Modul 1 – individuell innlevering

Konsepter, definisjoner og historien til interaksjon med AI og autonome systemer

Det var Dartmouth Research Project som først definerte AI, den gang som «få en maskin til å oppføre seg på en måte som ville blitt kalt intelligent dersom et menneske hadde gjort det». De ønsket å gjennomføre en studie hvor de ønsket å få en maskin til å simulere intelligens ved å gi maskinen en så nøyaktig forklaring som mulig. De ønsket å gi maskiner en slags «tilfeldighet» eller originalitet, men heller ikke bare ren tilfeldighet, men tilfeldighet basert på tidligere resultater, noe lignende det mennesker ville kalt en «hunch» (McCarthy et al., 1955).
Dette var derimot ikke første gangen noen tenkte på kunstig intelligens. Alan Turing blir ofte sett på som den første til å ta opp computere og dens evner til å gjøre oppgaver normalt gjort av mennesker (Grudin, 2009).

Definisjon:

1. "AI – defined as a systems ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan & Haenlein, 2019).
2. That of making a machine behave in ways that would be called intelligent if a human were so behaving (McCarthy et al., 1955).
3. Artificial intelligence leverages computers and machines to mimic the problem-solving and decision-making capabilities of the human mind (IBM, u.å.).
4. Kunstig intelligens er en datamaskin som er avansert nok til å lære av omverden og tilpasse seg deretter.
    a. Dette vil si at i motsetning til tradisjonelle datamaskiner vil kunstig intelligens lære av hvordan den blir brukt og forbedre seg.

Er autonomi, autonomi eller er det forskjell på dette? Mennesker er autonome, og denne autonomien er det som gjør at vi kan ta våre egne valg og fungere som individer, men har AI den samme kapasiteten til å ta egne valg? Vi bestemmer over våre egne handlinger, vi kan ta avgjørelser basert på akkurat det vi vil, og vi kan ta konsekvensene av disse handlingene og avgjørelsene. Dette er vanskeligere å få til hos AI fordi AI ikke har den samme kapasiteten til å ta avgjørelser. Totsching (2020) skriver at AI ikke kan få den samme autonomien som mennesker har fordi den mangler evnen til å endre målet sitt. Måten AI'en når målet sitt kan være uforutsigbar, men jeg tror ikke man kan sammenligne autonomi hos mennesker og AI, i alle fall ikke helt enda.

Automatisk vs autonomt. Automatisk betyr at noe gjøres uten at noen utenfor tar aktiv del i oppgaven, men oppgavene som skal gjøres er tydelig definert på forhånd, for eksempel en vaskemaskin.
Til motsetning betyr autonomt at noe ikke blir kontrollert av noen andre, det er uavhengig. Et autonomt system lærer av omgivelsene sine og vil endre seg deretter.

HRI

Denne artikkelen handler om Human Robot Interaction (HRI) og hvordan dette feltet og konseptet «robot» har endret seg de siste 20 årene. Artikkelen ser også fremover, og lurer på hvor mye vår definisjon av robot har endret seg om 50 år. Forfatteren (Dautenhahn) synes forskningen er på vei i feil retning, og at HRI heller burde fokuser enda mer på virkelige mennesker med virkelige problemer og hvordan de ville interagert med en virkelig robot. Dautenhahn mener også det burde forskes mer på interaksjoner over lenger tid i omgivelser som gir mening for denne interaksjonen. Men Dautenhahn er ikke bare skeptisk til HRI og dens utvikling, hun mener HRI har et åpent sinn og at det er fint at ting ikke er satt i stein, at de kan låne metodologier fra andre disipliner og at alle disse ulike disiplinene har mye ved seg som man kan ta med til HRI.

OpenAI

Jeg har valgt OpenAI som er et selskap som ønsker å utvikle AI som verden kan dra nytte av. Ut ifra nettsiden deres virker det som om AI er det eneste de jobber med, og det ser ut som de jobber med mange forskjellige produkter samtidig. De har også et API som de markedsfører hvor folk kan programmere med tekst som deretter blir omgjort til kode. OpenAI snakker om AI som et produkt, men det virker også som om det er noe de tar seriøst, og de skriver at de ønsker å hjelpe andre som også ønsker å utvikle AI løsninger.

Robots and AI systems

Kerstin Dautenhahn (2018) skriver at hva vi definerer som en robot er i stadig utvikling, og det vi så på som en robot for 20 år siden blir ikke nødvendigvis sett på som en robot i dag, og det vi ser på som en robot i dag vil ikke nødvendigvis ses på som en robot i fremtiden.
I 1979 definerte Robot Institute of America en robot som: “a reprogrammable, multifunctional manipulator designed to move materials, parts, tools, or specialized devices through various programmed motions for the performance of a variety of tasks” (Thrun, 2004). Mens IEEE definerer en robot som “*… an autonomous machine capable of sensing its environment, carrying out computations to make decisions, and performing actions in the*

real world." (Guizzo, 2020). Personlig ville jeg definert en robot som *en automatisk maskin som kan utføre de oppgavene den er programmert til å gjøre.* Dette skiller den tydelig fra AI, samtidig som det kommer fram at den kan gjøre oppgaver på egenhånd.

Robot og AI

Jeg mener at en robot ikke nødvendigvis er en AI, men en AI vil alltid være en type robot. Om vi ser på definisjonene fra AI delen ser vi at disse ikke nødvendigvis passer til en robot. En robot er ikke nødvendigvis smart, og den kan fint være programmert til å gjøre akkurat de oppgavene den skal, og ikke noe mer. AI derimot vil i større grad lære av oppgavene den gjør og så bruke denne kunnskapen for å forbedre gjennomførelsen og resultatet av oppgaven. Til og med den første definisjonen av AI av McCarthy er for avansert for en vanlig robot. En robot vil som regel være mer begrenset enn en AI, fordi en AI vil kunne ha de samme kvalitetene som en robot og en AI, mens en robot kun vil ha kvalitetene til en robot.

Robot i reklame

Roboten Moxie er en robot som det kan se ut som er laget for å interagere med barn. Barnet kan bruke den som en slags «levende chatbot» som svarer og som det er mulig å ha en samtale med. Foreldrene kan gjennom en app styre noen deler av roboten, som for eksempel å gi den beskjed om at barnet har en tannlege time. Barnet får en slags venn i Moxie, og litt som i «Her» kan det virke som poenget med roboten er å hjelpe barnet i hverdagen og oppføre seg som en menneskelig venn ville gjort, bare mer fornuftig.

Universelt design

Universal design: “The design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design” (Persson et al., 2015).
Jeg synes dette er en god beskrivelse av universelt design. Universelt design handler om inkludering av alle, uansett. Det skal ikke være nødvendig med ekstrautstyr eller programmer fordi du har dårlig syn eller hørsel. Det skal legges til rette for at alle skal kunne bruke de samme tjenestene.

AI har uendelig med bruksområder, og mange av disse er i forbindelse med mennesker. AIer som SIRI eller Alexa kan hjelpe personer som har vanskelig for å utføre ellers enkle handlinger som å skrive meldinger eller sette på en alarm. Andre typer AI som chatbotter kan brukes av personer som prøver å lære seg et nytt språk for å få en samtalepartner.
AI kan være et veldig godt verktøy for å inkludere ulike brukergrupper siden teknologien naturlig lar seg tilpasse ulike situasjoner.
AI kan også være ekskluderende ved å neglisjere enkelte brukergrupper, og ikke tilrettelegge

for alle brukere gjennom UU. OpenAI som jeg skrev om tidligere utvikler som sagt en tjeneste som lar deg skrive inn hva du ønsker å programmere, og så blir koden skrevet for deg basert på beskrivelsen. Her har man mange muligheter for å tilrettelegge for mange ulike brukergrupper, men det er også fort gjort å overse brukergrupper fordi man blir for opphengt i teknologien, og ikke i personene som skal bruke den. Så vidt jeg kan se har de ikke skrevet noe om at denne teknologien er mer tilrettelagt for spesifikke brukergrupper, noe den fort kunne vært dersom de hadde ønsket å tilrettelegge for det.

Forstå og forståelse

Forståelse kan være et vanskelig begrep fordi man kan ha så mange nivåer av forståelse. Jeg tror roboter kan forstå hva deres oppgave er, men jeg tror ikke de har noen forståelse utover det. I motsetning til AI som kan lære og utvikle seg, vil ikke en robot ha den samme forståelsen for sammenhenger mellom ting. En robot kan ikke vise forståelse på en annen måte enn å gjøre den oppgaven den er satt til å gjøre fordi den ikke klarer å sette ting i kontekst slik et menneske klarer.

Retningslinjer for menneske-ai interaksjon

06 Mitigate social biases.

> AI'en skal ikke ha noen form for bias eller forhåndsdømming basert på sosiale, økonomiske eller andre forhold som kan være med på å forme en mening om brukeren.

Normans sju prinsipper og microsofts retningslinjer for menneske-ai interaksjon har flere likheter enn forskjeller, begge har fokus på tydelighet og hva systemet skal gjøre når noe er feil, men microsofts retningslinjer har i tillegg med punkter som omhandler hva systemet skal få til over tid.

Individuell oppgave modul 2

Karakteristikker av AI-infunderte systemer

Amershi et al., (2019) beskriver AI-infunderte systemer som «systemer som utnytter AI-egenskaper som direkte eksponeres for sluttbrukeren», slike systemer har flere karakteristikker som går igjen.

I forelesningen 28.09.22 forteller A. Følstad om fire hovedkarakteristikker av AI-infunderte systemer. Dette er at systemet lærer, at systemet forbedrer seg, black box, og at systemet er drevet av store datasett. Vi ser at disse går igjen i artikler fra pensum også. Både Amerish et al., (2019) og Kocielnik et al., (2019) skriver at evne til læring og det at det skal forbedre seg over tid er viktige konsepter hos AI systemer. AI er avhengig av store datasett og bruk for å utvikle seg og lære, og dermed også forbedre seg. Videre skriver både Amerish et al., (2019), Kocielnik et al., (2019) og Yang et al., (2020) at AI ofte er lite nøyaktige, og at dette kan være utfordrende for brukerne og føre til usikkerhet. En annen karakteristikk er «black box». Black box er et konsept som handler om å vise for brukeren hvorfor systemet gjorde som det gjorde. Dette er viktig for brukerens forståelse av systemet, noe som igjen vil gi bedre brukerdata og igjen forbedre AI-en.

Et eksempel på et slik AI-infundert system er Spotify. Spotify tar inn masse data om brukerne sine og bruker dette for å tilpasse opplevelsen til hver enkelt bruker. Alle spotifybrukere får spillelister tilpasset seg og deres lyttevaner. Det kanskje beste eksempelet her er listen «Discover weekly» som er en liste som oppdaterer seg ukentlig og gir deg en liste med sanger spotify tror du vil like. Disse sangene velges ut basert på det du lytter til og det andre brukere med lik musikksmak som deg lytter til. Spotify tenker at brukere som hører på lik musikk også antageligvis har lik musikksmak, og dermed vil du få anbefalinger og sanger i listen din som andre med lik musikksmak som deg hører på.

Spotify bruker også andre måter å anbefale musikk til deg på, for eksempel gjennom filtrering av sanger via «tags» og analyse av en sangs trekk. Alt dette (og mer) vil gi deg den opplevelsen Spotify gir deg.

Vi kan se at de fire hovedkarakteristikkene er i full sving her. Systemet forbedrer seg hele

tiden, og anbefalingene man får blir mer og mer tilpasset deg, dette kommer blant annet fra bruk. Jo mer man hører på musikk, jo mer skjønner Spotify om deg, og dermed klarer den også bedre å forstå hva den skal anbefale deg. Spotify tar inn store datamengder fra brukere verden over hele tiden, i tillegg til at all musikken på platformen blir analysert for å gi brukerne den beste opplevelsen. Det er heller ikke veldig tydelig hvorfor sangene som blir anbefalt til deg blir anbefalt. Man kan jo alltids tenke seg fram til hvorfor noen sanger blir anbefalt basert på hva man har hørt på i det siste, men det er relativt lite gjennomsiktighet i hvordan spotify velger ut disse sangene til deg, og hvis sangene noen gang er «feil» og ikke passer til din musikksmak, så får man ikke vite noe om hvorfor akkurat disse sangene ble anbefalt.

Menneske-AI interaksjonsdesign

Amershi et al. (2019) har laget 18 designretningslinjer basert på rundt 20 år med forskning innen HCI og menneske-AI interaksjon. Disse retningslinjene har blitt utarbeidet og forbedret igjennom fire faser og forfatterne forklarer prosessen de har gått igjennom – hvilke metoder de har brukt, hva de ulike fasene fokuserte på og hvordan de har jobbet med dataen de har brukt. De har jobbet med både brukere og eksperter, og hatt utallige fagfolk gå over og jobbe med disse retningslinjene slik at de skal være så brukbare som mulig. Ønsket til forfatterne er at retningslinjene skal kunne brukes av designere for å få bedre og mer menneske-sentrerte AI-infunderte systemer, men de anerkjenner også at ulike prosjekter må gjøre sine egne vurderinger av retningslinjene, og hvordan de passer inn i deres prosjekt.

Kocielnik et al. (2019) skriver om folks forventinger til AI og hvordan høye forventinger kan påvirke folks mottakelse av AI-infunderte systemer. Forfatterne ønsker å se på metoder for hvordan de kan gi folk realistiske forventinger. Dette gjør de ved å bruke Scheduling Assistant, et AI-system laget for å finne ut om det er et møteforslag i mailen din. De bruker to ulike versjoner, en versjon som skal skjule feil og ha flere false negatives, mens den andre skal være mer åpen og ha flere false positives. Forfatterne utformet tre teknikker for å påvirke brukernes forventinger og aksept av AI-infunderte systemer, noe som fungerte veldig bra. De fant ut at brukerne heller ville ha false positives – de ville heller vite at systemet tok feil, enn å ikke vite.

Spotify sin tilnærming til retningslinjene

Jeg har valgt å fokusere på disse to retningslinjene fordi jeg synes Spotify gjør den ene veldig bra og den andre veldig dårlig.

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.

Her synes jeg spotify gjør en dårlig jobb. Det finnes ingen forklaringer på hvorfor sangene du får anbefalt er de du får anbefalt. Man kan tenke seg til hvorfor, eller man kan google og få et generelt svar på hvordan spotify sine algoritmer fungerer, men en vanlig bruker vil ikke få noen indikasjon på hvorfor noen sanger blir anbefalt. Det skjer ofte at spotify anbefaler sanger innen sjangere og kategorier man sjelden eller aldri hører på, så hvorfor er blir disse anbefalt? Jeg skjønner at det kanskje ikke er så lett å gi deg en begrunnelse på hvorfor alle sanger de anbefaler blir anbefalt, fordi det hadde fort blitt veldig rotete, men de kunne gjerne være kommunisert tydeligere hva som påvirker disse listene som er spesifikke for meg.

G13: Learn from user behavior.

Personalize the users experience by learning from their actions over time.

Her synes jeg Spotify gjør en god jobb. De har mange ulike lister og funksjoner som er der for å gi ny musikk du kan høre på, og jeg synes i alle fall at de veldig ofte treffer ganske bra. Det er ganske sjeldent at jeg får opp sanger jeg føler er helt bom. Listene blir også hyppig oppdatert, og det er ingenting man trenger å gjøre selv. Jeg synes også de er flinke til å anbefale musikk basert på det jeg lytter til *nå*, og ikke det jeg lytta til for et år siden.

Bender et al. (2021)

Bender et al. (2021) skriver om AI-systemer basert på store språkmodeller og gjennom artikkelen belyser de flere utfordringer med denne typen AI.

Det er lite støtte for over 90% av alle språkene i verden, så språkene til rundt en milliard av verdens befolkning har for øyeblikket lite eller ingen støtte for språkteknologi.

Man trenger også veldig mye energi for å trene opp språkmodeller, og dette er både veldig kostbart og veldig dårlig for miljøet. Dette kommer hovedsakelig fra ikke-fornybare energikilder og konsekvensene av disse utslippene vil ramme de mest utsatte gruppene – gjerne grupper som ikke vil få noen nytte av denne teknologien. Data hentet fra internett som kan ha problematiske karakteristikker som rasisme, stereotypi og nedlatende assosieringer om kjønn, rase, etnisitet og funksjonshemmede personer. Endringer i sosiale normer og syn over tid kan også være et problem, hvor språk som en gang var greit å bruke over tid endrer seg til å ikke lenger være innafor.

Chatbotter

Chatbotter kan være utfordrende å designe, men vi beveger oss mer og mer mot en verden hvor chatboter blir en sentral teknologi. Designere må tilpasse seg fra å designe en *forklarende* oppgave, altså forklare for brukeren hvordan man bruker systemet og hvordan brukeren skal navigere – til en mer tolkende oppgave, hvor man må forstå hva brukeren trenger og hvordan man skal tilby dette til brukeren (Følstad & Brandtzæg, 2017). Det er vanskeligere å designe for dette ettersom man ikke vet hvordan brukeren kommer til å bruke chatboten, og det er mer utfordrende å styre brukeren i en spesifikk retning, i motsetning til tradisjonelle brukergrensesnitt siden designeren ikke kan støtte seg på grafiske representasjoner for å kommunisere med brukeren.

Mange brukere kan miste tålmodigheten til en chatbot dersom den ikke fungerer slik de forvente og de kan ofte ha en forventing om at chatboten skal fungere så godt som perfekt, noe som er urealistisk. Det må derfor komme tydelig frem *hva* chatboten egentlig kan gjøre. En løsning på dette kan være å følge retningslinjer slik som for eksempel Amershi et al. (2020) foreslår.

Retningslinje G1 og G2 – hvordan kan de hjelpe med å løse dagens utfordringer med chatboter?

Som nevnt over kan det å følge retningslinjene foreslått i Amershi et al. (2020) hjelpe til med å løse noen av utfordringene man har med designing av chatboter. Retningslinje G1 *«Make*

clear what the system can do» kan hjelpe til med å sette realistiske forventinger til chatboter. Mange forventer at chatboter skal fungere som en livechat med et menneske i den andre enden, og blir derfor irriterte når de ikke får de svarene de vil ha.
Retningslinje G2 *«Make clear how well the system can do what it can do»* bidrar også med å sette forventingene til brukeren på et realistisk nivå, siden brukeren får en forståelse av hva systemet er designet for å gjøre.
Man hører mye om chatboter, og hvordan noen faktisk blir forelsket og er i forhold med chatboter, så når chatboten til et selskap ikke står til de forventingene man har fra før av er det fort gjort at brukeren kan bli frustrert. Dersom det kommer tydelig frem hvordan denne chatboten fungerer og hva den kan gjøre for å hjelpe brukeren er det en mye større sannsynlighet for at brukeren tolererer de feil og mangler chatboten måtte ha.

Kilder:

Amershi, Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction. Conference on Human Factors in Computing Systems - Proceedings, 1–13. https://doi.org/10.1145/3290605.3300233

Dautenhahn. (2018). Some Brief Thoughts on the Past and Future of Human-Robot Interaction. ACM Transactions on Human-Robotic Interaction, 7(1), 1–3. https://doi.org/10.1145/3209769

Følstad, A. & Brandtzæg, P., B. (2017*). Chatbots and the new world of HCI. interactions 24* (4)*,* 38–42. https://doi.org/10.1145/3085558

Grudin. (2009). AI and HCI: Two fields divided by a common focus. The AI Magazine, 30(4), 48–57. https://doi.org/10.1609/aimag.v30i4.2271

Guizzo, E. (05.08.2020). What is a robot? Ieee.org. https://robots.ieee.org/learn/what-is-a-robot/

IBM (u.å.) *Artificial intelligence (AI). https://www.ibm.com/cloud/learn/what-is-artificial-intelligence*

Kaplan, & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kocielnik, Amershi, S., & Bennett, P. (2019). Will You Accept an Imperfect AI? Conference on Human Factors in Computing Systems - Proceedings, 1–14. https://doi.org/10.1145/3290605.3300641

McCarthy, J., Minsky, M. L., Rochester, N., Shannon, C. (1955) A proposal for the Dartmouth summer research project on artificial intelligence. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Norman, D. A. (1990). The Design of Everyday Things.

Persson, Åhman, H., Yngling, A. A., & Gulliksen, J. (2015). Universal design, inclusive design, accessible design, design for all: different concepts—one goal? On the concept of accessibility—historical, methodological and philosophical aspects. *Universal Access in the Information Society*, *14*(4), 505–526. https://doi.org/10.1007/s10209-014-0358-z

Roboten Moxie: https://www.youtube.com/watch?v=7YRNjclHTHg

Sebastian Thrun (2004) Toward a Framework for Human-Robot Interaction, Human–Computer Interaction, 19:1-2, 9-24, DOI: 10.1080/07370024.2004.9667338

Totschnig, W. (2020). Fully Autonomous AI. Sci Eng Ethics 26, 2473–2485. https://doi.org/10.1007/s11948-020-00243-z

Yang, Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. Conference on Human Factors in Computing Systems - Proceedings, 1–13. https://doi.org/10.1145/3313831.3376301

Hvordan har jeg jobbet med tilbakemeldingene?

Tilbakemeldingene jeg har fått har vært gode, og de har hjulpet meg med å skrive en bedre oppgave. De har vært ganske konkrete, så det har vært lett for meg å skjønne hva de synes burde forbedres, samt hva de synes var bra. Jeg synes også det var ganske hjelpsomt å lese de andre sine oppgaver og gi de tilbakemelding, fordi det gjorde at jeg så stoffet på en litt annen måte, gjennom noen andres øyne. Det har også vært fint å få høre om hva jeg har gjort som har vært bra ettersom det har hjulpet meg enda mer med å skjønne hvordan jeg kan forbedre de delene som ikke har vært så bra. For eksempel så har jeg fått tilbakemelding på at jeg hadde «fine refleksjoner, men savner eksempler». Man blir ofte blind for egne feil, så det å få en tilbakemelding på det gjør at jeg tenker mer over dette neste gang.