible – resultater.

Noen av disse mulige systemene karakteriseres som ekspertsystemer. Ifølge House of Lords kan et ekspertsystem defineres som «et dataprogram som imiterer beslutningsprosessene til en menneskelig ekspert gjennom å følge preprogrammerte regler som 'hvis det skjer, gjør dette'».[13] Disse systemene er bygget opp ved hjelp av et stort antall logiske regler, noe som også betyr at systemene faller sammen dersom antallet mulige svar overstiger antallet regler som kan behandle dem. Men dersom man greier å kanalisere spørsmålene inn i de formene som systemet kan besvare, vil de også gi presise svar.

Et tidlig eksempel på slike systemer er programmet Eliza, som ble utviklet mellom 1964 og 1966 på MIT og er et av de første *natural language processing*-systemene i bruk. Forbausende mange mennesker trodde at systemet hadde menneskelige følelser, til tross for at det ikke hadde noen lærings- eller resonneringsfunksjoner.



Figur 1.2

Helt overfladisk kan Eliza beskrives på følgende måte:[14]

1. Programmet skanner brukerens inntastede setning på jakt etter forhåndsdefinerte nøkkelord. Nøkkelordet med høyest definert prioritet vil generere et svar:
    1. «can you» (kan du) → svar 1–3 anvendes
    2. «can I» (kan jeg) → svar 4–5 anvendes
    3. «you are» (du er) → svar 6–9 anvendes
    4. …
    5. (ikke funnet) → svar 106–112 anvendes
2. Noen av svarene genereres tilfeldig. Eksempelvis kan svarene til «can you» (kan du) være:
    1. «Don't you believe that I can<*» (Tror du ikke jeg kan …?)
    2. «Perhaps you would like to be able to<*» (Du vil kanskje gjerne være i stand til å …?)
    3. «You want me to be able to<*» (Du vil gjerne at jeg kan …?)

3. Setningen fullføres gjennom at den delen av brukerens setning som kommer etter nøkkelordet, kopieres og settes inn i svaret. I eksemplene ovenfor betyr det at «<*» vil bli erstattet av kopien av brukerens setning.
4. Så avsluttes svaret gjennom at første person (jeg, meg) erstattes av andre person (du, deg) og omvendt.

På denne måten kan man oppnå en svært troverdig samtale, uten at noe menneske er involvert, og uten noen læring eller resonnering. Men poenget er at systemet heller ikke gir noe svar eller noen informasjon.

På den annen side er det ikke gitt at dagens chatboter trenger å være veldig mye smartere enn dette.

En god del systemer innen kundeservice som er innført de siste tjue årene, er faktisk bygget på en rekke med forhåndsdefinerte spørsmål som henger logisk sammen med oppfølgingsspørsmål i lange rekker. Rekkene er forhåndsdefinerte, og det er ingen resonnering i systemet som avgjør hvilke oppfølgingsspørsmål som trigges av innholdet i kundens svar.

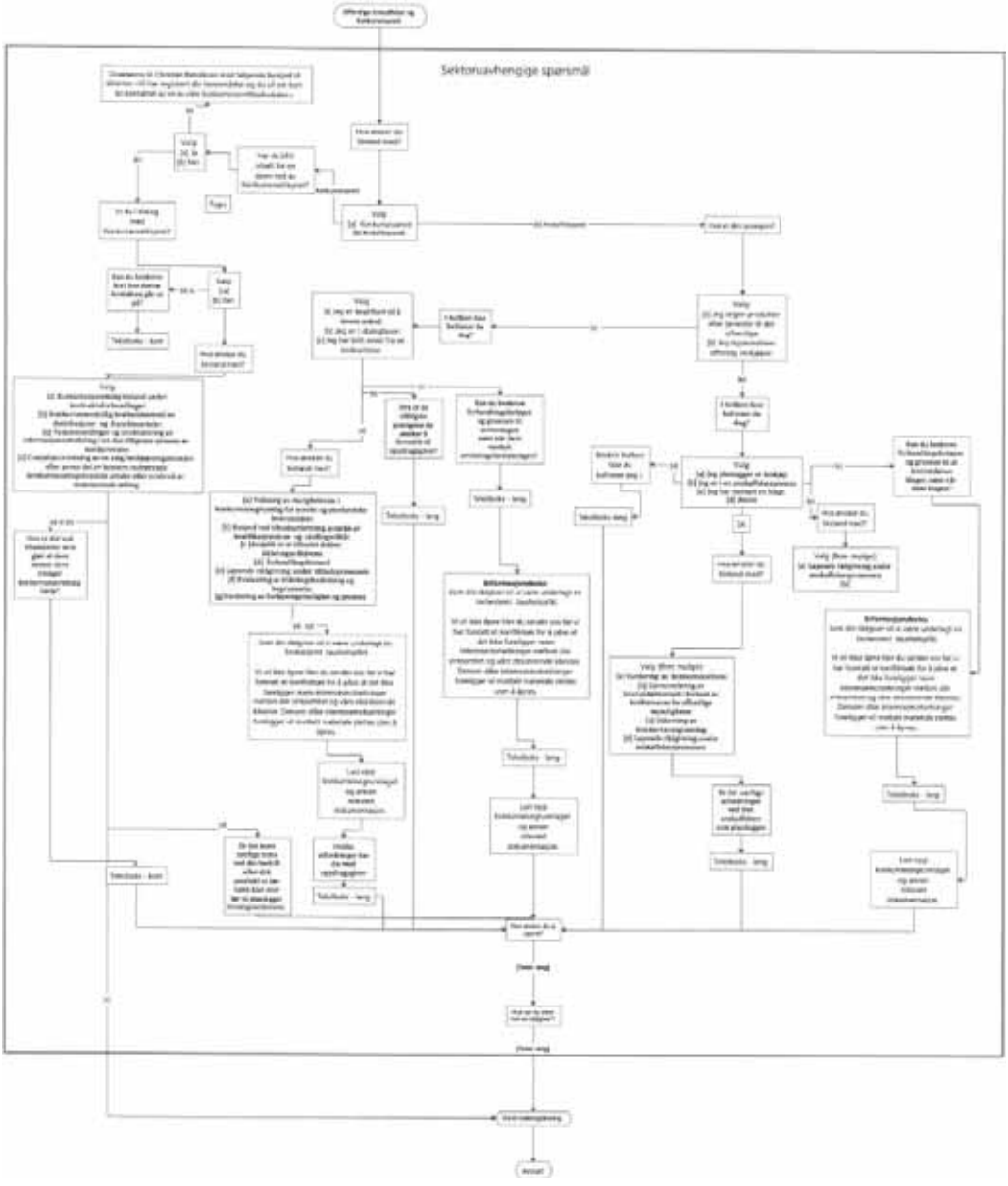
Likevel skal kunden oppleve situasjonen som om han eller hun interagerer med programmet. Dette gjøres gjennom svært komplekse rekker av spørsmål, hvor flere spørsmål kan brukes i flere rekker, og hvor sammensetningen av spørsmål er erfaringsbasert. På den måten oppleves spørsmålene som relevante for kunden.

Figur 1.3 viser et eksempel på kompleksiteten i spørsmålsflytene i et kundebehandlingsverktøy basert på en ekspertmodell.

Disse systemene vil etter hvert kunne suppleres av lærende og resonnerende AI.

En slik AI vil kunne undersøke hvilke spørsmål som gir mest omfattende svar, hvilke spørsmålsrekker som oftest får kundene til å avbryte før slutten, og hvor man samler inn de nyttigste dataene. Deretter vil AI-en kunne foreslå oppgraderinger av systemet. En slik AI gir da beslutningsstøtte til vedlikeholdet av ekspertsystemet, men går ikke selv inn i kundebehandlingen.

Eventuelt kan AI-en også erstatte deler av ja/nei-spørsmålene eller «velg fra liste»-spørsmålene som brukes for å få kunden til å definere hvilke rekker av spørsmål som skal brukes for å løse kundens problem. Eksempelvis kan en lærende AI stille mer åpne spørsmål – «Hva kan jeg hjelpe deg med?» – og lete etter triggerord i svarsetningen. Dersom kunden svarer «Jeg vil gjerne vite mer om boliglån», sender AI-en kunden videre på spørsmålsrekken som dreier seg om boliglån, i stedet for at kunden selv må velge fra en nedtrekksmeny eller en annen form for forhåndsdefinert liste.



Figur 1.3 Eksempel på kompleksiteten i et kundebehandlingsverktøy basert på ekspertmodell

1.1.2.2 Ulike former for selvlærende og resonnerende systemer

I det følgende skal vi gi en oversikt over ulike kategorier av selvlærende og resonnerende systemer. Hvordan disse systemene faktisk lærer og resonnerer, skal vi komme nærmere inn på i neste punkt.

Den første definisjonen man bør ha et forhold til, er forskjellen på smal eller svak AI og generell eller sterk AI: Et generelt AI-system kan utføre de fleste handlinger som et menneske kan gjøre. Smale AI-systemer kan kun utføre spesifikke oppgaver, men disse er det også mulig at systemet kan løse bedre enn et menneske. Et generelt AI-system vil ha evnen til å resonnere på et bredt felt av områder og vil i praksis være umulig å skille fra et menneske, selv om det kanskje også er bedre enn mennesker til generell problemløsning. Samtidig er det ingen slike generelle AI-systemer i drift i dag, og det er betraktelige utfordringer som må løses før man kan sette et slikt system i drift.[15]

De fleste selvlærende og resonnerende systemer baserer seg på en form for maskinlæring ved algoritmer. En algoritme er et presist beskrevet dataprogram bygget for å løse et helt konkret problem. Noen av algoritmene er statistikkbaserte, andre lager beslutningstrær, mens de mest dominerende algoritmene i dag er bygget på prinsipper som går under navnet nevrale nettverk.

Dette er nettverk av virtuelle synapser og nevroner som har en overordnet likhet med vår egen hjerne. Nevronene organiseres i lag, hvor hvert lag har en bestemt oppgave. Når det er et tilstrekkelig antall lag, kalles algoritmen ofte et dypt nevralt nettverk (*deep neural network*) eller dyp læring (*deep learning*). Ordene «synapser» og «nevroner» brukes som pedagogiske grep; i realiteten representerer hver synapse et tall, som vi kaller en vekt. Hvilket tall nevronene skal ha, avhenger av hvilken oppgave som skal løses. Vi bestemmer tallverdien gjennom å trene algoritmene. Med andre ord: Når noen trener et nevralt nettverk, er det for å bestemme hvor sterke vektene skal være, altså hvilke tall som skal ligge i hver node/nevron når den mottar data fra laget over.

De nevrale nettverkene mestrer mange oppgaver ingen andre algoritmer får til i dag, og er derfor sett på som spydspissalgoritmene innen kunstig intelligens.

Dernest bør man gå ned i detaljene. Som figur 1.1 fra AI HLEG viser, har kunstig intelligens flere underkategorier både av resonnering og av læring, og disse kan kombineres på flere ulike måter.

Resonneringsmodellene må være i stand til å konvertere data – enten i et treningssett eller i en datastrøm – til mening som systemet kan bearbeide og modellere.

Når dataene er tilgjengelige i en form som systemet kan bearbeide, må systemet kunne analysere tilgjengelige løsninger på problemet og ha modeller for å velge blant disse for å finne den optimale løsningen på det konkrete problemet dataene viser. Deretter må systemet kunne velge hvilken handling det skal ta, basert på skrittene foran.

Dernest bør man skille mellom rasjonelle AI-systemer og lærende rasjonelle AI-systemer. Forskjellen ligger i evnen til å evaluere virkningen av egen adferd opp mot systemets mål. Dersom adferden forrige gang ikke er best egnet for å oppnå det ønskede målet, vil et lærende rasjonelt system modifisere sine beslutningsmetoder og resonneringsregler. Et rent rasjonelt system vil derimot gjenta den samme adferden inntil regelsettet dets blir endret av et menneske (eller et annet system).

En variant av dette skillet ligger i hvordan systemet trenes: veiledet eller ikke-veiledet læring, eller en mellomting.[16]

I veiledet trening er algoritmene opplært på et sett med data som allerede har «riktige» verdier allokert til dem. Eksempelvis at dataene er strukturert i databaser med «navn», «adresse» og «telefonnummer.» AI-en vil da lære å gjenkjenne ulike verdier som «adresser» basert på fasiten i databasene. I eksempelet ovenfor lærer den å gjenkjenne en enkel telefonkatalog.

I praksis kan det kreve stor innsats å generere slike «merkede» datasett. Dette er grunnen til at mange hjemmesider nå krever at du skal «bevise at du er et menneske» gjennom for eksempel å gjenkjenne biler på 15 bilder. I realiteten gir innehaveren av siden blaffen i om du er et menneske, men hver gang får de generert et datasett som kan brukes til maskinlæring. Og det datasettet har verdi.

I ikke-veiledet trening er algoritmene overlatt til selv å finne sammenhenger og sekvenser i datasettet uten noen retningslinjer for hva de skal lete etter.

En mellomting mellom disse to læringsformene er «forsterket læring». I styrt læring brukes avviket fra fasitsvaret til å korrigere algoritmen, mens i forsterket læring blir tilbakemeldingen kun basert på hvor godt løsningen fungerte. Tilbakemeldingen gir med andre ord ingen tilbakemelding på hvor feilen var, eller hvordan den skal korrigeres.[17]

Dette kommer vi nærmere tilbake til i neste punkt, men det er verd å merke seg at i begge situasjonene er det uansett algoritmenes evne til å endre produksjonsresultat – output – basert på erfaringene som opparbeides under treningen, som gjør maskinlæringen til noe helt annet enn andre datasystemer.

1.2 Hvordan lærer kunstig intelligens?

1.2.1 Innledning

All lærende kunstig intelligens lærer fra et miljø.

I ikke-veiledet læring skal algoritmene oppdage mønstre i dataene i miljøet, men uten at noen veileder algoritmene. Eksempelvis kan algoritmene gruppere kunder i en nettbutikk basert på kjøpshistorikk. Her vil kundene i én gruppe ha et helt annet mønster enn kundene i en annen gruppe. Algoritmene vil gruppere kunder basert på kjøpshistorikk og i etterkant foreslå varer andre kunder i de samme gruppene har handlet.

Hvis algoritmen trenes opp med forsterkende læring, vil algoritmen hele tiden påvirke miljøet, som igjen påvirker algoritmen. Et typisk eksempel er brettspill som sjakk. En algoritme som spiller sjakk, vil ved hvert trekk endre på spillet. Spillet blir ulikt avhengig av om algoritmen flytter kongen eller bonden. Hvilke brikker som blir flyttet, påvirker hvilke brikker algoritmen kan flytte ved neste trekk.

En AI bygget med veiledet læring lærer fra et miljø som i de aller fleste tilfeller er bygget med eksempeldata. Algoritmene skal da forsøke å gjenskape egenskapene i dataene. Typiske eksempler er å kjenne igjen språk eller innhold i en tekst, oppdage kategorier i et bilde eller diagnostisere i medisinske data.

Algoritmen går i to faser. I den første fasen, treningsfasen, mates algoritmene med data, såkalte treningsdata. I den andre fasen, testfasen, testes algoritmene med data den aldri har sett før.

La oss si at vi ønsker å trene en AI til å skille mellom to språk, engelsk og norsk. Da må den først trenes med mye data, i dette tilfellet språkeksempler fra disse to språkene. En utvikler har kategorisert den norske teksten i en norsk virtuell bunke, og den engelske teksten i en engelsk virtuell bunke. Det er algoritmens oppgave å trene på disse dataene.

Etter at treningen er ferdig, havner vi i testfasen. Her vil algoritmen få et helt nytt eksempel som den ikke vet om er norsk eller engelsk. Hvis den kategoriserer teksten som engelsk, og det viser seg at den er engelsk, har den gjort det riktig.

Hver tekst som gis algoritmen, har nå en fasit, enten norsk eller engelsk, og i treningsfasen er det algoritmens oppgave å skille mellom disse.

I språkeksempelet vil mønstrene den plukker opp, antageligvis være ganske enkle. Typiske mønstre for norsk vil være ord som bare finnes på norsk, og ikke fin-