**Individuell innlevering 1**
**IN5480**
**Andrea Ulshagen**

# 1.1 Concepts, definition and history of AI and interaction with AI

First, write a section about how AI came about, the history of AI. When, and by whom, was the term first used?

Ideen om objekter som "våkner til live" som intelligente skapninger har vært aktuell i lang tid, helt fra gammel gresk mytologi har myter om robot eksistert (T. Lewis, 2014). Historier om, Talos, en stor bronserobot, Pandora, en kunstig dame og andre kunstige, menneskelige karakterer har gått igjen i legender i 2,700 år (A. Shashkevich, 2019). Det er en veldig spennende tanke at vi mennesker i så lang tid har fantasert om slike "skapninger" og at vi fremdeles ønsker å lage ting som oppfører seg som mennesker.

Feltet AI ble formelt etablert i 1956 av John McCarthy, på en konferanse på Dartmouth College, i Hanover, og regnes som grunnleggeren av disiplinen (Council of Europe). Tross positiviteten og forventningene til AI og dets fremtid, viste AI seg vanskelig å oppnå og epoker med "AI vinter" fulgte (T. Lewis, 2014). mye grunnet til at maskinene hadde lite minne (Council of Europe). Det var først når mikroprosessoren ble til, på slutten av 70 tallet at AI igjen fikk en ny boost (Council of Europe). Etter en ny "vinter" la store mengder tilgjengelig data og oppdagelsen av effektive computer graphic card processors til rette for en ny boom i i 2010.

Then, find three different definitions of AI. Describe and explain these three definitions, for example by when it was defined, by whom and in what community. Based on these three definitions, make one definition yourself - and describe and explain your definition.

Av John McCarthy og konferansen på Dartmouth College, som beskrevet tidligere, ble feltet beskrevet slik: *"The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."* (B. Marr, 2018).

En andre definisjon, publisert i en artikkel i fair observer om Elon Musk og hans holdning til AI, lyder som følger: *"AGI is a single intelligence or algorithm that can learn multiple tasks and exhibits positive transfer when doing so, sometimes called meta-learning, leading to recursive self-improvement."* (P. Isackson, 2019). Denne definisjonen baserer seg på rekursiv læring hvor maskinen tilegner seg data, følger mønstre, analyserer, lærer og tilegner mønsteret logikk (P. Isackson, 2019). Men med ubegrenset data kan maskinen i teorien ved hjelp av ny tilegnet logikk og "sannheter" bevege seg vekk fra den originale menneskeskapte algoritmen (P. Isackson, 2019). Dette kalles i noen sammenhenger for "strong AI" (P. Isackson, 2019). Denne definisjonen beskriver AI i større grad som en black box hvor det er vanskelig å finne ut hvordan maskinen er kommet dit den er (P. Isackson, 2019).

Alan Turing's, en kjent matematiker fra 1900-tallet, definisjon lyder som følger : "A computer can be said to possess artificial intelligence if it can mimic human responses under specific conditions." (P. Isackson, 2019) Slike definisjoner kalles ofte "svak AI"(P. Isackson, 2019).

Om vi ser på første og tredje definisjon kan vi se at disse er i større grad relatert til simulering, herming og etterligning av menneskelig oppførsel og intelligens heller enn, slik som det er forklart i definisjon to, en maskin med reelle menneskelig intelligens som tilegner seg mønstre og sannheter avhengig av den originale algoritmen. Slik jeg ser det er det et tydelig skille mellom Turing ogMcCarty sin definisjon og black box definisjon; En simulering av menneskelig intelligens vil ikke på samme måte ha sjulte logiske sluttninger som en "sterk AI" som beskrives i definisjon to. Det kan se ut til at definisjonen er skiftet med tiden og teknologiens muligheter.

Find one contemporary company that works with AI and describe how this company presents AI on their web pages. In what way does this company talk about AI, as a product, as a service, framework or "idea"?

IBM Cloud har en AI de kaller Watson, og i følge nettsiden deres har en nesten ubegrensede muligheter for hva en kan få til med Watson, en skikkelig superhuman virker det som.
Jeg siterer: "Unlock the power of AI with IBM Watson", "Turn unstructured data into intelligence and competitive advantages", "Improve decisions based on real-time trends",
"Quickly and securely build truly cognitive applications", "Oh, the things you can do with Watson on the IBM Cloud". De snakker om AI som en tjeneste i stor grad. Hentet fra:
https://www.ibm.com/uk-en/cloud/ai

Select one documentary or a fictional film, book or game that is about the use and interaction with AI. Describe with your own word how human interaction with AI is portrayed in this work.

*Upgrade*

En selvkjørende bil kolliderer og etterlater hovedpersonen paralysert av skadene. Han får så en chip implantert i hodet, som lar han kontrollere kroppen sin. Måten denne AIen interagerer med hovedkarakteren er da gjennom stemmer i hodet. AIen kalles STEM og kan ta kontroll over kroppen hans når han ønsker det for å gjøre han rask og sterk. Etterhvert blir denne AIen oppdatert, hacket og tar til slutt kontroll over hele kroppen til hovedkarakteren. Litt som venom i tech i steden for aliens.

## 1.2 Robots and AI systems

First, write a section about how the word Robot came about.

Karel Capek, en Tsjekkisk skuespillforfatter, introduserte ordet robot i et skuespill i 1920. Ordet er hentet fra ordet robota som betyr slaveri eller tvangsarbeid. Capek hentet inspirasjon fra blant annet Frankenstein når han skrev stykket som omhandlet roboter som gjør jobber som ingen mennesker vil gjøre. I følge Capek manglet disse robotene ingenting annet enn en sjel (Science Friday, 2011).

Then, find two different definitions of "robot". Describe and explain these definitions. Based on these definitions, make one definition yourself, and describe and explain this definition.

*"Robot, en datastyrt enhet som ved hjelp av sensorer kan motta data fra omgivelsene, bearbeide disse og reagere ved å iverksette handlinger i henhold til forhåndsprogrammerte regler."* (I. M. Liseter, SNL, 2018)

*"I would say that a robot is a physically embodied artificially intelligent agent that can take actions that have effects on the physical world," (M. Simon, 2017)*

Jeg vil påstå at begge disse definisjonene baserer seg på at roboten klarer å sanse omgivelsene og ta avgjørelser hva gjelder handlinger og respons basert på den bearbeidede informasjonen. I den første definisjonen vil jeg påstå heller mer mot forhåndsprogramerte responser til gitte situasjoner, mens definisjon nr to kanskje heller mer mot det "å tenke sjæl". Dette vil si at vi utelukker alle mekaniske objekter som ikke klarer å sanse omverdenen.

Jeg vil definere roboter som programmerte maskiner som kan handle uavhengig men basert på innebygd minne.

Discuss the relation between AI and Robots. Is "a robot" different from "an AI"? In what ways are they different and similar? Bring in the definitions that you described earlier about robots and AI for this discussion.

Slik jeg ser det vil en robot bare kunne handle på innebygd minne, og endringer eller lærdommer om du vil kan bare endres dersom du en gjør endringer direkte i koden eller systemet. På den andre siden vil en AI ikke kun ha tilgang til det satte minnet og det satte handlingsmønsteret for men menneskelignende intelligens skal den kunne tilegne seg egen kunnskap og selv utvikle seg på denne måten (E. Oberoi, 2019).

Find one contemporary physical robot, either described in a research article - or a commercial robot, and describe how this robot moves and how a human user is interacting and using the robot in a specific situation.

*Roboten Sophia*

Den store stjernen Sophia, den første roboten som har fått et statsborgerskap. I Dubai. Er laget på en slik måte at den skal se ut som en menneske, en humanoid, med menneskelig hud skal den kunne ha totalt ulike ansiktsuttrykk, dog er de litt slappe. Den kan "se" med kameraer i øynene. Den kan holde øyekontakt, kjenne igjen ulike mennesker, har tale og kan gå. Den fungerer som en chatbot og kan svare på forhåndsprogramerte spørsmål med forhåndsprogrammerte svar.Tanken bak roboten er at den skal kunne brukes innen helse, kundeservice, terapi og utdanning, foreløpig er ikke ansiktsutrykkene veldig terapautiske, men det er tanken. Foreløpig er den mest for show og det har vært diskutert om dette er AI eller kun en robot. Basert på tidligere definisjoner i denne oppgaven vil jeg konkludere med robot.

## 1.3 Universal Design and AI systems

Please find and describe a definition of Universal Design. Explain this definition, how you understand what Universal Design is about with respect to inclusion.

*"Universell utforming handlar om å utforme omgjevnadane slik at vi tek omsyn til variasjonen i funksjonsevne hos innbyggarane, inkludert personar med nedsett funksjonsevne. Når du lagar noko som er universelt utforma, når du alle målgruppene gjennom éi og same løysing." (DIFI)*

Universell utforming handler om å inkludere alle mennesker i samfunnet og på en naturlig måte. Alle mennesker skal ha en følelse av selvfølgelighet når de befinner seg i et miljø. Omgivelsene skal altså ikke være en påminner om nedsatt funksjonsevne men en invitasjon til alle innbyggere. Når du bygger en trapp skal det være en naturlig del av designet som skal være tilrettelagt for rullestolbrukere og det samme gjelder med IT systemer.

Describe the potential of AI with respect to human perception, human movement and human cognition/emotions. You are encouraged to use examples.

Jeg vil tro at man i stor grad kan herme etter menneskelige følelser og lære opp maskinen til å reagere "riktig" på ulike hendelser. Men måten vi føler på, som grunner i persepsjonen og sansene vi får inn, skapes av ulike signaler i hjernen som skiller ut forskjellige stoffer i hjernen, enten som dopamin som gir en brennende følelse, eller adrenalin som gjør oss på alerten. En AI kan lære seg å reagere på en gitt måte basert på sansene de får fra omverdenen, men den vil ikke kunne ha det samme følelsesregisteret. Følelsene våre er jo egt bare hjernen vår som påvirker oss til å ville gjøre mer av noe og mindre av noe annet og kanskje rømme om det er noe farlig. En robot kan "lære seg" å føle, men ikke føle. Tror og håper jeg.

Describe the potential of AI for including and excluding people. You are encouraged to use examples.

Som det har vært snakket om i forelesning så er AI også svak for bias og skjevheter i dataen den får inn og analyserer. Men lignende menneskelig intelligens i form av metalæring og rekursiv selvforbedring vil en AI kunne gå vekk fra den opprinnelige algoritmen og tilegne seg nye "sannheter". Dette kan resultere i at den lager seg uante mønstre som gjør at eksempelvis noen kulturelle grupper ikke får boliglån. Med tilgang til ubegrenset data kan vi også kanskje se at AIen får menneskelige feil og opererer på bias vi mennesker besitter og overfører til AIen. Dersom den i tillegg opererer som en black box er det vanskelig å få svar på hvorfor den trakk den sluttningen den gjorde. Dette kan skape inkludering og skiller dersom dette er gjort på feil grunnlag, men uekte sannheter og bias som grunnlag.

In WCAG 2.1 principles and in the Human AI-Interaction guidelines the concept “understand” and “understanding” is used. Explain briefly in what way you make sense of the concept “understand” and “understanding”. Then address the question: Do machines understand?

I utgangspunktet tenker jeg at forståelsen til mennesker å AI er vidt forskjellig. Men når jeg begynner å legge frem definisjoner blir det vanskeligere. Vi mennesker tar til oss inntrykk fra omgivelsene våre gjennom sansene, så er det persepsjonen vår som setter disse inntrykkene sammen til en forståelse av verden basert på mønstre, tidligere erfaringer og kognitive rammeverk. Når vi snakker om hvordan AI prosesserer informasjon sier vi også har de tilegner seg data følger mønstre analyserer, lærer og dermed tilegner omgivelsene noe form for logikk. Basert på dette burde en maskin også kunne skape seg et “bilde av verden” og omgivelsene dersom den han et rammeverk å sette det inn i forhold til. Jeg er fremdeles ikke komfortabel med å kalle det en AI gjør å forstå verden får den tar utgangspunkt i allerede satte algoritmer. Men jeg synes også det er vanskelig å skulle diskutere for forskjellen ettersom vi også ser og opplever verden basert på tidligere erfaringer og kognitive rammeverk.

## 1.4 Guideline for Human-AI interaction

Please select one of the 18 guidelines from Microsoft, and describe this guideline with a different example than what is given by Microsoft. Search, and find one set of HCI design guidelines. Discuss briefly similarities and differences between the HCI design guidelines and the Human-AI interaction guidelines.

Jeg ser en del likheter mellom HCI design guidelines og Human-AI interaction guidelines. HCI designprinsippet visibility handler om at viktig funksjonalitet skal komme klart frem i grensesnittet fordi usynlige og automatiske funksjoner kan være forvirrende. Human-AI interaction guidelines sier at man skal “make clear what the system can do, help users understand what the AI system is capable of doing”. I tillegg legger sistnevnte også til at man må “make clear how well the system can do what it can do” Fellestrekket her handler om å vise brukeren hvordan systemet fungerer.

Consistency og affordance er to andre HCI designprinsippet som vi kan se en sammenheng til med Human-AI interaction guidelines hvor “match relevant social norms. Ensure the experience is delivered in a way that users would expect given their social and cultural context.” Samtlige omhandler at man ta i betrakting brukerens erfaringer, kunnskaper og sosiale og kulturelle kontekster når man designer et system. Feedback er også et viktig aspekt i begge guidelines, hvor det å gi brukeren en tilbakemelding på hva og hvorfor systemet gjorde som det gjorde.

Den store forskjeller mellom de to guidelinene slik jeg ser er at Human-AI interaction guidelines handler mye om fremtiden og hvordan man skal håndtere et feilende system, lære for fremtiden og informere brukeren om disse lærdommene og endringene. Ulike scenarioer de er laget for men i bunn og grunn veldig likt.

**Individuell innlevering 1. interasjon 2**
**IN5480**
**Endringer etter feedback**

**Ser du på "AGI" som det samme som "AI"?**
Jeg ser på AI som noe som er preprogrammert til å utføre en "menneskelig" oppgave. AGI ser jeg på som noe som forventes at kan være like smart som mennesker.

**Jeg fikk litt inntrykk av at du ser på mønstergjenkjenning som en nødvendig del av AI. Tror du det kan finnes AI som ikke er basert på mønstergjenkjenning?**
Slik jeg ser det, må en ai meste noen form for mønstergjenkjenning ettersom de skal kunne sanse omgivelsene. Sansene våre blir satt sammen til noe meningsfullt ved hjelp av kognisjon vår. Dersom dette ikke er til stede vil jeg argumentere for at det ikke er en menneske-spesifikk oppgave.

**Du siterer M. Simon, 2017: "I would say that a robot is a physically embodied artificially intelligent agent that can take actions that have effects on the physical world" og du skriver så: "Jeg vil påstå at begge disse definisjonene baserer seg på at roboten klarer å sanse" Tolker du da sansing som en del av begrepet "physically embodied", eller kan noe være "physically embodied" uten å kunne sanse?**
Spennende tanker, for å svare på dette ønsker jeg å reflektere noe ytterligere. For at effekten den skal kunne ha på den ytre verden er basert på at roboten leser og responderer til sine omgivelser må den kunne sanse. Men en robot kan også være kun tangible og bare operere i den fysiske verden, slik som ferbi.

**Spennende å høre hvordan du definerer "Robot"! Jeg ble interessert i hva du legger i "uavhengig" i din definisjon. Tenker du uavhengig av en kontrollør?**
Ja, at den kan handle uten at det er noen som forteller den at den skal gjøre akkurat dette. Enter på forhånd som en del av koden eller ved en kontroller.

**Du inkluderte ikke noe element av sansing i robot definisjonen din om jeg forstod riktig. For meg hadde det vært spennende å høre hva du tenkte rundt det.**
Man kan kanskje skille mellom maskin og roboter. Maskiner er typ robothundene (slike definisjoner er kanskje det som gjør det vanskelig å bruke rett vokabular) en hadde når man var liten, som gikk og bjeffet men hadde null styring på hvordan omverdenen var og kunne ikke gjøre noen endringer i forhold til dette. En robot vil da gjerne kunne kjenne at den har veltet eller krasjet og derfor slutte å gå.

**Og hvor åpent er kravet om å kunne "handle"? Kan det å summere to tall i prosessoren til systemet være en tilstrekkelig "handling"?**
Ja, det vil jeg si.

**Vil din definisjon av roboter f.eks. definere en røykvarsler som en robot?**
- **En optisk røykvarsler er programmert**
    - **(til å reagere etter en viss mengde lys blir forstyrret, og si fra om det er svakt batteri)**
- **Den er en maskin**

- **(eller hva legger du i "maskin"?)**
  - **snl: Maskin, apparat som med tilført energi utfører et visst arbeid.**

- **Den handler uavhengig av en kontrollør**
  - **(etter at den har fått et batteri. Men det er vel tilfellet for de fleste kjente roboter også)**
- **Den har en form for innebygd minne**
  - **(som sier hvor mye lyset kan forstyrres før alarmen skal gå) Jeg tar opp røykvarsler som eksempel siden det er en ting jeg ikke tenker på som en robot, men likevel passer den med hvordan jeg forstår din definisjon av en robot. Det kan da kanskje bidra til å utfordre eller tydeliggjøre din definisjon av en robot, hvis du mener at røykvarslere ikke burde være i kategorien "robot". Men det er kanskje ikke noe grunn til at en røykvarsler ikke skal kunne kategoriseres som roboter?**

Spennende diskusjon du tar opp her. Jeg mener skillelinjen mellom robot og maskin er veldig vag. Men jeg står på min definisjon om at en robot klarer å tolke omgivelsene sine til en viss grad og opererer i den reelle verden. Man kan kanskje trekke skillelinjen mellom hvaslags handlinger. At mer kompliserte handlinger tilhører en robot mens mindre avanserte, som røykvarsleren, som kun har en funksjon (å måle lyset) Forblir en maskin.

Takk for spennende og interessant tilbakemelding.

**Individuell innlevering 2**
**IN5480**

# 1.a.

Amershi et. al. (2019) adresserer tre ulike karakteristikker ved "AI-infused systems". Den første er "uncertainty"; Amershi et. al. påstår at ikke bare er feil vanlig, men at feil-forebygging er vanskelig å oppnå. Neste er "inconsistency"; Amershi et al mener at "AI-infused system" ofte bryter med etablerte usability guidelines, eksempelvis "consistency" ettersom systemet lærer og utvikler seg over tid. Faktum er at systemet kan oppføre seg ulikt over tid og fra bruker til bruker (Amershi et. al., 2019). Den siste karakteristikken er "Behind the scenes personalization", karakteristikker som opaque/black box/lite transparent gjør at mange handlinger skjer "behind the scenes".

Kocielnik et. al. (2019) adresserer tre ytterligere karakteristikker ved "AI-infused systems". Den første er "probabilistic" – "underlying algorithms driving AI functionalities, such as natural language understanding, sensorbased inferences, web behavior prediction, or object recognition in video or images, are probabilistic and almost always operate at less than perfect accuracy." (Kocielnik et. al., 2019). Videre karakteristikker er "Impacted by user actions" og "Transparency issues". Begge i samme gate som karakteristikkene til Amershi et. al. og dermed adressert ovenfor.

Karakteristikker adressert av A. Følstad (2020) er følgende; Learning, Improving, Black box, Fuelled by large data sets. "Computer systems [are] learning and improving on the basis of large data sources" (A. Følstad 2020). Stadig lærende betyr også dynamiske og improving betyr også at feil er uunngåelig. Disse karakteristikkene adresserer også Amershi et. al. og Kocielnik et. al. poengterer at "users don't expect apps to act inconsistently or imperfect [...] leads to disappointment and abandonment." Black box og opaque er også blitt adressert. Relatert i Data gathering through interaction så poengterer Kocielnik et. al. følgende: "Prior work has shown that letting users contribute to a system's behavior may make them more accepting of the systems mistake." Dette er et argument for å bevisst lukke gapet mellom brukerens forventninger til systemet og deres faktiske opplevelse. "User satisfaction and acceptance of the system is directly related to the difference between initial expectations and their actual experience." (Kocielnik et. al., 2019). For å få brukeren til å bruke nok tid på systemet til å forstå og akseptere feil som vil forekomme må brukerens forventninger og faktiske opplevelse være like nok til at de ikke forlater systemet.

# 1.b.

Jeg ønsker å trekke frem to eksempler basert på karakteristikken "fuelled by large data sets".

Den første er en "art agent" kalt Creative Adversarial Networks (CANs). CANs viser eksempler på at AI mestrer egenskaper man har tenkt var reservert for mennesker, nemlig kreativitet. Det er nylig blitt gjennomført et studie som tok utgangspunkt i en teknologi som prøver å male så realistiske bilder som mulig. Teknologien kompilerer en rekke algoritmer hvor det ene nettverket genererer ideer og den andre dømmer resultatet. Algoritmen looper frem og tilbake til en decent resultat er nådd.

Denne AIen ble matet med over 81,000 malerier fra WikiArt databasen fra det 15. til 20. Århundre. Resultatet viste at mennesker kunne ikke skille mellom kunsten som var skapt av algoritmen og

contemporary kunstneren.

Og publikumet ranket algoritmens kunst høyere enn kunsten laget av mennesker. Da kan man jo spørre seg selv om kreativitet eksklusivt en menneskelig kvalitet.

Neste eksempel jeg vil ta frem er en AI som fungerer som designer. Videoen forklarer bedre enn jeg kan, så se videoen her. AI kan booste våre analytiske evner og våre evner til besluttningstaking ved å gi oss riktig informasjon til riktig tid. De kan også da lett hjelpe oss å booste kreativiteten.

Dette er da to eksempler på at AI samler inn store datasett og bruker dette som sitt grunnlag. Her er det ikke meningen at AIen skal være en uperfekt erstatter til menneskets intelligens og handlekraft. Det er rett og slett ment som en ekstensjon av vår egen kognitiv og kreativitet. En bidragsyter heller enn en erstatter og feil vil ikke være en så stor påkjenning på mennesket fordi vi ikke legger all tillit i systemet men heller utnytter det til å prosessere disse store datasettene og komme med en haug ulike forslag på mye kortere tid enn vi hadde klart.

## 2.

Jeg velger eksempel nr. 2 for å besvare denne oppgaven.

*G11 - Make clear why the system did what it did.* Enable the user to access an explanation of why the AI system behaved as it did. At systemet ikke har en “black box” slik at all tidligere design den har hentet inspirasjon fra kommer tydelig frem. Det bør altså mulig å gå tilbake å se hvorfor AIen kom frem til den konklusjonen som den gjorde.

*G4 - Show contextually relevant information.* Display information relevant to the user’s current task and environment. AIen’s eneste oppgave er å svare på informasjon i forhold til brukerens ønsket oppgave og miljøet den skal være i. Om det er en bro eller en pil og bue, det er altså viktig for brukbarheten og nyttigheten til systemet at den svarer til skopet brukeren ønsker og henter inspirasjon fra lignende “ting”.

# 3.

*G1 - Make clear what the system can do. Help the user understand what the AI system is capable of doing.*

I artikkelen til E. Luger og A. Sellen (2016) har en av deltakerene i undersøkelsen deres følgende utsagn: “I think by playing with it and understanding what it could do well and what it could do badly….through that I found that I developed a series of things that I used it for that were a quite discrete number of its functions” og forfattere utdyper “In this way, users’ satisfaction with, and trust in, the product had a strong relationship to the extent to which they were prepared to invest time in both understanding what their CA could do, and practicing those interactions.” Med tid vil erfaringer gi brukerne et inntrykk av hva systemet kan gjøre, og hvilke oppgaver det faktisk kan hjelpe med. Men dette krever som nevnt at brukerne investerer tid til en tosidig læringsprosess.

Det å gi brukeren en pekepinn på hva systemet klarer kan være en hjelp på veien, som kanskje gjør at de ikke forlater systemet før de har gitt det et forsøk. Som nevnt tidligere, å minimere gapet mellom forventninger og faktisk opplevelse gjør det mindre sannsynlig at brukeren forlater systemet før de får sjansen til å lære om det. Mange chatboter gir brukeren en indikasjon på hva de kan svare på og at de er under opplæring.

Et annet eksempel fra samme artikkel “Asides from the two most frequent users who tended to be more experimental and forgiving, all of those interviewed raised issues of trust as limiting the tasks they would ask their CA to perform.” Det er et karakteristikk ved AI at de gjør mye feil. Så det er bedre med en tilgivende adferd der brukeren er klar over at alt kanskje ikke er perfekt. Igjen snakker vi om forventningsavklaring. Dette bringer oss over til neste guideline.

*G2 - Make clear how well the system can do what it can do. Help the user understand how often the AI system may make mistakes.*

De to går litt over i hverandre. Men det å være ærlig på muligheter og begrensninger gjør det lettere for brukeren å forvente det “riktige” fra systemet, som vi nå har etablert at bidrar til brukervennligheten.

I følge Kocielnik et. al. (2019): “transparency techniques are intended for post fact explanations about the decisions of an AI or ML system, rather than for adjusting user expectations of an AI-powered system prior to its use.” Om man hadde snudd om på dette hadde man kanskje klart i større grad å imøtekomme tidligere nevnte guidelines.

**Individuell innlevering 3**
**IN5480**

# 1.

*Philips at al. (2016) give a taxonomy and examples of human-robots collaboration. Choose 2- 3 examples.*

## 1.1 Human-Animal Teams: Physical Benefits

Dette eksempelet i Philips at al. (2016) beskriver hvordan dyr har fungert som en eksternalisert fysisk evne til mennesker og gitt fordeler i teamets kapasitet ved både å erstatte, multiplisere eller øke menneskers fysiske evne. Eksempler på dette kan være mer effektiv transport, frakt av varer, løfting og trekking av tyngre objekter eller ting som krever større utholdenhet. Disse teamene har ofte gitt et fordel til domener som produksjon og industri.

## 1.2 Human-Animal Teams: Emotional Benefits

Dyr, i stor del kjæledyr, kan bistå mennesker og gi en komfort i form av en følgesvenn. Philips at al. (2016) skriver også at dyr kan øke mennesker følelsesmessige kompetanse og minske følelsen av ensomhet. Philips at al. (2016) påstår også at "søte" dyr skaper et raskere bånd av tillitt og kjærlighet. Dette har blitt overført til sosio-emosjonelle roboter som mimikerer søte dyr. Eksempelvis "Paro, the small seal" som skal minske depresjon ved å fungere som en følgesvenn.

Dyr brukes også til å lære sosiale skills og bli kvitt angst og dysleksi. Når karakteristikkene til et dyr som kan bidra til et "team" på denne måten blir overført til en robot er det mulig at dette har de samme effektene på mennesker.

# 2.

*Describe their levels of autonomy as described in Shneiderman (2020) and reflect on advantages and disadvantages if we decrease/increase their current level of autonomy*

Shneiderman (2020) diskuterer ulike grader av "computer automation" og påstår at godt designet teknologi tilbyr høy grad av "human control" og høy grad av "computer automation". Shneiderman (2020) påstår at kun da kan de bidra til å bedre menneskelig prestasjon, sørge for økt adapsjon og gjøre maskinen tillitsverdig, pålitelig og trygg.

Sheridan and Verplank's (1978) presenterer ti nivåer av "human control" og "computer automation.

Jeg har valgt å ta for meg to ulike eksempler fra Philips at al. (2016); Provide comfort og maskinen Paro og Multiply physical capabilities, industrielle roboter. Jeg vil påstå at Paro befinner seg på level 9 i Sheridan and Verplank's (1978) nivåer: "The computer informs the human only if it, the computer, decides to". Jeg vil påstå at industrielle roboter befinner seg på en blanding av 6 og 7: "The computer executes automatically, then necessarily informs the human and the computer allows the human restricted time to veto before automatic execution".

Hva gjelder figuren Shneiderman (2020) presenterer vil jeg vil påstå at Paro befinner seg et sted mellom de to rutene helt til høyre, da den i liten grad lar seg styre av mennesker. Den fungerer som

den gjør, oppsøker interaksjon når den gjør og ber brukeren lade den om det er nødvendig. Altså ikke full pott på tillitsverdig, pålitelig og trygg. Dersom den i større grad hadde åpnet for menneskelig kontroll ville den nådd opp til øverste rute til høyre. Men dersom den skulle blitt styrt av mennesker hadde den ikke lenger tilbudt det samme som den gjør i dag. Jeg vil påstå at bruksområde til maskinen har mye å si, en kan ikke svart hvitt si utifra en slik figur om det er en god og brukervennlig maskin, man må også se ann bruksområdet.

Industrielle roboter derimot, vil jeg påstå befinner seg øverst til høyre. Med en stor grad av menneskelig kontroll men også stor grad av autonomitet. Dersom man hadde satt den autonomiteten hadde den ikke på samme måte kunne erstatte menneskelig evne slik som den gjør i dag, men kanskje multiplisere eller øke dersom den gjør ting raskere eller løfter tyngre.

## 3.

*Reflect on their current and needed explainability (Hagras, 2018; Smith-Renner et al. 2020).*

Jeg ønsker å påpeke at diskusjonen rundt disse systemene er vanskelig uten konkrete eksempler og god nok innsikt i disse. Jeg har gjort meg opp en ide om hvordan de ulike systemene fungerer og forsøkt å diskutere ut fra dette.

Ettersom industrielle roboter i stor grad er forhåndsprogrammerte og tilbyr brukeren mulighet til "veto before automatic execution" er gjennomsiktigheten god her. Dette vil selvfølgelig variere, men her tar jeg utgangspunkt i mitt eget bilde av slike maskiner. Det skal sies at det kanskje bare er noen mennesker som kan forstå formatet og språket gjennomsiktigheten vises gjennom men dette vil i så fall kunne læres for de som er nye i jobben.

Bias som ligger i slike maskiner, basert på forhåndsprogrammering og innstillinger vil i så fall ha grunn i brukerens bilde av verden. Og en kan være sikker på at den ikke har introdusert noen egen form for bias her, da de ikke er laget for å lære, men kun gjennomføre. Her er det en forskjell på maskiner og AI infused systems.

Jeg mener den menneskelige kontroller også spiller inn på causality, fairness og safety.

Dersom jeg skulle kommentert på noe som kunne vært forklart bedre så er det nok for de menneskene som ikke har programmert maskinen og ikke er godt kjent med hvordan den fungerer. Det kan være vanskelig å sette seg inn i en ny maskin og man må gjerne ha noen erfarne som kan være der og lære de nye hvilke beskjeder som betyr hva og hvordan man setter inn innstillinger på den.
Det er et viktig poeng at valgene maskinen tar ikke er selvsagt for alle. Det vil være nødvendig med feedback for å vite at riktig valg er tatt, men også er forklaring for å vite at riktig valg er tatt basert på riktig avgjørelse (Smith-Renner et al., 2020).

For å få en litt interessant diskusjon her så ønsker å se for meg at Paro er en ai basert sel, ikke bare en robot. La oss for diskusjonens skyld si at den kan reagere på hvordan mennesker behandler den.

Ettersom dette er en sel som ikke har tale eller noen skjerm som kan si noe om hvorfor den tar de avgjørelsene den tar. Dersom den begynner å nynne eller ule så er det vanskelig å vite hva som forårsaket dette, annet enn at man kan regne med at den vil ha de samme karakteristikkene som et dyr og ule om den er sulten og nynne når den blir kost på.

Dersom selen lærer hvordan den skal interagere med brukeren og brukeren oppfører seg hårreisende kan den lære seg til andre måter å reagere på handlinger fra mennesker enn det vanlige dyr ville gjort og kunnskapen er kanskje ikke overførbar til nye brukere.

Denne manglende gjennomsiktigheten og mulighet for biasen kan gå igjen i kausaliteten. De modellene brukeren lærer seg for å interagere med et dyr kan være lite overførbart til virkeligheten. På bakgrunn av disse to "manglene" kan man ikke være sikker på at de modellene man har lært fra AI er overførbar til andre situasjoner da vi ikke får undersøkt fenomenene som ligger til grunn for interaksjonen med selen.

# Kilder

Council of Europe, "History of Artificial Intelligence". Hentet fra: https://www.coe.int/en/web/artificial-intelligence/history-of-ai (09/09/2020)

Digitaliseringsdirektoratet, "Kva er universell utforming?". Hentet fra: https://uu.difi.no/kva-er-universell-utforming (10/09/2020)

P. Isackson, (Feb 27, 2019), "AI and the New Dimensions of Hyperreality". Hentet fra: https://www.fairobserver.com/region/north_america/ai-elon-musk-artificial-intelligence-open-ai-business-news-today-34802/ (09/09/2020)

Tanya Lewis, (December 04, 2014), "A Brief History of Artificial Intelligence". Hentet fra: https://www.livescience.com/49007-history-of-artificial-intelligence.html (09/09/2020)

I. M. Liseter, SNL, (20. februar 2018) "robot". Hentet fra: https://snl.no/robot (10/09/2020)

B. Marr, (Feb 14, 2018), "The Key Definitions Of Artificial Intelligence (AI) That Explain Its Importance". Hentet fra: https://www.forbes.com/sites/bernardmarr/2018/02/14/the-key-definitions-of-artificial-intelligence-ai-that-explain-its-importance/#13f5ea364f5d (09/09/2020)

E. Oberoi, (2019-10-14), "Difference between robotics and artificial intelligence". Hentet fra: https://www.skyfilabs.com/blog/difference-between-robotics-and-artificial-intelligence (10/09/2020)

Science Friday, (04/22/2011), "The Origin Of The Word 'Robot'". Hentet fra: https://www.sciencefriday.com/segments/the-origin-of-the-word-robot/ (10/09/2020)

Alex Shashkevich, (Februar 28, 2019), "Stanford researcher examines earliest concepts of artificial intelligence, robots in ancient myths". Hentet fra: https://news.stanford.edu/2019/02/28/ancient-myths-reveal-early-fantasies-artificial-life/ (09/09/2020)

M. Simon, (08.24.2017), "What Is a Robot?". Hentet fra: https://www.wired.com/story/what-is-a-robot/ (10/09/2020)

Orge Castellano, 3. Sep. 2018. “Will the Next Picasso Be a Robot?”. Hentet fra: https://medium.com/s/story/will-the-next-picasso-be-a-robot-9438482b4208 (13.10.20)

A. Følstad, 22. sept. 2020, Forelesningsfoiler - “Interacting with AI”. Hentet fra: https://www.uio.no/studier/emner/matnat/ifi/IN5480/h20/Undervisningsmateriale/interacting-with-ai-2020---module-2---session-1---handouts.pdf (13.10.2020)

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil og Eric Horvitz, 2019. “Guidelines for Human-AI Interaction”. Hentet fra: https://www.microsoft.com/en-us/research/uploads/prod/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf

Rafal Kocielnik, Saleema Amershi og Paul N. Bennett, 2019. “Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems”. Hentet fra: https://www.microsoft.com/en-us/research/uploads/prod/2019/01/chi19_kocielnik_et_al.pdf (13.10.2020)

Ewa Luger og Abigail Sellen, 2016. “Like Having a Really bad PA”: The Gulf between User Expectation and Experience of Conversational Agents”. Hentet fra: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/08/p5286-luger.pdf (13.10.20)

Hagras, H., Toward Human-Understandable, Explainable AI, Computer, 51, 9, 2018, 28- 36 https://ieeexplore.ieee.org/document/8481251

Phillips, E., Ososky, S., Swigert, B. and Jentsch, F. Human-animal teams as an analog for future human-robot teams, Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol 56, Issue 1, (2016) pp. 1553 – 1557 DOI: https://doi.org/10.1177/1071181312561309

Shneiderman, B., Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy, arXiv.org (February 23, 2020). https://arxiv.org/abs/2002.04087v1 (Extract from forthcoming book by the same title)

Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D.S., and Findlater. L. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. DOI: https://doi.org/10.1145/3313831.3376624