

Når bør vi bruke datamaskiner og når bør vi bruke hodet? En kritikk av Dreyfus og Dreyfus' (1987) myter om ekspertise

Geir Kirkebøen
Universitetet i Oslo

Innledning

Allerede på 50-tallet foreslo enkelte at datamaskinen burde brukes til å utvikle "kunstig intelligens" (KI). Denne tidlige KI-visjonen har i alt for stor grad preget diskusjonene rundt spørsmålet om hva vi kan og bør bruke datamaskiner til. Debatten i informatikkmiljøer har vært sterkt polarisert. På den ene siden har vi hatt de som har forsvart de svulstige KI-visjonene. På den andre siden, har vi hatt KI-kritikerne som svært ensidig har argumentert for et "mind over machine"-syn.

KI-visjonærene og KI-kritikerne har hatt til felles at de har basert sine oppfatninger om hhv. datamaskinens muligheter og menneskets fortreffeligheter, på spekulasjoner. KI-visjonen baserer seg på en ren spekulasjon, nemlig antagelsen om at "a physical symbol system has the necessary and sufficient means for general intelligent action" (Newell og Simon, 1981, s. 41). Denne hypotesen er igjen basert på en antagelse om at kognisjon er samme type prosess som informasjonsprosessering i datamaskiner. KI-kritikerne (f.eks. Dreyfus, 1972; Winograd og Flores, 1986) har - med gode argumenter - avvist denne antagelsen, men de har fulgt opp med en like ensidig og spekulativ antagelse om menneskelig ekspertise som noe høyt hevet over maskiner.

Spørsmålet om en programmert datamaskin kan sies å være intelligent eller ikke, har liten relevans for problemstillingen jeg reiser i artikkelens tittel. Jeg mener det nå er på tide at informatikere vurderer hvilken rolle datamaskiner kan og bør spille i beslutningstaking, problemløsning, utøvelse av ekspertise osv. ut fra hva man i kognitiv psykologi vet om styrker og svakheter i eksperters skjønn og beslutningstaking. Denne kunnskapen er i hovedsak avdekket i en tradisjon i psykologien som går tilbake til Paul Meehls (1954) bok *Clinical versus Statistical Predictions*. I hva vi kan kalle Meehl-tradisjonen har man systematisk sammenlignet menneskers evne til å foreta skjønnsmessige vurderinger med mekaniske rutiner. Paradokset er at mens KI-tradisjonens forsøk på å utvikle avanserte KI-programmer har avdekket menneskers styrke i forhold til datamaskiner og altså gitt næring til en generell "mind-over-machine"-myte, så har den systematiske sammenligningen av mennesker mot helt enkle mekaniske rutiner avdekket dramatiske svakheter i menneskelig skjønn og vurderingsevne.

Jeg vil først vise at det ekspertisesyn Dreyfus og Dreyfus (1987) forfekter, blir en myte sett i lys av de siste førti års beslutnings- og vurderingsforskning. Deretter vil jeg

forsøke å få fram at den type funn som denne forskningen har avdekket, er svært relevant for utforming av menneske-maskin samspill og beslutningsstøttesystemer.

Dreyfus og Dreyfus' (1987) myter om ekspertise

I 1947 ble den første programmerbare digitale datamaskinen satt i drift. Samme år startet Allan Turing debatten om muligheten for å utvikle "kunstig intelligens" (KI). Han foregrip idéen som ligger til grunn for den moderne KI-visjonen, en idé som gjerne tilskrives Newell, Shaw og Simon (1958). Turing (1947) antok at kognisjon essensielt er (symbolsk) problemløsning og at problemløsning kan reduseres til "the form 'Finding a number n such that ...' ... We should not go far wrong if we assumed that all problems are reducible to this form." (s. 22). Turings mest markante motdebattant i 1947 var kjemiprofessoren Michael Polanyi, en venn og nabo av Turing på denne tida. Polanyi var sterkt uenig med Turing. Han hevdet at menneskelig kognisjon slett ikke lot seg spesifisere eller simulere ved hjelp av formelle systemer. Diskusjonene inspirerte hhv. Turings (1950) artikkel *Computing Machinery and Intelligence* og Polanyis (1958) bok *Personal Knowledge*. Den velkjente "Turing-testen" for å vurdere om en programmert maskin kan sies å være intelligent, er et sentralt bidrag til KI-tradisjonen i Turings artikkel. Polanyi avviste testen som et kriterium på intelligens. For han var det en *a priori* kjennsgjerning at menneskelig kognisjon og intelligens er vesensforskjellig fra programmerte datamaskiner. Ingen empiri kan endre dette, hevdet Polanyi. Den fundamentale forskjellen mellom mennesker og maskiner kommer særlig til uttrykk i hans begrep "taus" kunnskap (*tacit knowledge*). I dette begrepet ligger det at menneskelig kognisjon dels er basert på kunnskap som ikke lar seg beskrive. Begrepet "taus kunnskap" har siden stått sentralt i KI-kritikken. (For debatten mellom Turing og Polanyi, se f.eks. Hodges, 1983)

Filosofen Hubert Dreyfus (1965, 1972) overtok på 60-tallet Polanyis rolle som toneangivende KI-kritiker. Han fokuserer, som Polanyi, på hva maskiner *ikke* kan gjøre. Mens Polanyi bl.a. argumenterte ut fra Gödels ufullstendighetsteorem, er utgangspunktet for Dreyfus' kritikk de tidlige faktiske forsøkene på å utvikle KI-programmer. Dreyfus retter en skarp, og i stor grad berettiget, kritikk mot vyene mange hadde om å "kopiere" menneskelig intelligens og ekspertise i form av programmerte datamaskiner.

Verre blir det når Dreyfus og Dreyfus (1987) i boka *Mind over Machine* forklarer hva menneskelig eksperter *kan* gjøre som maskiner ikke kan. De sammenfatter sin forståelse av hva som skiller eksperter fra programmerte datamaskiner og nybegynnere slik:

"It seems that a beginner makes inference using rules and facts just like a heuristically programmed computer, but that with talent and a great deal of involved experience the beginner develops into an expert who *intuitively sees* what to do without applying rules." (Dreyfus og Dreyfus, 1987)

I dette sitatet kommer de to mytene jeg mener Dreyfus og Dreyfus forfekter, til uttrykk. Den ene er påstanden om at ekspertise *generelt* har en gestaltkarakter, at eksperter "intuitivt ser" helheter i mønstre av data/informasjon. Vi kan kalle dette *myten om det gode skjønn*. Dette er nemlig en myte. Psykologisk forskning har avdekket at menneskelige beslutningstakere, eksperter inkludert, ofte ikke er i stand til dette.

Den andre relaterte myten som kommer til uttrykk i sitatet, kan vi kalle *myten om erfaringslæring*. Med dette mener jeg troen på at man på alle kunnskapsområder har mulighet til å erverve seg et godt skjønn gjennom erfaring. Som vi skal se, har psykologisk forskning klart vist at dette langt fra er tilfelle.

Hva bygger Dreyfus og Dreyfus (1987) sin forståelse av ekspertise på?

Dreyfus og Dreyfus (1987) peker korrekt på at KI-visjonene baserer seg på spekulasjoner: "the grandiose claims and predictions made by Simon and associates were not based on sound empirical research" (s.8) Hva bygger så Dreyfus og Dreyfus (1987) sin egen forståelse av ekspertise på? I hovedsak baserer de seg på "the seemingly plausible arguments of Merleau-Ponty, Heidegger, and Wittgenstein, which I had come to accept" (Dreyfus & Dreyfus, 1987, s.7). Den empiriske forskningen på erfaringslæring og skjønnsmessige vurderinger, som klart strider mot deres eget syn, neglisjerer eller bortforklarer de: "To forsake rationality in favor of unrationalized know-how is to sail on uncharted seas, and there will always be those ... who challenge the wisdom. A number of academic psychologists have gone so far as to create experiments purporting to show ... consistent flaws in human decision-making" (s. 41)

I den grad Dreyfus og Dreyfus baserer seg på empiriske observasjoner, avgrensner de seg i hovedsak til studier og observasjoner av senso-motoriske ferdigheter som sykling, svømming og flyvning. Av kognitive ferdigheter er det stort sett studier av (og uttalelser fra) sjakkspillere de viser til. Hva de i bokas undertittel karakteriserer som "the power of human intuition" belegger de ikke med stort annet enn eksperters egne opplevelser:

"There is no choosing. It happens unconsciously, automatically, naturally" (Japansk kunstner, s. 32);

"When I say to a doctor, 'the patient is psychotic', I don't know how to legitimize the statement. But I am never wrong. Because I know psychosis from inside out. And I feel that, I know it, and I trust it". (Psykiatrisk sykepleier, s. 34);

"To these elementary laws there leads no logical path, but only intuition, supported by being sympathetically in touch with experience". (Einstein, s. 41)

Dreyfus og Dreyfus appellerer dessuten til lesernes "common sense":

"You need not merely accept our word but should check to see if the process by which you yourself acquired various skills reveals a similar pattern." (s. 20)

Jeg benekter ikke at Dreyfus og Dreyfus' ekspertisesyn er utbredt. Jeg vil imidlertid vise at mange tiårs psykologisk forskning klart viser at deres tro på både vår evne til å lære av erfaring og "intuisjonens kraft", ikke er "common sense". Den er - i mange og viktige sammenhenger - "common non-sense".

Dreyfus og Dreyfus' (1987) feilslutninger

På vesentlige områder er menneskelig kognisjon klart overlegen informasjonsprosesser i datamaskiner. Det gjelder for eksempel diverse former for persepsjon og bruk av naturlig språk. Felles for våre perseptuelle evner og våre språkevner er at de er utviklet gjennom millioner av år. Vår hjerne er spesielt "designet" for å realisere slike ferdigheter.

Når det gjelder for eksempel ansiktspersepsjon er vår "mind" høyt hevet over "the machine". Vi vet blant annet at vi alle har en ekstrem evne til å "lese" ansikter. Vi vurderer vanligvis raskt humøret en person er i ut fra personens ansiktsuttrykk. Selv om vi er svært dyktige til dette, kan vi ofte bare vagt beskrive de trekk eller tegn vi kjenner igjen. Polanyi (1958) brukte nettopp slike persepsjonseksempler i sin argumentasjon for den "tause", ikke-beskrivbare kunnskapens betydning. Dreyfus gjør det samme. I *What computer's can't do* (1972) innleder han typisk nok med et sitat der filosofen Pascal kritiserer Descartes' forståelse av nettopp persepsjon: "Mathematicians wish to treat matters of perception mathematically, and make themselves ridiculous ... the mind ... does it tacitly, naturally, and without technical rules."

Persepsjon har åpenbart en "gestalt-karakter": Ser vi deler av en sirkel, "ser" vi hele sirkelen. Dreyfus og Dreyfus' (1987) hevder at også eksperter har en slik gestaltkarakter; dvs. at eksperter i sin alminnelighet er i stand til å "se" helheter i mønstre av data/informasjon. De generaliserer altså fra persepsjon til vurderinger generelt. Dette er en feilslutning.

Det som skiller mange av de typer problemer som eksperter i vår moderne tid stilles overfor fra problemene knyttet til å "lese" ansikter og forstå naturlig språk, er at vi ikke har noen evolusjonsmessige preferanser for å løse den førstnevnte type problemer. Det gjelder for eksempel problemet en lege eller psykolog har som på grunnlag av et sett av data, skal vurdere sannsynligheten for at en pasient har en bestemt diagnose. En slik "kunstig" integrering av informasjon er et "nytt" problem. Vår hjerne er ikke "kablet" for slike oppgaver. Som vi skal se, har da også psykologisk forskning nærmest entydig vist at vi er meget dårlige til dette.

Også når det gjelder erfaringens betydning, overgeneraliserer Dreyfus og Dreyfus (1987). Deres modell for utviklingen fra nybegynner til ekspert er basert på studier av "the skill-acquisition process of airline pilots, chess-players, automobile drivers and adult learners of a second language" (s. 20). Når det gjelder disse områdene, er det liten grunn

til å betvile at Dreyfus og Dreyfus (1987) "observed a common pattern in all cases, which [they] call the *five stages of skill acquisition*" (s. 20). Feilen de gjør, er igjen at de generaliserer fra ervervelse av senso-motoriske ferdigheter (og sjakkferdigheter) til kognitive ferdigheter i sin alminnelighet.

Psykologisk forskning har avdekket at bestemte betingelser må være til stede for erfaringslæring. Læringsbetingelsene er tilstede i de spesielle kunnskapsområdene Dreyfus og Dreyfus hovedsakelig henter sin empiri fra. De generaliserer imidlertid til kunnskapsområder der disse betingelsene ikke er tilstede. Vi skal se at på slike områder opparbeider heller ikke "ekspertene" seg et stadig bedre skjønn med erfaring. Først vil jeg redegjøre for empirisk forskning som klart indikerer at Dreyfus og Dreyfus (1987) grenseløse tro på "the power of human intuition", er en myte.

Tradisjonen etter Meehl (1954): Myten om det gode skjønn avlives

Alle klager på sin hukommelse. Det er sjelden man hører noen klager på sine vurderinger. Dette er blant annet uttrykk for at folk flest vet svært lite om hva psykologien har avdekket om systematiske svakheter ved menneskers - eksperter inkludert - skjønnsmessige vurderinger. I dette avsnittet vil jeg forsøke å få dette fram.

Det hele begynte som nevnt med Meehls (1954) lille bok *Clinical versus Statistical Predictions*. Som det framgår av tittelen på boka, er den overordnede problemstillingen å avgjøre hva som er best av kliniske (skjønnsmessige) vurderinger og aktuariske metoder. Dette hadde lenge vært et stridsspørsmål i psykologien. Meehl argumenterer for at spørsmålet bare kan avgjøres på empirisk grunnlag. Han setter fram metodologiske regler for hvordan kontroversen "klinisk versus statistisk" kan avgjøres empirisk. Med *statistisk* prediksjon mener Meehl en vekting og kombinasjon av data på grunnlag av statistiske trender som er avdekket i en eller flere normative grupper. *Klinisk* prediksjon svarer til at data kombineres intuitivt eller skjønnsmessig. På grunnlag av 24 empiriske studier slutter Meehl at kliniske vurderinger (Har pasienten hjerneskade? Er pasienten schizofren? osv.) foretatt vha. formelle statistiske prosedyrer alltid er minst like nøyaktige, og ofte mer nøyaktige, enn skjønnsmessige vurderinger.

Innvendingene mot Meehls studier var av tre typer. Det ble hevdet:

- 1) Vurderingene er ikke representative for den type vurderinger som klinikerer står overfor i en virkelig klinisk praksis (f.eks. McArthur, 1956).
- 2) Klinikerer har ikke tilgang til informasjon som er essensiell for kliniske vurderinger, som intervjudata, møte med pasienten osv. (f.eks. Holt, 1958)
- 3) Konklusjonene holder ikke for "konfigurale" vurderinger.

Holdbarheten av disse tre innvendingene er testet ut i en rekke studier. Konklusjonene er nærmest entydige: Innvendingene holder ikke. La oss se på noen "klassiske" og representative studier som indikerer dette.

MMPI er den mest brukte personlighetstesten. Testresultatene summeres i en "profil" bygd opp av ti tallverdier som hver representerer testpersonens nivå på én skala eller "personlighetsdimensjon" (introvert-ekstravert; maskulinitet-feminitet; psykopati osv.). Det er alment akseptert at MMPI-tolkninger krever at man ikke trekker slutninger bare på grunnlag av verdien på én enkelt skala. Man må se på relasjonene mellom de ti skårene. Klinikere må altså foreta en "konfigural" vurdering av MMPI-profilen (jfr. innvending 3). Vurdering av MMPI-profiler er forøvrig også representativ for den type vurderinger klinikere hyppig møter i klinisk praksis (jfr. innvending 1).

Goldberg (1965) studerte erfarne klinikers evne til å vurdere om en pasient var nevrotisk eller psykotisk på grunnlag av slike MMPI-profiler. Han vurderte klinikerene opp mot hva som siden er blitt kalt "Goldbergs regel". Denne regelen legger sammen skårene på tre utvalgte skalaer, og trekker fra skårene på to andre skaler. Er resultatet under 45 er pasienten nevrotisk, ellers psykotisk. 861 MMPI-profiler skulle vurderes. Det viste seg at "Goldbergs regel" med god margin utkonkurrerte de mange erfarne klinikerne som ble studert. Goldbergs studie, og en rekke andre studier, står i skarp kontrast til den psykiatriske sykepleieren som Dreyfus og Dreyfus (1987) ukritisk tar til inntekt for sitt ekspertisesyn: "When I say to a doctor, 'the patient is psychotic', I don't know how to legitimize the statement. But I am never wrong." (s.34)

Sawyers (1966) omfattende review-studie viser klart at "naturlig" klinisk informasjon ofte ikke bidrar til bedre vurderinger (jfr. innvending 2). Studiene Sawyer tok for seg, viser blant annet at dersom klinikere gis mulighet til å basere sine vurderinger på intervjuer med pasientene i tillegg til pasientenes testdata, så faller kvaliteten på vurderingene i forhold til om klinikerne bare baserer seg på testdata. En rekke senere studier viser det samme. Generelt har det vist seg at dersom en beslutningstaker får tilgang til informasjon ut over de to-tre mest prediktive data, så faller kvaliteten på vurderingene. Slike studier viser altså hvor lett menneskelige beslutningstakere "drukner" i informasjon.

Goldbergs studie viste at én enkel regel utkonkurrerte erfarne klinikere på en type vurdering som disse rutinemessig foretar. Dawes' (1971, 1979) studier viser at enda enklere aktuariske ligninger enn "Goldbergs regel" kan slå ut erfarne menneskelige beslutningstakere. Dawes studerte blant annet inntakskomiteer ved amerikanske colleger. Disse komitéene skal vurdere hvilke studenter fra high school som skal tas inn, dvs. de skal predikere framtidige skoleprestasjoner. Dawes sammenlignet komitéenes vurderinger med helt trivielle lineære ligninger. Han fant at selv ligninger som bare trenger informasjon om én variabel (elevenes standpunktkaraktersnitt) predikerer

framtidige skoleprestasjoner bedre enn komitéen bestående av erfarne skolefolk. Komitéen hadde da ut over standpunkt-karaktersnittet også tilgang til elevenes eksamenskarakterer, kjennskap til kvaliteten på den skolen de kom fra, anbefalingsbrev fra denne skolen, samt et lengre intervju med elevene.

Stikk i strid med Dreyfus og Dreyfus' (1987) påstand om at eksperter generelt "intuitivt ser" hva de bør gjøre, har tradisjonen etter Meehl avdekket at intelligente, motiverte og erfarne menneskelige beslutningstakere i mange sammenhenger systematisk overgås av usofistikerte statistiske prosedyrer. Hundrevis av studier går i samme retning som Goldbergs, Sawyers og Dawes' studier (se Dawes, Faust og Meehl, 1989 for en oversikt). Dreyfus og Dreyfus (1987) ser helt bort fra denne empiriske forskningen. De holder seg til "the seemingly plausible arguments of Merleau-Ponty, Heidegger, and Wittgenstein", og gjentar de grundig tilbakeviste innvendingene mot Meehl (1954) fra 50-tallet, for eksempel: "But does poor performance of most subjects in the experiment really show that people deal poorly with new evidence in real-world situations?" (s. 42)

Hvorfor lærer vi (ofte) ikke av erfaring?

Dreyfus og Dreyfus (1987) avviser Dawes' (1971) funn med den begrunnelse at komitémedlemmene ikke er fulltids-eksperter på å vurdere elevens framtidige skoleprestasjoner. De mener: "It would be interesting to compare the predictive ability of models against those professionals responsible on a full-time basis for the admission decisions at elite undergraduate colleges. Our guess is that full-timers would fare better." (s. 45) Hvorfor de gjetter er ikke så lett å forstå. Det er nemlig gjort en rekke studier av betydningen grad av erfaring har for vurderingers kvalitet.

For eksempel, i en stor undersøkelse nylig basert på et representativt utvalg på 600 av USAs drøyt 3400 nevropsykologer, konkluderte man slik: "Except for a possible tendency among more experienced practitioners to overdiagnose abnormality, no systematic relations were obtained between training, experience, and accuracy across a series of neuropsychologic judgments." (Faust, Guilmette, Hart et al., 1988).

Dette er ikke en spesielt valgt studie. Man har lenge visst at kvaliteten på klinikers vurderinger ikke bedres med erfaring. Allerede Wiggins (1973) konkluderte på grunnlag av en omfattende studie av klinikere med ulik grad av erfaring: "Surprisingly, there is little empirical evidence that justifies the granting of 'expert' status to the clinician on the basis of his [or her] training, experience, or information-processing ability" (s. 131) I en bredt anlagt meta-studie nylig bekrefter Garb (1989) i hovedsak Wiggins konklusjoner.

Funnene gjelder ikke spesielt for klinikere. De gjelder alle kunnskapsområder der:

- 1) man ikke har en klar forståelse av hva som er en god vurdering.

2) man ikke får umiddelbar, utvetydig og konsistent tilbakemelding når man gjør feil.

På en rekke kunnskapsområder er kvaliteten på den tilbakemeldingen man får på sine vurderinger ikke i nærheten av å tilfredsstille de minimumskrav som er nødvendige (f.eks. Balzer, Doherty og O'Connor, 1989) Da finner man heller ingen forskjell på kvaliteten til erfarne og uerfarne klinkere (Kirkebøen, 1995). Grunnen til at få klager på sine egne vurderinger, skyldes selvfølgelig mangel på tilbakemelding.

Dreyfus og Dreyfus' baserer sin stadiemodell på hvordan vi lærer oss senso-motoriske ferdigheter som sykling, svømming og flyvning. Ved læring av slike ferdigheter får vi tilbakemelding som åpenbart tilfredsstiller de to betingelsene. Også i sjakk, som Dreyfus og Dreyfus ofte viser til, oppfylles i stor grad de to kravene til tilbakemelding. Derfor oppøver sjakkspillere etterhvert evnen til "intuitivt å se" gode, fornuftige trekk. I svært mange av de kunnskapsområder vi i et moderne samfunn har eksperter, er derimot betingelsene ikke oppfylt.

Hva er så årsakene til at intelligente, motiverte og erfarne menneskelige beslutningstakere systematisk overgås av usofistikerte statistiske prosedyrer? Det er to hovedgrunner til dette:

- 1) Våre kognitive begrensninger.
- 2) Strategiene vi benytter oss av for å overkomme disse begrensningene.

La oss først se litt på hvordan informasjonsteknologien bidro til å avdekke menneskers kognitive begrensninger.

Informasjonsteknologien og menneskets kognitive begrensninger

Filosofene og psykologer fra Platon til Descartes og Freud forklarte feilvurderinger, dårlig skjønn etc. med at våre høyere intellektuelle prosesser ble blokkert eller forstyrret av "lavere" drifter eller emosjoner. Informasjonsteknologien bidro på flere måter til å endre dette. Indirekte avdekket denne teknologien svakhetene ved behavioristisk psykologi, som dominerte fram til 50-tallet. Atferdspsykologene så bort fra menneskets "indre" begrensninger. De psykologene som fikk jobben med å tilpasse operatører til den nye komplekse krigsteknologien, fant imidlertid raskt ut at operatørene gjorde feil i omgang med den nye teknologien, uansett hvor mye læring de ble utsatt for. Dette reiste nettopp spørsmålet om hvilke "indre" eller kognitive begrensninger vi har. Blant annet slik bidro informasjonsteknologien til den såkalte kognitive revolusjon i psykologifaget (se Kirkebøen, 1993).

Først med informasjonsteknologien (i vid betydning) ble det mulig systematisk å forklare kognitiv feilfungering uten å måtte henvise til ikke-kognitive faktorer. Shannons (1948) informasjonsteori var kanskje det viktigste bidraget, blant annet fordi teorien gir et

presist kvantitativt mål på "informasjon". Teorien inspirerte dessuten psykologer til å betrakte mennesket som en informasjonskanal. Det naturlige prosjektet ble da å bestemme denne "kanalens" evne til å overføre "informasjon". Vi kan altså si at Shannons teori indirekte gjorde det mulig for psykologer å betrakte tenkning kvantitativt. Millers (1956) klassiske artikkel "The magical number seven plus or minus two", var motivert av Shannons informasjonsteori: "[T]hese experiments would not have been done without the appearance of information theory." (s. 81) I artikkelen gir Miller nettopp et kvantitativt mål (i *bit*) på begrensningen ved vår umiddelbare hukommelse. Snart ble metaforen "mennesket som informasjonskanal" erstattet med metaforen "mennesket som informasjonsprosessor". Moderne kognitiv psykologis hovedprosjekt har siden vært å bestemme vår såkalte "kognitive arkitektur", dvs. de kvantitative egenskapene til denne "informasjonsprosessoren" antatte bestanddeler.

Herbert Simon (1955) var forøvrig den første som forklarte menneskelig irrasjonalitet ut fra begrensningene i vår kognitive arkitektur. Han lanserte begrepet "bounded rationality". Idéen bak begrepet er nettopp at mennesker ikke tenker rasjonelt (sammenlignet med økonomenes normative modeller) nettopp fordi den informasjonsbehandling som kreves dersom vi skal foreta vurderinger i overensstemmelse med normative kriterier, overskrider vår kognitive kapasitet. Dette, sammen med de etterhvert mange studiene som viste eksperteres dårlige vurderingsevne, reiste nye spørsmål: Hvilke strategier bruker vi for å redusere kompleksiteten? Hvorfor gir disse strategiene i mange sammenhenger dårlige resultater?

Strategier og svakheter i menneskelig beslutningstaking

Pionerene var psykologene Daniel Kahneman og Amos Tversky. Deres studier viste at mennesker som beslutningstakere opererer etter regler, eller heuristikker, som avviker fra elementære (og ikke fullt så elementære) statistiske prinsipper (f. eks. Tversky og Kahneman, 1974). Disse reglene fungerer i mange situasjoner bra, men noen ganger - ikke minst i "kunstige" profesjonelle situasjoner - gjør de ikke det. Kahneman og Tversky viste ikke bare at menneskelige beslutningstakere avviker fra normative modeller. De viste at intuitivt baserte vurderinger og prediksjoner er *systematisk* avvikende.

Jeg vil kort redegjøre for de to sentrale slutningsstrategiene eller heuristikkene vi automatisk bruker for å forenkle komplekse vurderinger, nemlig tilgjengelighetsheuristikk og representativitetsheuristikk. Med "heuristikk" i denne sammenheng menes en prosedyre, tommelfingerregel eller "short-cut" som kraftig reduserer mulige løsninger på et problem eller antall mulige svar på et spørsmål.

Med *tilgjengelighetsheuristikk* (TH) menes tendensen til at vurderinger av hvor hyppig "noe" forekommer (i forhold til noe annet) påvirkes av hvor lett tilgjengelig dette

"noe" er (det vil både si hvor lett det er å legge merke til, hvor lett det er å huske og hvor lett det er å forestille seg). TH bidrar ofte til gode vurderinger av hyppighet. Problemet er at det er mange faktorer som *ikke* er korrelert med hyppighet, som kan påvirke tilgjengeligheten. Et eksempel på hvordan dette slår ut i eksperimentsituasjoner: Man blir spurt om hvilke ord det er flest av på engelsk: a) ord som begynner på R, eller b) ord som har R som tredje bokstav. De fleste svarer da a). Grunnen, antar man, er at det er lettere å forestille seg ord som begynner på R, selv om det er veldig mange ganger flere ord som har R som tredje bokstav.

Et annet eksempel. Forsøkspersoner får lest opp for seg en liste med kjente kvinnenavn og ukjente mannsnavn. De blir så spurt: Hva forekom hyppigst av kvinne- eller mannsnavn? Selv om det er like mange av hver, vil forsøkspersonene systematisk svare at det var flere kvinnenavn i lista. Kjente navn er lettere å huske - lettere tilgjengelig - enn fra før ukjente navn.

TH fungerer bra så lenge tilgjengelighet og hyppighet samvarierer. Antagelig fungerte TH meget bra i en før-moderne verden. Da var stort sett det som var viktig, både farer og muligheter, "slående" hendelser, og derfor lett både å legge merke til og huske. I vår moderne verden derimot, er store mengder statistisk informasjon langt mer reliabelt enn personlige erfaringer. Vi - eksperter inkludert - har en sterk tendens til å legge overdreven vekt på de siste. Det skyldes blant annet at vi automatisk bruker TH.

Bruk av representativitetsheuristikk (RH) eller "representativ tenkning" kommer til uttrykk på mange ulike måter. Generelt viser bruk av RH seg ved at vi har en tendens til å kategorisere objekter og fenomener på grunnlag av i hvilken grad slående trekk ved objektet/fenomenet ligner eller synes å være "representative" for kategorien. RH innebærer å anvende enkle likhetskriterier på kategoriseringsproblemer, dvs. at vi "reduserer" vurderinger til gjenkjenning.

Ett eksempel. Blir forsøkspersoner spurt: "En professor liker å skrive poesi, er ganske sky, og er liten av vekst. Hva tror du er hans felt? a) Kinesiske studier; b) Psykologi", så svarer de fleste a). Grunnen er at beskrivelsen er mer "typisk", eller representativ, for hvordan man forestiller seg en sinolog enn hvordan man forestiller seg en psykolog. Svært få vil ta hensyn til at det er mange, mange ganger flere professorer i psykologi enn i sinologi, dvs. man ser bort fra baseratene.

La oss se hva representativ tenkning innebærer i forhold til Bayes teorem, som kan betraktes som en idealisert modell eller formel for diagnostisk beslutningstaking. Teoremet kan uttrykkes i formelen: $P(D/S) = P(S/D) * PD/PS$. Relatert til klinisk diagnostikk, kan teoremet forstås slik: Vi kan tenke oss at vi står overfor en pasient som har et symptom S. Vi ønsker å bestemme sannsynligheten for at pasienten da også har sykdommen D. Denne sannsynligheten er det som i formelen betegnes $P(D/S)$. For å kunne beregne $P(D/S)$, må vi vite hvor ofte symptomet S er til stede hos de som har den

bestemte sykdommen D , altså sannsynligheten for S gitt D eller $P(S/D)$. Dette må multipliseres med sannsynligheten for at en tilfeldig person i den pasientgruppen (populasjonen) klinikerer behandler, har denne sykdommen, altså med PD . Videre må det deles med hvor vanlig det er at symptomet S , isolert sett, forekommer i denne populasjonen, dvs. med sannsynligheten for symptomet: PS .

Bruk av RH innebærer at man betrakter $P(D/S) = P(S/D)$, altså at man ser bort fra baseratene PD og PS . På grunn av bruk av RH, vil man ofte se - også blant eksperter - en sterk tendens til å sidestille for eksempel $P(\text{Traumatiske barndomsopplevelser}/\text{Psykiske problemer})$ med $P(\text{Psykiske problemer}/\text{Traumatiske barndomsopplevelser})$. Vi kan si at bruk av RH introduserer en symmetri i tanken som ikke eksisterer i virkeligheten.

Dette var en liten smakebit. Poenget er at det i stor grad er klarlagt hvilke ulike typer heuristikker eller strategier som automatisk styrer våre beslutninger og vurderinger (se for eksempel Nisbett og Ross, 1980 for en oversikt).

Relevansen for informatikere

Jeg vil nøye meg med å liste opp noen funn som jeg mener gir klare føringer for utforming av for eksempel beslutningsstøttesystemer:

- Mennesker har en unik evne til å "se" hva som er relevant informasjon i en situasjon, men vi har meget begrenset evne til å integrere den informasjonen som ligger i flere (unike) observasjoner.
- Informasjon ut over de 2-3 dataene med høyest prediktiv verdi resulterer i dårligere vurderinger når informasjonen kombineres klinisk (skjønnsmessig), men øker samtidig eksperters egen tro på vurderingenes kvalitet. (f.eks. Oskamp, 1965).
- Når eksperter hevder å ha brukt kompleks konfigural analyse for å nå bestemte vurderinger, så har man alltid kunnet konstruere enkle lineære modeller som adekvat dupliserer ekspertenes vurderinger (f.eks. Goldberg, 1968).

Samlet indikerer disse punktene hvilke typer vurderinger man bør tenke på å automatisere. Dersom man *a priori* kan spesifisere hvilken type informasjon eller input som er mest relevant for vurderingen, så er det god grunn til å tro at man kan finne fram til en formel som foretar en bedre integrering av input-informasjonen enn det menneskelige beslutningstakere er i stand til.

- Dersom menneskelige beslutningstakere i tillegg til informasjon om problemet gis tilgang til den aktuariske formelens konklusjon, og så blir bedt om å justere den siste, så resulterer det systematisk i dårligere vurderinger (f.eks., Goldberg, 1968; Sawyer, 1966)

Det vil altså ofte være en dårlig løsning å la menneskelige beslutningstakere korrigere resultatet av mekaniske rutiners integrering av informasjon.

- Skjønnsmessige vurderinger er avvikende på systematiske og predikerbare måter.

Det siste er spesielt viktig. Det gir grunnlag for å tro at det er mulig å utforme effektive beslutningsstøttesystemer (se f.eks. Kahneman og Tversky, 1979).

Når bør vi bruke hodet og når bør vi bruke datamaskiner?

Bayes teorem indikerer hvor komplisert beregning som ideelt sett må foretas når for eksempel en lege på grunnlag av flere data som alle sier noe om sannsynligheten for en diagnose, skal vurdere hva disse data samlet sier om sannsynligheten for diagnosen. Legen bruker svært ofte ingen hjelpemidler for å foreta denne vurderingen eller beregningen. Få synes å bekymres over at leger foretar slike vurderinger skjønnsmessig og intuitivt. Dette til tross for at de fleste er klar over hvor svak den menneskelige hjerne er til å vekte og kalkulere. Alle ville derfor antagelig reagere dersom kassamannen på RIMI tok handlekurven på øyemål. Han gjør ikke det. Han summerer. Derimot sitter fortsatt leger landet rundt og foretar langt mer komplekse vurderinger - som gjelder liv og død - på skjønn. Dette gjelder så godt som alle vurderinger. Alltid.

Paul Meehl (1957) stilte spørsmålet: "When shall we use our heads instead of the formula?" "Heads" refererer til en skjønnsmessig, intuitiv, subjektiv behandling av informasjon. "Formula" refererer til en matematisk, statistisk eller mekanisk kombinerings av informasjon. Meehl konkluderte i 1957 med at dersom vi har tilgang til en formel, så bør vi bruke hodet veldig, *veldig* sjelden. Grunnene til det er ikke blitt færre siden da. Meehl (1986) ser det nå slik:

"When you are pushing 90 investigations, predicting everything from the outcome of football games to the diagnosis of liver disease and when you can hardly come up with half dozen studies showing even a weak tendency in favor of the clinician, it is time to draw a practical conclusion." (s. 374)

Informatikere har mulighet til å trekke de praktiske konklusjonene. Forutsetningen er imidlertid at informatikere som utformer samspillet mellom teknologi og mennesker, gjør seg kjent med hva vi i dag vet om menneskelig skjønn og beslutningstaking.

Til slutt, en etisk betraktning

Mange vil kanskje peke på etiske problemer med å erstatte menneskelig skjønn med mekaniske rutiner. Dawes (1988) ser det slik:

"Friends tell me that important human judgment is often ineffable, unsystematic, and intuitive. I agree. And it is, therefore, often bad. Friends tell me that decisions that are effable, systematic, and explicit are dehumanized decisions. I agree. But they are "dehumanized" only for the decision maker, and I am concerned with the consequences for the people affected by the decisions. Bad decisions are dehumanizing for them." (s. 150)

Referanser

- Balzer W.K., Doherty M.E. og O'Connor R. (1989) Effects of cognitive feedback on performance. *Psychological Bulletin*, 106, 410-433
- Dawes R.M. (1971) A Case Study of Graduate Admissions: Application of Three Principles of Human Decision Making. *American Psychologist*, 26, 180-88
- Dawes R.M. (1979) The robust beauty of improper linear models in decision making. *American Psychologist*, 34, 571-582.
- Dawes R.M. (1988) You can't systematize human judgment: Dyslexia. I J. Dowie & A. Elstein (eds.) (s. 150-162) *Professional Judgment. A reader in clinical decision making*. Cambridge: Cambridge University Press.
- Dawes R. M., Faust D. og Meehl P. E. (1989) Clinical versus actuarial judgment. *Science*, 243, 1668-1674.
- Dreyfus H.L. (1965) *Alchemy and Artificial Intelligence*. The RAND Corporation Paper P-3244, December
- Dreyfus H. (1972) *What computers can't do*. New York: Harper & Row .
- Dreyfus H.L og Dreyfus E. (1987) *Mind over Machine. The power of human intuition and expertise in the era of the computer*. New York: The Free Press
- Faust D., Guilmette T.J. Hart K., Arkes H. R., Fishburne F. J. og Davey L. (1988) Neuropsychologists' training, experience, and judgment accuracy. *Arch Clinical Neuropsychology*, 3, 145-163.
- Garb H.N. (1989) Clinical judgment, clinical training, and professional experience, *Psychological Bulletin*, 105, 387-92.
- Goldberg L. R. (1965) Diagnosticians versus diagnostic signs: The diagnosis of diagnosis psychosis versus neurosis from the MMPI. *Psychological Monographs* 79 (9, hele no. 602)
- Goldberg L.R. (1968) Simple models or simple processes? Some research on clinical judgements. *American Psychologist*, 23, 483-496
- Hodges A. (1983) *Alan Turing: the Enigma*. London: Burnett.
- Holt R.R. (1958) Clinical and statistical prediction: A reformulation and some new data. *Journal of Abnormal and Social Psychology*, 56, 1-12
- Kahneman D. og Tversky A. (1979) Intuitive predictions: Bias and Corrective procedures. *TIMS Studies in Management Science*, 12, 313-17
- Kirkebøen G. (1993) Det mekaniske og det mentale. I T. Rasmussen og M. Sjøby (Eds.): *Kulturens digitale felt*. Oslo: Aventura, s.339-376
- Kirkebøen G. (1995) En bombe under soveputen? *Tidsskrift for Norsk Psykologforening*, 32, 426 -434
- McArthur C.C. (1956) The dynamic model. *Journal of Counseling Psychology*, 3, 168-71
- Meehl P.E (1954) *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.
- Meehl P. (1957) When shall we use our heads instead of the formula? *Journal of Counseling Psychology*, 4, 4, 268-273

- Meehl P.M. (1986) Causes and effects of my disturbing little book. *Journal of Personality assessment*, 50 (3), 370-375
- Miller G. A. (1956) The magical number seven plus or minus two: some limits in our capacity for processing information. *Psychological Review*, 81- 97
- Newell A., Shaw, J.C. og Simon H.A. (1958) Elements of a theory of human problem solving. *Psychological review*, 65, 151-166
- Newell A. og Simon H. A. (1981) Computer science as empirical inquiry: symbol and search. I Haugland (1981a, s. 35-66)
- Nisbett R.E. og Ross L.E. (1980) *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, New Jersey: Prentice-Hall
- Oskamp S. (1965). Overconfidence in case study judgment. *Journal of Consulting Psychology*, 63, 81-97
- Polanyi M. (1958) *Personal knowledge*. London: Routedledge&Kegan
- Sawyer J. (1966) Measurement and prediction, clinical and statistical. *Psychological Bulletin*, 66, 178-200.
- Shannon C. (1948) A mathematical theory of communication. *Bell Systems Tech. Journal*, 27, 379-423
- Simon H.A. (1955) A behavioral model of rational choice. *Quarterly Journal of Economics*, 69, 99-118
- Turing A. (1947/1969) Intelligent Machinery. I B. Meltzer og Michie D. (eds.) (1969, p. 3-23) *Machine Intelligence*, 5. Edinburgh: Edinburgh University Press.
- Turing A. (1950) Computing machinery and intelligence. *Mind*, 59, 433-460
- Tversky R. H. og Kahneman D. (1974) Judgement under uncertainty: heuristics and biases *Science*, 185, 1124-31
- Wiggins J. S. (1973) *Personality and Prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.
- Winograd T. og Flores F. (1986) *Understanding computers and cognition: a new foundation for design*. N. J.: Ablex Publishing Corp.