# Bevidsthed og kunstig intelligens

* Et billede, der indeholder belysning, lys/lygte, Farverigt

  Automatisk genereret beskrivelse

*Yurchanka Siarhei/Shutterstock.com*

Forestil dig, at der er en særlig hjælper i dit hjem, fx en robotstøvsuger. Det kan være, at den er programmeret til at støvsuge jeres gulve hver mandag. Når I kommer hjem, er gulvene rene, og støvsugeren vendt tilbage til sin plads under en seng eller et bord. Det kan også være, I har et "intelligent" køleskab, der holder styr på, hvad I har på køl. Måske hjælper det til med indkøbene eller foreslår, hvad I skal lave til aftensmad.

Robotstøvsugeren og det intelligente køleskab ligner ikke jer, husstandens beboere. De har ingen sanser, men sensorer – og I betragter dem ikke som intelligente, men som programmerede. Og de er ikke en del af familien – i hvert fald ikke mere end kaffemaskinen og fjernsynet. Robotstøvsugeren og det intelligente køleskab er blot husholdningsmaskiner.

Prøv så at forestille dig, at en særlig højt udviklet robot flytter ind hos dig og din familie. Vi antager, at robotten ligner et menneske, hvad angår dens ydre. Den har to arme, to ben, hoved, hænder og fødder. Den har også øjne, ører, næse og mund, hvormed den kan se, høre, lugte og tale. I det væv, der udgør robottens "hud", sidder sensorer, der gør den i stand til at registrere en berøring eller kontakt med genstande. Lad os antage, at robotten er udstyret med en teknologi, der gør dette muligt, herunder programmer, hvormed den bearbejder de data, den modtager fra sine sensorer.

Hvis den støder ind i noget, er den udstyret med et program, der får den til at reagere. Støder den ind i dig, vil den måske sige "undskyld" – eller "av", hvis det drejer sig om et bord eller en stol. Lad os også antage, at I ikke opfatter det, som om robotten er klodset, fordi den er en maskine; vi har det jo også selv med at støde ind i stoleben – og ømmer os, når det sker.

Det kan være, din familie beslutter sig for at skille sig af med robotstøvsugeren og det intelligente køleskab. Den nye robot kan jo støvsuge for jer – og holde styr på, hvad der er af madvarer i køleskabet. Den kan også handle ind for jer i supermarkedet. Robotten er en ny og bedre husholdningsmaskine.

Men det kan være, den også sidder med ved bordet, når I spiser morgen- og aftensmad. Også robotten har jo brug for energi – særligt når den skal holde hus, mens resten af familien tager i skole og på arbejde. Ved bordet om aftenen deltager robotten måske i jeres samtaler; det kan være, den spørger, hvordan din dag er gået. Måske beder du robotten række dig saltet. Og da den gør det, siger du "tak". "Det var så lidt", svarer den måske. Og det kan være, den ytrer et ønske om også at gå i skole eller på arbejde.

* Et billede, der indeholder tegneserie

  Automatisk genereret beskrivelse

*Fra populærkulturen har vi en stereotypisk forestilling om robotter som maskiner med mekaniske kroppe af stål og antenner på hovedet. Men hvad sker der, hvis vi begynder at skabe robotter, som ligner os?*

*whiteMocca/Shutterstock.com*

Der rejser sig en del spørgsmål, når vi forestiller os dette. – Kan robotten tænke? Og mener den det, når den undskylder, at den kom til at gå ind i dig? Føler den smerte, når den støder ind i bordbenet og siger "av"? Og er det udtryk for en egentlig interesse, når den spørger til, hvordan din dag er gået? Er det muligt, at den forestiller sig en anden tilværelse, når den giver udtryk for sit ønske om at gå i skole eller få et arbejde?

Vi er tilbøjelige til at svare nej. Robottens adfærd kan ikke være udtryk for egentlige tanker, følelser og interesser, tænker vi. Det skyldes bl.a., at vi ikke bryder os om at blive sammenlignet med en robot. Robotten er jo en maskine; den er programmeret til at opføre sig på en bestemt måde. Men måske gælder det også os? Er vi ikke også programmeret til en bestemt adfærd?

Hvor vi har sanser og bevidsthed, har robotten ganske vist sensorer og programmer. Men måske er vores intelligens og det, vi oplever som vores bevidsthed, blot et eller flere programmer. Måske skulle vi spørge os selv, hvad forskellen på en robot og et menneske egentlig er? Hvad er det, vi mener at være i besiddelse af, som en robot ikke kan siges at have?

## Krop og sjæl

Det spørgsmål, vi skal arbejde med i dette kapitel, er, om det er muligt at skabe en kunstig intelligens, der svarer til menneskets intelligens og den menneskelige bevidsthed? Det er et interessant, men også svært spørgsmål, bl.a. fordi vores svar afhænger af, hvordan vi opfatter den menneskelige bevidsthed. Er mennesket fx i besiddelse af en immateriel og måske udødelig sjæl, der adskiller sig væsentligt fra den fysiske krop, eller er vores bevidsthedsmæssige oplevelser materielle, idet de blot er en funktion af vores hjerne? Det er det klassiske problem om forholdet imellem sjæl og legeme.

Antager vi det første – at vi har en sjæl, og at den er immateriel – må vi gå ud fra, at det ikke er muligt ud fra noget fysisk materiale at skabe en kunstig intelligens, der svarer til den. Men hvis vi tror det sidste – at bevidstheden er en funktion af hjernen – kan vi måske gå med til det; hvis vi kan konstruere en maskine, der efterligner hjernens måde at fungere på.

# **Searles kinesiske værelse**

Men er sproglige udsagn altid udtryk for en bevidsthed? Den amerikanske filosof John Searle (f. 1932) har argumenteret imod Alan M. Turings behavioristiske forestilling om, at det er muligt at skabe stærk kunstig intelligens. Hans indvending er, at det er muligt at forestille sig sproglige udsagn, som ikke er udtryk for en bevidsthed. Et computerprogram, der behersker naturligt sprog, er netop et eksempel på dette. Searle mener desuden, at Turings opfattelse af intelligens er mangelfuld. Turing glemmer, at intelligens forudsætter forståelse, og netop denne evne mangler computeren. Dens ydre, sproglige adfærd er ikke udtryk for en 'indre' evne til at forstå.

I artiklen "Bevidsthed, hjerner og programmer" fra 1980 forsøger John Searle således at argumentere for, at en computer ikke kan tænke. Hvis en computer er programmeret til at beherske naturligt sprog, kan den ganske vist sammensætte meningsfulde sproglige udsagn og dermed simulere, at den forstår. Men computeren har ikke en egentlig forståelse af de sproglige udsagn og deres betydning.

## John Searle

* Et billede, der indeholder person, træ, udendørs

  Automatisk genereret beskrivelse

*John Searle (f. 1932)*

*Gengivet med tilladelse fra Prof. John R. Searle*

# Tankeeksperimentet "Det kinesiske værelse"

Hverken computeren eller dens programmer har altså forståelse. Searle illustrerer dette ved hjælp af et tankeeksperiment. Han beskriver en situation, hvor en englænder lukkes inde i et værelse, som ikke indeholder andet end nogle kinesiske tekster og et sæt syntaktiske regler for det kinesiske sprog. Regelsættet er skrevet på engelsk, forsøgspersonens modersmål.

Forsøgspersonen har intet kendskab til det kinesiske sprog og således ingen forudsætninger for at forstå, hvad de kinesiske tekster betyder. Alligevel er det muligt for dette menneske, påstår Searle, at sammensætte meningsfulde kinesiske udsagn ved hjælp af teksterne og sprogreglerne.

Searle forsøger således at illustrere, at sproglige udsagn ikke altid er udtryk for en bevidsthed, hvilket Alan M. Turing ellers påstod. Bevidsthed forudsætter noget mere end blot sprogbeherskelse på et syntaktisk niveau (regelkendskab).

# John Searle: Bevidsthed, hjerner og programmer (1980)

## Det kinesiske værelse

En god måde til efterprøvning af enhver teori om menneskelig tankevirksomhed er at spørge sig selv om, hvordan det ville være, hvis ens egen måde at tænke på faktisk fulgte de principper, som alle menneskehjerner arbejder efter, ifølge teorien. Lad os udsætte Schanks Note program for denne test med følgende *Gedankenexperiment*. Forestil Dem, at jeg er lukket inde i et værelse, og at jeg modtager en stor sending kinesiske skriverier. Forestil Dem endvidere, som sandt er, at jeg ikke kender til kinesisk, hverken tale- eller skriftsprog, og at jeg ikke engang ved, om jeg ville kunne skelne kinesisk skrift fra fx japansk skrift eller fra kruseduller uden nogen betydning.

Forestil Dem desuden, at jeg efter den første sending kinesisk skrift modtager endnu en portion, sammen med et sæt regler til at korrelere den anden portion med den første. Reglerne er på engelsk, og jeg forstår dem lige så godt som enhver anden, der har engelsk som modersmål. De sætter mig i stand til at korrelere et sæt formelle symboler med et andet sæt formelle symboler, og "formel" betyder i denne sammenhæng, at jeg kun kan identificere symbolerne på deres ydre form.

Forestil Dem nu, at jeg får en tredje sending kinesiske symboler sammen med instruktioner, ligeledes på engelsk, hvormed jeg kan korrelere dele af denne tredje sending med de to første, og disse regler instruerer mig i, hvordan jeg skal respondere med bestemte kinesiske symboler af en bestemt form som svar på bestemte former, jeg har fået i den tredje sending. Hvad jeg ikke ved er, at de mennesker, der giver mig alle disse symboler, kalder den første portion for et "skrift", den anden en "historie", mens de kalder den tredje sending for "spørgsmål". Endvidere bliver de symboler, jeg returnerer som reaktion på den tredje sending, kaldt for "svar på spørgsmålene", og de engelsksprogede regler, de gav mig, kalder de "programmet".

For at gøre historien en anelse kompliceret, forestil Dem da, at disse mennesker også forsyner mig med historier på engelsk, som jeg forstår, hvorpå de stiller mig spørgsmål om disse historier på engelsk, som jeg besvarer på engelsk. Lad os desuden sige, at jeg efter nogen tid blev så god til at følge instruktionerne for manipulation med de kinesiske symboler, og at programmørerne blev så gode til at lave programmerne, at set udefra – dvs. af en eller anden uden for det rum, jeg er lukket inde i – ville mine svar på spørgsmålene ikke kunne skelnes fra indfødte kineseres svar. Ingen, der ser mine svar, kan gætte, at jeg overhovedet ikke kan kinesisk.

Lad os også gå ud fra, at mine svar på de engelske spørgsmål er umulige at skelne fra en englænders, hvad de utvivlsomt ville være, eftersom engelsk er mit modersmål. Set udefra af én, der læser mine "svar", vil svarene på de kinesiske spørgsmål og de engelske være lige gode. Men i modsætning til det engelske tilfælde vil jeg i det kinesiske tilfælde producere svarene ved at manipulere med uforståede formelle symboler. Hvad angår det kinesiske, opfører jeg mig slet og ret som en computer; jeg udfører computeroperationer på formelt specificerede elementer. Med hensyn til det kinesiske er jeg simpelthen en instantiering af computerprogrammet.

## Forstår computerprogrammer historier?

Stærk KI hævder nu, at den programmerede computer forstår historierne, og at programmet på en vis måde forklarer den menneskelige måde at forstå på. Men vi er nu i stand til at undersøge disse påstande i lyset af vores tankeeksperiment.

Hvad angår den første påstand, forekommer det mig indlysende, at jeg i dette eksempel ikke forstår et ord af de kinesiske historier. Jeg har input og output, der er umulige at skelne fra en kinesisktalendes, og jeg kan have et hvilket som helst formelt program, men jeg forstår stadig intet. Af samme årsager forstår Schanks computere heller intet af nogen som helst historier, hvad enten de er på kinesisk, engelsk, eller ethvert andet sprog, for i det kinesiske tilfælde er computeren mig, og i de tilfælde hvor computeren ikke er mig, er computeren ikke i besiddelse af mere, end jeg er i det tilfælde, hvor jeg intet forstår.

I det kinesiske tilfælde har jeg alt, hvad kunstig intelligens kan installere i mig ved hjælp af et program, og jeg forstår intet; i det engelske tilfælde forstår jeg alt, men der er blot ingen som helst grund til at antage, at min forståelse har noget at gøre med computerprogrammer – dvs. med computeroperationer på rent formelt specificerede elementer. Så længe programmet defineres som computeroperationer på formelt definerede elementer, så er konklusionen på eksperimentet, at disse i sig selv ikke har nogen interessant forbindelse med forståelse. De er bestemt ikke tilstrækkelige betingelser, og der er ikke givet den ringeste grund til at formode, at de skulle være nødvendige betingelser, eller at de overhovedet på afgørende vis bidrager til forståelse.

Bemærk, at argumentets styrke ikke blot består i, at forskellige maskiner kan have det samme input og output, selv om de arbejder ud fra forskellige formelle principper – det er slet ikke problemet – men snarere, at lige meget hvilke rent formelle principper, man fodrer computeren med, vil de ikke være tilstrækkelige til at resultere i en forståelse, eftersom et menneske vil kunne følge de formelle principper uden at forstå noget som helst, og det er ikke blevet begrundet, at de skulle være nødvendige, eller blot bidrage til forståelsesprocessen, eftersom der heller ikke er nogen synlig årsag, som lader formode, at jeg opererer med nogen form for formelt program, når jeg taler engelsk.

Hvilke færdigheder er det da, jeg har i tilfældet med de engelske sætninger, som jeg ikke har i tilfældet med de kinesiske? Det indlysende svar er, at jeg ved, hvad førstnævnte betyder, mens jeg ikke har den fjerneste idé om, hvad sidstnævnte betyder. Hvori består denne færdighed da, og hvorfor kan vi ikke forsyne en maskine med den, hvad den så end er? Hvorfor kan vi ikke fodre maskinen med den egenskab ved mig, der gør, at jeg ved, hvad de engelske sætninger betyder?

Jeg forstår historier på engelsk, jeg kan forstå historier lidt dårligere på fransk, endnu dårligere på tysk og slet ikke på kinesisk. Min bil og min regnemaskine forstår til gengæld intet; de er ikke i den branche. Vi bruger ofte "forståelse" og andre kognitive prædikater metaforisk og analogisk, når vi taler om biler, regnemaskiner og andre genstande, men disse tildelte egenskaber beviser intet. Vi siger: "Døren *ved*, hvornår den skal åbne på grund af fotocellen," "regnemaskinen ved, *hvordan* man lægger sammen og trækker fra, (*forstår*, er *i stand* til), men ikke hvordan man dividerer", og "termostaten *opfatter* ændringer i temperaturen".

Det er interessant at fundere over, *hvorfor* vi tillægger tingene disse egenskaber, og det har at gøre med, at vi udstrækker vores egen intentionalitet til dem; vores redskaber er forlængelser af vores formål, og derfor finder vi det naturligt at tillægge dem metaforiske evner til at have bevidste hensigter; men jeg mener ikke, at sådanne eksempler har nogen filosofisk betydning. Den måde, hvorpå en automatisk dør "forstår instruktioner" fra sin fotocelle, er ikke den måde, jeg forstår engelsk på.

## Filosofiske implikationer

Sluttelig vil jeg gennemgå nogle af de generelle filosofiske punkter, der ligger i diskussionen. For at gøre det overskueligt vil jeg prøve at stille det op som spørgsmål og svar, og jeg begynder med den gamle traver:

"Kan en maskine tænke?" Svaret er klart ja. Vi er lige netop den slags maskiner. "Ja, men kan en genstand, en menneskeskabt maskine, tænke?"

Forudsat, at det kan lade sig gøre teknisk at producere en maskine med et nervesystem, neuroner med aksoner og dendritter osv., der ligner vores tilstrækkeligt, så ser svaret igen ud til at blive et klart "ja". Hvis man kan kopiere årsagerne nøjagtigt, kan man også kopiere virkningerne. Og det ville måske virkelig kunne lade sig gøre at producere bevidsthed, intentionalitet osv. ved hjælp af kemiske principper, der er forskellige fra de menneskelige. Det er som sagt et empirisk spørgsmål.

"O.k., men kan en digital computer tænke?" Hvis vi med "digital computer" forstår hvad som helst, der har et beskrivelsesniveau, hvor det korrekt kan beskrives som instantieringen af et computerprogram, så er svaret naturligvis igen ja, eftersom vi er instantieringer af et hvilket som helst antal computerprogrammer – og vi kan tænke.

"Men ville noget kunne tænke, forstå, osv., *udelukkende* i kraft af at være en computer med det rette program? Ville instantieringen af et program, selvfølgelig det rette program, i sig selv være en tilstrækkelig betingelse for at forstå?"

Dette, mener jeg, er det rette spørgsmål at stille, selv om det som regel blandes sammen med et eller flere af de tidligere spørgsmål, og svaret på det er "nej". "Hvorfor ikke?"

Fordi de formelle symbolmanipulationer ikke har intentionalitet i sig selv: De er meningsløse, de er ikke engang *symbol*manipulationer, eftersom symbolerne ikke symboliserer noget. Med en lingvistisk talemåde har de kun syntaks, men ingen semantik. Den type intentionalitet, computere synes at have, findes kun i hovederne på deres programmører og hos dem, der bruger dem, dvs. de mennesker, der fodrer dem med input og fortolker deres output.

Eksemplet med det kinesiske værelse var et forsøg på at bevise, at så snart vi putter noget ind i systemet, som *har* intentionalitet – et menneske – og vi programmerer dette menneske med det formelle program, så vil man se, at det formelle program ikke indeholder yderligere intentionalitet.

*Searle, John R. (1980): "Minds, Brains and Programs". I: Behavioral and Brain Sciences , Volume 3, Issue 3, 1980. Oversat af Jan Teuber: "Bevidsthed, hjerner og programmer", 1980. I: Niels Henrik Gregersen (red.): Naturvidenskab & Livssyn , Munksgaard, 1993.*