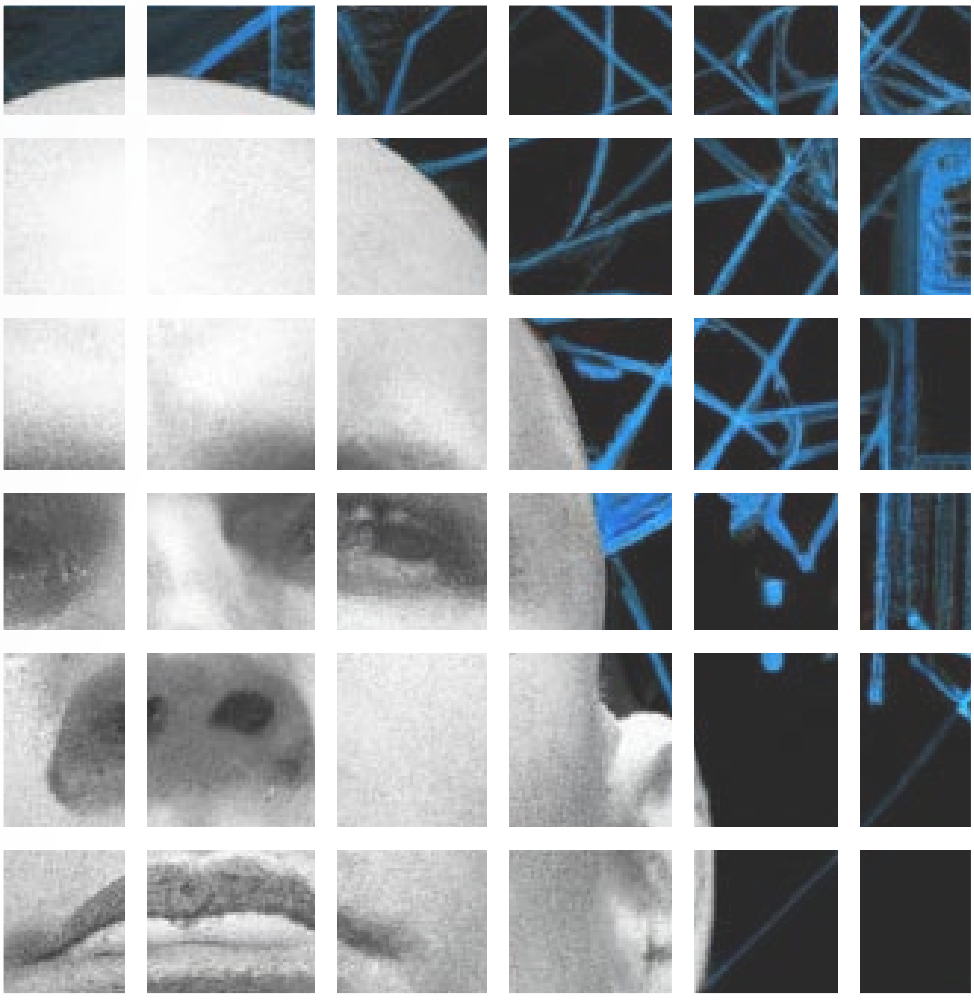
# BAGGRUND

## TIL LÆREREN



## BALANCEN MELLEM UBEKYMRET TEKNOLOGI-FASCINATION OG DYSTOPI

I dette projekt har vi valgt at lytte til det flertal af forskere og eksperter indenfor kunstig intelligens, der ikke mener, at det indenfor overskuelig fremtid vil blive muligt at udvikle kunstige intelligenser, der har bevidsthed og fri vilje. Andre mener dog godt, at kunstige intelligenser kan blive selvstændige, og at det måske ikke ligger så langt ude i fremtiden. Blandt andet har Elon Musk trukket overskrifter med advarsler om kunstig intelligens, der får fri vilje.

Senest har Elon Musk i marts 2023 været medunderskriver på et åbent brev, der opfordrer til at sætte udviklingen af kunstig intelligens på pause, indtil der bliver udviklet retningslinjer og regler på området. I det åbne brev er det dog ikke først og fremmest frygten for kunstige intelligenser med fri vilje, der fremføres, men en række andre bekymringer. Blandt andet stiller brevet disse 4 retoriske spørgsmål:

- *Should we let machines flood our information channels with propaganda and untruth?*
- *Should we automate away all the jobs, including the fulfilling ones?*
- *Should we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us?*
- *Should we risk loss of control of our civilization?*

Foruden Elon Musk har en lang række tech-iværksættere og forskere skrevet under på brevet. Så der er også blandt mange fagfolk en reel bekymring for, hvad konsekvenserne af den hurtige udvikling af kunstig intelligens kan føre med sig.

I dette forløb har vi forsøgt en balanceret tilgang. På den ene side nedtoner vi (særligt i fase 1) risikoen for den rene dystopi med maskiner der overtager verdensherredømmet. På den anden side arbejder vi med en tilgang, hvor udvikling og brug af kunstig intelligens ikke – som det i vid udstrækning er tilfældet i øjeblikket – udelukkende bør styres af økonomiske interesser og en trang til udforske, hvor langt man teknologisk kan nå.

Eleverne skal i deres vurdering af teknologier med kunstig intelligens arbejde med det "benspænd", som er udviklet på AU: *"Research Unit for Robophilosophy and Integrative Social Robotics":*

> *Robotter (her kunstig intelligens) bør kun udføre opgaver, som mennesker bør udføre – men som mennesker ikke kan udføre.*

Dette krav eller benspænd er i tråd med det krav, som stilles i det føromtalte åbne brev:

> *Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable.*

*https://futureoflife.org/open-letter/pause-giant-ai-experiments*

## DRØMMEN OM EN "GENEREL KUNSTIG INTELLIGENS"

Som det fremgår af elevteksten i kopiark 2, så er de kunstige intelligenser, der findes i dag alle sammen "specifikke"– dvs. kunstige intelligenser, der kun er udviklet til at kunne mestre et bestemt område. En skakcomputer kan fx spille skak, og ChatGPT er ekspert i sprog.

Den menneskelige hjerne er en "generel intelligens", fordi den kan alle mulige forskellige ting – tænke logisk, bruge sproget, styre vores krop, afkode billeder osv. Indtil videre har man ikke kunnet udvikle en "generel kunstig intelligens". De kunstige intelligenser er altså stadig "ensporede".

Men det er helt klart et mål for nogle af udviklerne at nærme sig generel kunstig intelligens. (På engelsk kaldes generel kunstig intelligens AGI - artificial generel intelligens) Firmaet bag ChatGPT4 hedder Open AI, og de har en afdeling på deres hjemmeside, der hedder "Planning for AGI and beyond".

Her skriver de blandt andet:

> *AGI has the potential to give everyone incredible new capabilities; we can imagine a world where all of us have access to help with almost any cognitive task, providing a great force multiplier for human ingenuity and creativity.*
>
> *On the other hand, AGI would also come with serious risk of misuse, drastic accidents, and societal disruption. Because the upside of AGI is so great, we do not believe it is possible or desirable for society to stop its development forever; instead, society and the developers of AGI have to figure out how to get it right.*

Hos Open AI virker de ret sikre på, at generel kunstig intelligens er en mulighed. I hvert fald forskes der massivt for at opnå det. De fortæller også, at de har nogle principper der skal sikre "den gode anvendelse". Her er første princip:

> *We want AGI to empower humanity to maximally flourish in the universe. We don't expect the future to be an unqualified utopia, but we want to maximize the good and minimize the bad, and for AGI to be an amplifier of humanity.*

Det er jo visioner på den helt store klinge: "to be an amplifier for humanity". Men når man så samtidig i medierne kan se, hvordan tech-giganter som Microsoft og Google konkurrerer intenst om at være forrest i kapløbet, kan man godt tvivle på, hvorvidt de fine principper og etiske overvejelser har vægt nok til at holde den konkurrencedrevne udvikling i ave. Microsoft har fx i marts 2023 fyret hele det team, der stod for "responsible AI" dvs. deres etik-afdeling.

Så noget tyder på, at regulering ikke skal komme fra tech-firmaerne men fra politikerne. EU arbejder i øjeblikket på et regelsæt om regulering af kunstig intelligens. Planen er, at der kan laves en aftale om dette i løbet af 2023.

*https://www.computerworld.dk/art/282155/midt-i-kaempe-satsningen-paa-chatgpt-microsoft-fyrer-team-med-ansvar-for-etisk-brug-af-ai https://openai.com/blog/planning-for-agi-and-beyond*

*https://openai.com/blog/planning-for-agi-and-beyond*

## GENEREL KUNSTIG INTELLIGENS FÅR (MÅSKE) IKKE BEVIDSTHED

Generel kunstig intelligens er ikke det samme som en kunstig intelligens med fri vilje og bevidsthed. Man kan i princippet godt forestille sig, at den type algoritmer, der styrer fx ChatGPT, udvides til andre områder. Sandsynligvis vil det, også efter en evt. udvikling af generel kunstig intelligens, stadig være sådan, at en menneskelig bevidsthed og en kunstig intelligens på "indersiden" er to meget forskellige ting. Men hvis en generel kunstig intelligens "på ydersiden" kan meget af det samme som et menneske – fx også vurdere komplekse situationer og træffe beslutninger – så er vi ved at være der, hvor maskiner kan fungere som selvstændige enheder, der træffer egne valg. Også selvom de ikke har en bevidsthed, der kan sammenlignes med den menneskelige.

Så måske skal vi ikke helt afskrive risikoen for, at maskinerne kan vende sig imod os. Det sker måske ikke som en stor sammensværgelse, men som enkeltstående fejlbeslutninger, der kan få alvorlige konsekvenser. Det vil fx næppe være en god ide at sætte en kunstig intelligens til at holde øje med eventuelle atommissilaffyringer fra fjenden – og samtidig give den ret til at iværksætte et modangreb.

Hvis det lykkes at udvikle generel kunstig intelligens, vil det afføde helt nye etiske udfordringer. Men også de specifikke kunstige intelligenser, der allerede findes, nødvendiggør etisk stillingtagen. Fx handler alle de 6 cases i dette forløb om teknologier, der anvender specifik kunstig intelligens.

## "BLACK BOX" OG BIAS PROBLEMATIKKERNE

For de fleste fagfolk ligger de åbenlyse farer ved kunstig intelligens dog andre steder end frygten for kunstige intelligenser, der bliver selvstændige. En af de gennemgående problematikker er den såkaldte "Black box"-problematik. De kunstige intelligenser er langt bedre end mennesker til at behandle gigantiske datasæt. Samtidig er de kendetegnet ved at udvikle sig løbende, altså blive bedre til at behandle data. Det betyder, at det for mennesker er utrolig svært at gennemskue hvilke overvejelser, der ligger til grund for den kunstige intelligens' vurderinger og anbefalinger. Denne manglende indsigt kaldes for "Black box".

Vi kan fx forestille os en kunstig intelligens, der skal vurdere og prioritere kræftbehandlinger. Måske finder den nogle kriterier, der er objektive, men som vi egentlig ikke ønsker skal ligge til grund for en vurdering. Det kunne fx være BMI. Dermed kommer folk med højt BMI (hvilket der kan være mange årsager til) længere nede i behandlingskøen end andre.

Uhensigtsmæssige vurderinger kan også skyldes skævvridninger i datasættene, som den kunstige intelligens bliver fodret med. Det kalder man med en fagterm for bias. Et eksempel kan være en kunstig intelligens, der skal planlægge en national indsats i forhold til hedeslag. Hvis man fodrer den kunstige intelligens med for mange data fra befolkninger i fx Australien, vil den kunstige intelligens sandsynligvis tro, at vi danskere reagerer langt mindre på varme, end det er tilfældet. Det er en bias.

## ETISK VURDERING AF KUNSTIG INTELLIGENS

Vi har valgt ikke at introducere begreberne pligtetik og nytteetik/konsekvensetik i dette forløb. Men hvis eleverne i forvejen har arbejdet med disse begreber, vil det være oplagt at inddrage dem her også.

I vurderingen af ny teknologi, er det ofte **nytteetik**, der fylder mest. Nytteetikken vil altid vælge den løsning, der giver mest lykke og mindst lidelse. Man oplister således de positive og negative konsekvenser. Herefter giver løsningen sig selv, da det i princippet bare er et regnestykke, der måler lykke minus lidelse.

Så simpelt fungerer det bare sjældent i virkeligheden. For det første er det svært at forudse og vurdere, hvor meget lykke og lidelse en given teknologi vil bibringe. For det andet er det tit sådan, at en ny teknologi vil forøge noget godt – men samtidig måske reducere noget andet, som vi også synes er godt. Fx vil øget overvågning kunne skabe større tryghed – men det kan reducere vores frihed. Her har vi en værdikonflikt og et ægte dilemma, hvor to hver for sig gode værdier/principper står i modsætning til hinanden.

**Pligtetik** handler om, at vi altid skal handle ud fra et overordnet princip – og holde fast i det, uanset hvad.

Benspændet fra den før omtalte forskergruppe på AU, der bruges i dette forløb, er pligtetisk. Også selvom der er vide rammer for fortolkning af princippet.

*Robotter/kunstig intelligens skal kun udføre opgaver, som mennesker bør udføre - men som mennesker ikke kan udføre.*

Andre pligtetiske principper i forhold til kunstig intelligens kunne være:

- Vigtige beslutninger skal altid tages af et menneske.
- Det skal altid være tydeligt, at man har med en kunstig intelligens at gøre.

Vurderingsskemaet, som eleverne i fase 3 skal bruge til at vurdere de forskellige teknologicases, er uden etiske fagudtryk og forholdsvis enkelt. Punkterne med vurdering af fordele og ulemper er en nytteetisk tilgang. Men skemaet spørger også efter, om teknologien kan have negative konsekvenser for særlige grupper. Og det er inspireret af pligtetikkens princip om, at et menneske altid skal behandles som et mål i sig selv, aldrig kun som et middel.

Hvis klassen tidligere har arbejdet med etik og kender til fx nytteetik og pligtetik, så er det bare at justere skemaet, så det flugter med, hvordan I tidligere har arbejdet med etiske problematikker.

## DET UERSTATTELIGE

Den danske filosof Peter Kemp har skrevet en bog om teknologi-etik, *Det Uerstattelige.* Tanken om, at det enkelte menneske er uerstatteligt, at det har absolut værdi og en iboende værdighed, mener Peter Kemp, er et godt udgangspunkt, når nye teknologier skal vurderes.

Ifølge Kemp, som blev ideen om individets absolutte værdi og uerstattelighed lanceret i kristendommen. Fx i lignelsen om hyrden der forlader flokken på 100 får, for at gå ud og lede efter det fortabte får. Eller i fortælling om den fortabte søn, der fejres stort, da han vender tilbage.

Hos Kant blev ideen om uerstattelighed løsrevet fra kirkesproget. Det førnævnte krav om, at et menneske altid skal behandles som et mål i sig selv, aldrig kun som middel, handler netop om menneskets uerstattelighed og værdighed som noget, der ikke må krænkes.

I forbindelse med kunstig intelligens er det tankevækkende, at teknologien netop bruges til at erstatte mennesker indenfor konkrete arbejdsopgaver. Men hvad det angår, adskiller kunstig intelligens sig jo ikke fra andre teknologier som gravemaskiner og industrirobotter. Det er bare en anden type opgaver/jobs, som kunstig intelligens kan erstatte.

Alligevel er kunstig intelligens-teknologien noget ganske særligt, fordi den efterligner noget af det, som vi plejer at opfatte som det særligt menneskelige; nemlig vores sprog og vores tænkning. Og hvad med menneskelige følelser og vores evne til at indgå i relationer? Er det uerstatteligt, eller kan kunstig intelligens også erstatte mennesker her? Det kan godt være, at en kunstig intelligens ikke selv har følelser "på indersiden", men når en kunstig intelligens som fx en Replika-chat-ven "på ydersiden" minder så meget om et menneske, at nogle personer udvikler både venskaber og kærlighedsforhold til algoritmen, så er vi måske ved at erstatte noget uerstatteligt med teknologi?