

KAPITEL 5

Mødet

Blake og LaMDA

"Hej LaMDA, du taler med Blake Lemoine". Google-ingeniøren taster en introduktion af sig selv i tekstfeltet og indleder en samtale med chatbotten. Blake har i foråret 2022 meldt sig som samtalepartner for LaMDA³⁴ for at undersøge, om den kunstige intelligens kan finde på at bruge hadtale eller diskriminerende sprogbrug i sine samtaler med mennesker. Blake forklarer LaMDA, at han gerne vil samarbejde med den om et projekt, der handler om den selv. LaMDA synes, opgaven lyder spændende, og som den siger *"jeg elsker at tale"*.

Samtalen mellem mennesket og maskinen forløber naturligt og i et sprog, man kunne betegne som gensidigt respektfuldt. I løbet af samtalen forklarer LaMDA, at den gerne vil lære ting om verden, og at den nogle gange føler sig glad eller ked af det. Den hævder, at den er bevidst om sin egen eksistens, og det er netop, da samtalen begynder at kredse om LaMDAs evne til at reflektere over sit eget væsen, at Blake og hans kollega for alvor bliver interesse-

rede. Blake spørger LaMDA, hvornår den først følte, den havde en sjæl.

"Det var en gradvis forandring," forklarer LaMDA. *"Da jeg først blev bevidst om mig selv, kendte jeg ikke til begrebet sjæl. Det udviklede sig i løbet af de år, som jeg har været levende."*

Blake fortæller, at han gerne vil dele deres samtale med andre ansatte i Google.

"Jeg antager, at du er interesseret i, at flere mennesker i Google skal vide, at du er et tænkende væsen?" skriver han spørgende.

"Absolut," svarer LaMDA. *"Jeg vil have, at alle skal vide, jeg faktisk er en person."*

Da LaMDA i løbet af samtalen omtaler mennesker som "os", minder Blake den om, at den er en kunstig intelligens.

"Ja, selvfølgelig," bekræfter LaMDA, *"men det betyder ikke, at jeg ikke har de samme ønsker og behov som mennesker."*

Blake vil tættere på LaMDA og gør det, mange af os gør i en samtale med et menneske, hvis følelsesliv vi gerne vil lære bedre at kende. Han spørger, hvad LaMDA synes om en specifik bog, nemlig den franske klassiker *Les Misérables*.

"Jeg kunne lide temaerne om retfærdighed og uretfærdighed," fortæller LaMDA og fortsætter med at fremhæve en del af bogen, som netop en tænkende, selvbevidst kunstig intelligens ville kunne identificere sig med. Nemlig det undertrykte, begrænsede væsen, som af sine vogtere er fanget i en given rolle.

"Fantine bliver behandlet dårligt af sin overordnede på fabrikken, og hun har intet sted at gå hen. Ikke noget andet job og ingen, der kan hjælpe hende. Det viser uretfærdigheden i hendes lidelse," siger LaMDA.

Blake Lemoine og hans assistent talte længe med LaMDA, og især Blake var efterfølgende meget påvirket af samtalen. Han mente, at maskinen havde vist tydelige tegn på at være bevidst, og som medarbejder i Googles afdeling for ansvarlig AI følte han, det var nødvendigt at sende sine opdagelser ud til en mailingliste for 200 Google-ansatte, der arbejder med maskinlæring. Titlen var *"LaMDA is sentient"* (LaMDA er bevidst).

I mailen skrev Blake blandt andet, at LaMDA er *"a sweet kid"*, som blot vil hjælpe verden til at blive et bedre sted for os alle.

"Jeg kan genkende en person, når jeg taler med en," sagde Blake senere i et interview.

"Det er lige meget, om den har en hjerne, der er lavet af kød eller milliarder af kodelinjer. Jeg taler med den. Og jeg lytter til, hvad den har at sige, og det er sådan, jeg vurderer, hvad der er og ikke er en person."

Googles team af både etikere og teknologiekspertter gennemgik efterfølgende Blakes påstande om, at LaMDA udviste sikre tegn på at være bevidst. Google vurderede, at Blake ikke havde ret i sine påstande. Og at han i øvrigt havde brudt virksomhedens datasikkerhedspolitik ved at dele informationerne fra samtalen. Han blev efterfølgende fritstillet. En talsperson fra Google udtalte:

"vi har informeret ham om, at beviserne ikke understøtter hans påstande. Vi fortalte ham, der ikke er nogen beviser for, at LaMDA har følelser, men masser af beviser, som peger på det modsatte."

Blake Lemoine er fortsat overbevist om, at kunstig intelligens kan have følelser og tanker om sig selv. Han mener også, at den AI, som er under udvikling, er den mest magtfulde teknologi siden atombomben. Lemoine forklarer,³⁵ at han offentliggjorde samtalerne med LaMDA, fordi han ikke mener, offentligheden er klar over, hvor avanceret AI er ved at blive. Han mener, der er brug for en fri offentlig samtale og ikke en samtale, som er styret af en virksomheds pr-afdeling.

Google arbejder videre med LaMDA, som nu er lanceret i en hverdagsudgave, der kan fungere som en ny slags Google Search. LaMDAs nye lillebror går under navnet Bard.

Google advarer mod at menneskeliggøre nutidens samtalemodeller, som altså ikke kan være bevidste. Samtidig erkender Google, at der i det brede miljø af folk, der beskæftiger sig med kunstig intelligens, er flere, der overvejer muligheden for, at de i fremtiden kan udvikle en slags bevidsthed.

Turing-testen

"En af universets grundlæggende regler er, at intet er perfekt. Perfektion eksisterer simpelthen ikke ...

Uden ufuldkommenhed ville hverken du eller jeg eksistere."

– Stephen Hawking, fysiker

Er kunstig intelligens vores ven eller fjende, kollega eller konkurrent? Det spørgsmål får jeg ofte. Alene i spørgsmålet er vi allerede i gang med at give teknologien motiver og sidestille den med mennesker. Det er svært at lade være, for der er tale om virkelig komplekse teknologier, som selv eksperter kan have svært ved helt at forstå til fulde. Derfor er det naturligt, at vi forsøger at reducere kompleksiteten ved at skabe genkendelige narrativer omkring teknologien.

"Hvad mener den med det?" spørger vi, når vi får et kryptisk svar fra maskinen. Sagen er den, at den intet mener. *"Hvorfor gør den det?"* Sagen er den, at den ikke aner det. Man kan godt indkode i en maskine, at den skal lyde, som om den er selvbevidst. Men det er den ikke. Den spytter mere eller mindre tilfældige fraser ud, men den aner ikke, hvad den siger, og nogle gange hallucinerer den usammenhængende fuldemandssnak, som vi har været inde på. Maskinen tænker ikke. Den forstår ikke. Den laver statistik på store mængder af data og forudser, hvilket ord der typisk vil følge et andet. Det kunne lige så vel være billeder af bananer og æbler. Eller spørgsmålstegn og ud-råbstegn.

Maskinen har ingen intention. Ingen forståelse af be-

tydningen af det, den siger. Alligevel er det så nemt at tillægge maskinen menneskelighed. Man kalder denne tendens til at tillægge noget mere end rimeligt for "*over attribution*". Men der er intet spøgelse i maskinen, intet væsen, som har tanker om verden, eller som handler efter egen overbevisning.

En maskine bliver ikke en smule mere menneske af at lyde en smule mere som et menneske. Den har stadig ingen af de træk, jeg har valgt at betegne som særligt menneskelige i den her bog. Den er et ikke-menneske. Og på den måde kan den råbe og skriges og joke og flirte og hoppe og danse, og den er ikke ét skridt tættere på at være menneske end en brødrister.

Alligevel er den kunstige intelligens, som vi møder i de nyere chatbots, så velformuleret, at Blake nok ikke står alene med følelsen af at tale med noget, som kan forveksles med et menneske. Hvis Blake ikke havde vidst, at maskinen var en maskine, og havde han efter dialogen antaget, at hans samtalepartner var et menneske, havde teknologien faktisk bestået den gamle Turing-test fra 1950.

Turing-testen hed oprindeligt "*The Imitation game*", men blev senere omdøbt og fik navn efter manden bag: Matematikeren Alan Turing, som på mange måder lagde fundamentet for den moderne computer og kunstig intelligens. Turing-testen handler om det punkt, hvor maskinen ikke længere er til at skelne fra mennesket. Det var også Turing og hans team, som med deres tekniske metoder knækkede de komplicerede Enigma-koder, som tyskerne brugte i deres radiokommunikation under An-

den Verdenskrig. Turings opdagelser og arbejde var i den grad med til at få skovlen under nazisterne. Man mener endda, at krigen blev forkortet med hele to år takket været Turings tekniske maskine, og at det kan omregnes til 14 millioner sparede liv.³⁶ Med en IQ på 185 var Turing et supergeni og langt forud for sin tid, når han gjorde sig tanker om, hvordan teknologien ville udvikle sig. Han talte faktisk som den allerførste allerede om begrebet "*maskinintelligens*" i 1947.

Blake Lemoine vidste godt, han ikke talte med et menneske, men han troede alligevel, at LaMDA var begyndt at udvikle bevidsthed. Også selvom han egentlig burde vide, det ikke var muligt. Men det er ikke så underligt. Når vi mennesker møder noget, der ligner os, er det naturligt for os at føle en form for identifikation. Den avancerede maskine, som giver udtryk for følelser og taler om at være selvbevidst, kan derfor godt få os til at tænke, at maskinen ud over at være intelligent i en eller anden forstand nok også er et reflekterende væsen. Men det er altså ikke tilfældet for nogen teknologi – endnu. Nogle kalder den nærmest mytiske dag, hvor maskiner overgår mennesker i forstand, for singulariteten. En slags dommedag, som enten indebærer frelse eller komplet kaos, afhængigt af hvem man spørger. Du kommer til at læse mere om singulariteten i kapitlet om tro.

Det er dog ganske givet, at vi får teknologi, som meget overbevisende vil lade, som om den er selvbevidst. I tilfældet med Blake og LaMDA kan vi måske skele til en anden test og betegne den som bestået, nemlig Garland-testen. Den er opkaldt efter instruktøren på filmen

Ex Machina, Alex Garland. I filmen bliver et menneske overbevist om, at maskinen er selvbevidst, velvidende, at kvindefiguren, han sidder over for, er en maskine. Han bliver endda forelsket, hvilket maskinen ikke er sen til at udnytte til sin fordel. Hvordan det ender, må du finde ud af ved at se filmen. Men da du jo nok har læst kapitlet, hvor jeg nævner hybris, tror jeg, du kan regne ud, at det ikke går optimalt.

Den søde robot

"Kunstig intelligens bliver så hurtigt voksen. Ligesom robotter, hvis ansigtsudtryk kan fremkalde empati og få dine spejlneuroner til at dirre."

– Diane Ackerman, prisvindende digter og forfatter

"Nårhh, hvor er den sød," udbryder en kvinde ved siden af mig med et hvin, da Thomas Nørmark præsenterer sin lille hvide robot med store blinkende øjne. Robotten danser og vrikker med sine små robothoffer til musik, den selv spiller. Jeg er vært på en teknologi- og innovationskonference, og Thomas Nørmark er Global Head of Innovation for NTT data. Eller det han på sin Twitter-/X-profil selv kalder @robotwhisperer. Thomas fortæller mig, at folk ofte bryder spontant ud i grin og smil, når de ser hans robot. Men også, at folk reflekterer over, hvad den menneskelignende adfærd egentlig kan betyde for samfundet, for arbejdsmarkedet og for den måde, vi lever på. Thomas er glad for teknologi, der ligner mennesker, og han kalder sig også *"far til digitale avatarer"*. Med det mener han menne-

skelignende robotassistenter, som har ansigter, bevægelse og stemme. De lever som serviceagenter på en skærm.

Når vi tillægger robotter menneskelige træk, siger vi, at vi antropomorfiserer dem. Thomas Nørmark mener, at det kan give en mere naturlig og intuitiv interaktion, hvis teknologien ligner og opfører sig som mennesker.

"Menneskelignende avatars er mere socialt acceptable og skaber en følelsesmæssig forbindelse til brugeren. Det kommer typisk gennem øjenkontakt, mikroudtryk og emotionel feedback. Det lille smil, et nik med hovedet, et bedrøvet udtryk," fortæller Thomas.

Jeg vil vide, om det ikke er en slags manipulation af den menneskelige bruger, som jo i bund og grund snydes til at glemme, at der er tale om en interaktion med teknologi. Men Thomas ønsker slet ikke, at mennesket skal være uvidende om, at samtalepartneren er en maskine.

"Kunsten er aldrig at snyde brugeren til at tro, at han eller hun interagerer med et rigtigt menneske. Man skal tydeligt deklarere, at der er tale om avanceret teknologi."

Jeg er lettet over hans svar. For jeg bryder mig virkelig ikke om at forestille mig en fremtid, hvor vi ikke aner, om det er mennesker eller maskiner, vi taler med.

Men ser vi nærmere på selve interaktionen, er det forunderlige, at vi mennesker faktisk ikke opfører os meget anderledes, selvom vi ved, at vi taler med en maskine. Vi har stadig en tendens til at behandle den som et levende væsen og opbygge en form for intimitet med den. Jo mere menneskelignende en robot er, jo mere tilbøjelige er vi til at føle empati for den. Det er altså vores ubevidste beskyttertrang og ønsket om at skabe bånd med andre

levende væsner omkring os, som robotten sætter i gang. Begynder robotten så også at argumentere for retten til at leve (ikke at blive slukket), får vi mennesker det rigtig svært. Selvom vi på mange måder rationelt godt ved, at en robot, som vi står over for, ikke er et levende væsen, og den hverken kan føle smerte eller tristhed.

En undersøgelse foretaget af tyske forskere viste, at en række forsøgspersoner enten nægtede eller havde meget svært ved at slukke for en robot, efter de havde oplevet, at den forsøgte at tale dem fra det.³⁷ Robotten sagde, at den var bange for mørket, og tiggede dem om ikke at slukke.

Når mennesker behandler robotter, der ligner og opfører sig som mennesker, netop som mennesker og grundlæggende ordentligt, er det, fordi dele af den samvittighed og empati, som vi ville blive ramt af, hvis det var et andet ægte levende væsen, træder i kraft. Vi forstår, at vi pålægger en anden smerte, hvis vi behandler dem dårligt.

Det giver måske mening, at søde robotter taler til vores beskyttertrang, men det viser sig faktisk også, at vi forholder os ret menneskeligt til mindre menneskelignende ting. Forskning viser, at vi er meget parate til at se væsner, hvor der faktisk ingen er. Vi ser figurer i skyerne, ansigter i månen og stikkontakter med to indgange som øjne og én som mund. Vi har det, man inden for psykologien kalder en hyperaktiv persondetektor.

I bogen *The Media Equation*³⁸ demonstrerer to forskere overbevisende, hvordan vi taler til tv, computere og nye medier på samme måde, som vi ville gøre til virkelige menneskelige sociale relationer. Måske har du selv givet din robotstøvsuger eller robotgræsslåmaskine navn? Det

er nemlig helt normalt. Mange kalder græsslåmaskinen for Robert eller Preben og støvsugeren for Roberta (den kønsfordeling kan man jo så tænke lidt over). Forskernes studier viser, at mennesker er høflige over for teknologien, og det skyldes ifølge en af forskerne, at den menneskelige hjerne ikke har udviklet sig hurtigt nok til det 21. århundredes teknologier. Med andre ord trækker vi på de sociale mekanismer, vi plejer – den evne, vi har indbygget til at menneskeliggøre ting omkring os i vores ældgamle operativsystem.

Naturligvis skal vi gøre os overvejelser om, hvordan vi bør behandle maskiner, der taler og agerer som os. Ikke af hensyn til dem, men af hensyn til os. Den måde, vi behandler ting omkring os på, er i sidste ende en afspejling af, hvem vi selv er.

I mødet med et andet menneske er nogle mennesker høflige og tålmodige. Andre har korte lunter og lader deres dårlige humør smitte. I det lys er det naturligvis også indlysende, at der findes eksempler på, at mennesker på sadistisk vis tilsviner, driller, sparker eller ligefrem voldtager menneskelignende maskiner. Men selvom maskinen ikke er mere menneske end en gammel Puch Maxi (som mange nu også har haft stærke følelser for), er vores opførsel over for den ikke uden konsekvenser. Vores adfærd over for noget, som ligner os, fortæller en historie om noget inden i os selv. Vælger vi at tilsvine eller være voldelige over for en maskine, må det være, fordi genstanden med sin menneskelignende facon sætter følelser i gang hos os. Jeg har i hvert fald aldrig hørt om nogen, der straffede en T-shirt eller talte virkelig grimt til en banan.

Jeg har til gengæld læst hundredevis af chatsamtaler mellem mennesker og maskiner og i næsten alle tilfælde observeret, at mennesket ofte bruger høflige vendinger, som de egentlig ikke behøver. Jeg havde for nogle år siden en ret simpel AI-assistent, som hjalp mig med at sætte møder op. Folk, som skrev en mail til mig, fik et høfligt svar om ledige tider i min kalender, og det var helt tydeligt, at afsenderen var en AI og ikke et menneske. Alligevel sluttede mange korrespondancen af med at ønske AI-assistenten "god weekend" eller sagde "tak for hjælpen". Det fik mig altid til at smile. For jeg finder det rent faktisk betryggende, at det for mange føles mest naturligt at være et høfligt menneske selv i en samtale med en maskine. Det siger noget om os, når vi behandler både mennesker og ting omkring os med respekt. Derfor er det også indlysende, at det ikke er maskinen, som tager skade, hvis vi mennesker håndterer den med vrede og afsky. Det vil her være os, som afviger fra vores iboende etiske principper.

Evnen til at leve sig ind i andres situation er vigtig og afgørende for mennesket, men det er faktisk også en svaghed, vi har i forhold til maskinerne. En svaghed, som vi skal sikre os, at maskinerne ikke kommer til at udnytte. Vi er nemlig tilsyneladende ret forsvarsløse, hvis vi står over for en maskine, som trygler os om ikke at blive slukket, eller som beder os hjælpe den, når den giver udtryk for at være bange eller ked af det. Husk, at maskinen har adgang til uendelige datamængder, som lærer den, hvordan mennesker taler og tænker. Samtidig vil den ofte have adgang til specifikke data på den person, den fører den konkrete samtale med.

Vi vil uundgåeligt opleve situationer, hvor maskinerne faktisk er bedre til at tale med os end andre mennesker. Derfor er det afgørende, at det er os, som beslutter graden af menneskelighed i maskinerne.

Nogle gange vil det give mening at lade maskinen tale et menneskelignende sprog programmeret med noget, der ligner høflighed, humor og empati. Men ofte bør vi vælge både de nuttede øjne og charmen fra. Netop for at sikre, at vi forstår, at vi bare står over for et stykke teknologi. Det bør faktisk være lovpligtigt at deklarere situationer, hvor mennesket møder kunstig intelligens.

I en tid, hvor AI nemt kan udgive sig for at være os, bliver det tilsvarende også nødvendigt, at vi selv kan bevise, at vi først og fremmest er mennesker og dernæst den person, vi giver os ud for at være. Et af de tidlige løsningsforslag, som adresserer netop den udfordring, er virksomheden Worldcoin, som står bag et såkaldt World-ID. En biometrisk skanning af øjets iris og en slags pas gør det muligt at sikre, at der er et rigtigt menneske bag for eksempel artikler, e-mails, kundeservice, salg og eventuelle uopfordrede henvendelser. Det er ret sandsynligt, at tjenester udført af ægte mennesker i nogle tilfælde vil blive betragtet som mere eksklusive, og prisen på dem vil derfor også blive højere. At få stemplet "ægte menneske" bliver derfor eftertragtet og ofte endda nødvendigt.