IN5620 Individuell oppgave

1. iterasjon - Høst 2022 - Fay Connie Haugrud

1.1 Konsepter, definisjon og historie med interaksjon av AI og autonome systemer

Før kunstig intelligens ble et begrep, fant man flere eksemplet på menneskelige ønsker om å skape egne selvstyrte maskiner. Selve begrepet ble skapet i 1956 under en seks uker lang konferanse ved Dartmouth college kalt *Summer Research Prosjekt on Artificial Intelligence*. Her ble det samlet forskere innen automasjonsteori, hjernens nevrale nettverk og studier av intelligens. Etableringen av begrepet kommer fra John McCarthy, som var en sentral foredragsholder på konferansen (Tidemann, 2022). Etter denne konferansen var det mye optimisme i forskningsmiljøene. Det var flere som i 1960 hadde stor tro på at maskinene ville bli like smarte som menneskene om få år. Likevel innså man noen år senere at resultatene uteble og i 1970 gikk man inn i en såkalt «AI-vinter» der forskningen stoppet opp. I 1980 kom kunstig intelligens tilbake igjen. I Hollywood ble kunstig intelligens introdusert i offentlighetens bevissthet gjennom store filmer som Terminator (1984) og Short Circuit (1986). Men man klarte heller ikke her å levere slik som ønsket. Det ble derfor en ny «AI-vinter» på slutten av 1980 tallet og utover 1990-tallet (Tidemann, 2022).

Gjennombruddet for AI kom på 2010. Metodikken, regnekraften og datamengden muliggjorde at man virkelig kunne ta skrittet ut av forskningslaboratoriene og inn i den industrielle verden (Tidemann, 2022). I 2012 kom er viktig vendepunkt i forbindelse med konkurransen ImageNet. Den årlige testen for automatisk gjenkjenning av objekter i en stor bildedatabase. Det var Alex Krizhevsky som utviklet en etablert teknikk som baserte seg på nevrale nettverk som videre senket feilraten på bildegjenkjenning med oppsiktsvekkende stor margin (Tidemann, 2022). Denne metoden viste seg å kunne brukes på tolkning av sekvenser som lys og tekst. Dette gjorde at kunstig intelligens kom tilbake for fullt, både i folks bevissthet, men også i den industrielle sammenheng. Hvorfor gjennombruddet kom i 2012 og ikke før ettersom nevrale nettverk hadde eksistert i flere tiår har i hovedsak to faktorer ifølge Tidemann:

- Man hadde tilgang til kraftigere datamaskiner som kunne trene opp store modeller relativt raskt (dvs. i løpet av dager eller uker)
- Man hadde store datamengder som kunne trene opp slike modeller. Det krevde mye data, ettersom de var så store

Kunstig intelligens har blitt definert flere ganger siden 1956. Bratteteig & Verne (2018) definerer AI som et «*underfelt av informatikk rettet mot å spesifisere og lage datasystemer som etterligner*

menneskelig intelligens eller uttrykker rasjonell atferd, i den forstand at oppgaven ville kreve intelligens hvis den ble utført av et menneske». Regjeringen (2022) definerer AI som «*systemer som utfører handlinger, fysisk eller digitalt, basert på tolkning og behandling av strukturerte eller ustrukturerte data, i den hensikt å oppnå et gitt mål*». En tredje definisjon hentet fra det Store norske leksikon definerer AI som et «*informasjonsteknologi som justerer sin egen aktivitet og derfor tilsynelatende fremstår som intelligent*» (Tidemann, 2022). De tre definisjonene vektlegger ulike aspekter ved kunstig intelligens. Til felles har alle definisjonene en maskin eller robot som klarer å utføre handlinger eller prosesser som til vanlig er forbeholdt mennesker.

Basert på dette har jeg utformet min egen definisjon: «*Kunstig intelligens er et digitalt system med menneskeskapt oppfatningsevne, behandlingsevne og oppgaveløsning*». Det er med andre ord en menneskeskapt digital løsning som har evne til å oppfatte, tenke og utføre problemløsning ved bruk av forskjellige fagfelt innen informasjonsteknologi.

Muligens den største forskjellen mellom roboter og andre fysiske enheter som eksempelvis husholdningsapparater – er autonomi. Thrun refererer til en robots autonomi som «*evnen til å imøtekomme variasjoner i miljøet*» (Thrun, 2011). Menneskelig autonomi refereres som «*en persons evne til å selv bestemme sine handlinger og grunnlaget for handlingene*» (Sagdahl, 2019). For en robot må graden av autonomi måles ved å relatere til hvilken grad miljøet kan varieres til gjennomsnittstiden mellom feil og andre faktorer som indikerer robotytelse. For mennesker er autonomi når grunnene for hvordan en person handler – personens egne grunner. Jeg vil si det er her man ser den største ulikheten mellom menneskelig og maskin autonomi. Vi mennesker gjøre egne beslutninger basert på hva vi vil, ønsker, tenker og føler. En maskin tar ikke egne handler basert på hva «den» vil. Den blir enten overstyrt, programmert eller påvirket gjennom AI til hvordan den skal håndtere situasjoner og hendelser.

Vi omgir oss mer og mer automatiserte og autonome systemer, som eksempelvis droner, på land og på sjøen, autonome skip og selvkjørende biler. Forskjellen på automatisk og autonom, er at automatisk er i strand til å fungere uten ekstern kontroll eller inversjon mens autonomt virker eller forekommer ufrivillig, uten noen bevisst kontroll.

Jeg har valgt artikkelen «*Some brief thoughts on the pas and future of Human-Robot Interaction*" av Kerstin Dautenhahn. Denne artikkelen tar for seg noen tanker om fortiden og fremtiden til menneske-robot interaksjonen (HRI). Den tar for seg utfordringer om hvordan man kan utføre og analysere forskningen videre når det kommer til HRI. HRI er veldig annerledes fra menneske-menneske interaksjon, menneske-data interaksjon og menneske-dyr interaksjon. Ettersom robot er et bevegelig mål, er det vanskelig å stedfeste hvordan utviklingen vil utarte seg i årene fremover. Dautenhahn

skriver også om bekymringer når det kommer til hvordan man kan adressere virkelige mennesker, i virkelige miljøer samhandle ansikt til ansikt med en ekte robot i økologiske gyldige miljøer. Da spesifikt i sammenhenger der roboten er et komplekst, intelligent og multitasking system. Hun skriver at dette er et vanskelig område å forske på, nettopp fordi det mer tidskrevende. Likevel påpeker hun at man ikke kan vike unna vis man vil at HRI-resultatene skal ha implikasjoner og sosial innvirkning utenfor laboratoriene. Videre skriver hun om optimisten hun har for HRI og at det er viktig å holde liv i diskusjonen for å komme videre i utviklingen og finne beste måten å gjøre det på. Jeg stilte forså vidt ingen spørsmål til artikkelen. Jeg ble mer interessert i å finne ut av hvordan disse problemstillingene løses og resultatet av dem. Det virker både kult og skummelt å se hva fremtiden kan by på når det kommer til hvor «menneskelig» roboter kan bli og hvordan de kommer til å påvirke samfunnet.

No Isolation er et norsk selskap, grunnlagt i 2015 som arbeider med AI. De ser på kunstig intelligens som et verktøy for å redusere ensomhet og sosial isolasjon gjennom å bruke varm teknologi og kunnskap. På nettsiden deres ser man fort hvilke produkter de presenterer: AV1 og Komp. AV1 er en kommunikasjonsrobot for barn og unge med langtidssykdom, og Komp er en datamaskin med kun en knapp for de med lite teknologisk erfaring.

Black Mirror er en britisk science fiction- og thrillerserie laget av Charlie Brooker. Black Mirror utforsker de mørke sidene ved å bruke teknologi. Hver episode er unik og deler en annen historie om hvordan sosiale medier og teknologi forbruker livene våre. Et eksempel på mennesker og AI kan sees i S4E2 som handler om at en mor utstyrer sin datter med en hjernebrikke som lar henne overvåke handlingene til datteren. Alt som gir dattera en form for stress, sensurerer moren gjennom AI. Så hvis dattera ser noe voldelig på TVen, kan moren sensurere innholdet. Med andre ord har Black Mirrir brukt AI som et middel for sensur.

1.2 Roboter og AI-systemer

Selve ordet robot stammer fra et tsjekkisk teaterstykke fra 1920, men robot lignende skapninger har i flere år hundre vært en del av menneskets historie. Men det var først i 1954 at den første programmerbare digitale roboten kom (Søraa, 2022). Det var den amerikanske George Devol som lagde den første programmerbare digitale roboten. Denne ble tatt i bruk av Ford-fabrikken til å løfte varmt metall (Søraa, 2022).

Russell & Nordvig (1995) definerte en robot som «*en om programmerbar, mulitifunksjonell manipulator designet for å flytte materialer, deler, verktøy eller spesialiserte enheter gjennom ulike programmerte bevegelser for å utføre en rekke oppgaver*» (Thrun, 2011). Dette er en mekanisk definisjon som ser på roboten som en maskin som har til hensikt å utføre spesifikke oppgaver. Det er

ingen rom for autonomi. En annen definisjon fra Merriam Websters collegiate dictionary (1993) så en robot som «*en automatisk enhet som utfører funksjoner som normalt tilskrives mennesker eller en maskin i form av et menneske*». Denne definisjonen går mer på menneskelige og intelligente aspekter. Begge disse definisjonene er enig i at en robot er noe som er fysisk, forskjellen fra dem er at den første definisjon tar utgangspunkt i en automatisert maskin, mens den andre definisjonen sier at roboter gjør oppgaver som er forbeholdt menneskelige aktiviteter. Min definisjon på hva en robot er vil være følgende «*en fysisk enhet som er programmert til å automatisk gjøre oppgaver eller handlinger*".

En robot kan inneholde større eller mindre grad av AI. Tar vi utgangspunkt i definisjonene fra Russel & Nordvig så kan en robot fungere uten noen form for AI (dette fordi den er programmert til å gjøre spesifikke oppgaver). Det er ikke forventet at den skal gjøre mer, med mindre koden endres. En robot som inneholder større grad av AI vil være mer avanserte og gjøre mer kompliserte oppgaver en dem uten. For eksempel vil den lære seg nye ting og utføre oppgaver bedre og bedre for hver gang. Ser man på forholdet mellom roboter og AI så er AI noe en robot kan ha mer eller mindre av.

AV1 er en skolerobot som sørger for at barn med langtidssykdom, skolevegring eller andre utfordringer ikke trenger å gå glipp av undervisning, friminutt eller annen sosialisering. Roboten plasseres i klasserommet, hvor eleven kan koble seg til hjemmefra gjennom nettbrett eller telefon. Roboten har et innbygd kamera, mikrofon og høyttaler, i tillegg til at den kan rekke opp hånden. Det er kun eleven som kan styre roboten, noe som gjør at eleven kan delta og se hva som foregår. Den har støtte for 4G noe som gjør at den har få begrensninger når det gjelder involvering.

1.3 Universell utforming og AI-systemer

FNs konvensjon om rettigheter til mennesker med nedsatt funksjonsevne har følgende definisjon av universell utforming (UU): «*Utforming av produkter, omgivelser, program og tjenester på en slik måte at de kan brukes av alle mennesker, i så stor utstrekning som mulig, uten behov for tilpasning og en spesiell utforming*». Denne definisjonen handler om som nevnt ovenfor at alle, også de med nedsatt funksjonsevne skal ha mulighet til å ta i bruk produkter, omgivelser, program og tjenester på lik linje som andre. Inkludering betyr å høre til og ha en plass i et fellesskap. I faglige sammenhenger handler det om samspill, tilgang på ressurser, roller, avbildning og nettverk. Med andre ord handler UU om at alle skal ha like muligheter og bli sett.

AI har bidratt til at andre som eksempelvis sårbare i samfunnet har tilgang til deltakelse på en annen måte enn tidligere. Jeg kan igjen trekke inn AV1 som har muliggjort at langtidssyke barn kan delta i undervisning og bursdag, selv om de fysisk ikke kan være til stede. Et annet eksempel er selvstyre bilder som har muliggjort at mennesker av ulike årsaker ikke kan kjøre bil. Men gjennom å kjøre en

selvstyrt bil kan de være selvstendige og mindre avhengige av andre. Svakenheten til AI er at maskiner kun lærer av den dataen man mater den. Det betyr at systematiske bias som oppstår i samfunnet i dag, vil videreføres gjennom maskinlæring. Et eksempel på dette kan være ansettelsesprosesser. Dersom de fleste interaksjonsdesignerne er hvite kvinner på 40 år, så kan kanskje systemet lære seg at det er denne type profilen man skal se etter når man ansetter en ny interaksjonsdesigner. Heldigvis er det vi mennesker som skal utvikle hvordan AI-et funker, og dette gir oss mulighet til å rette opp i skjevhetene. Det krever at man har kunnskap for å utelukke biaser som er en del av samfunnet nå og i fremtiden, men så lenge man vet om det, kan man bidra til å skape et mer inkludere samfunn som gir like muligheter for alle.

Når det kommer til begrepene «*understand*» og «*understanding*», forstår jeg dem på følgende måte: *Understand* vil si at du skal forstå noe. Man begriper, fatter, innser og oppfatter noe. Mens *understanding* betyr å komme til en forståelse av noe. Det du prøver å forstå, blir forstått. En datamaskin kan eksempelvis forstå en input innenfor de rammene vi har gitt den. Den vil kunne kjenne det igjen og gjøre operasjoner fortere enn et menneske kan gjøre det. Men forskjellen på et menneske og en datamaskin er at den ikke vil ha forståelse for de kulturelle og sosiale kontekstene.

Det er mange systemer som har blitt utviklet med sikte på kunstig intelligens, likevel er det få som har eksplisitt målrettet forståelse. Så `*do machines understand?´*. Kanskje det mest eksplisitte forsøket på å fylle maskiner med forståelse kan kanskje være arbeid med sunn fornuft, som likevel har avledet oppmerksomheten bort fra forståelse, som er rettet til menneskesentriske forestillinger om «sunn fornuft». Vi er mennesker og ikke maskiner, og maskiner er maskiner og ikke mennesker. Vi opplever rasjonelle og emosjonelle tanker og følelser, noe en maskin ikke gjør. Det er likevel noen likheter, og analytisk kan vi se på disse områdene:

Fysisk	Menneske: kroppen	Datamaskin: maskinvare
Sensorer	Menneske: syn, øyne	Datamaskin: kamera/sensorer
Effektorer	Menneske: stemme	Datamaskin: høyttaler
Argumentasjon	Mennesker: tenkning	Datamaskin: beregninger
Følelser	Menneske: følelser	Datamaskin: simuleringer

Vi mennesker lærer, resonnerer, føler og drømmer. Dette gjør ikke maskiner. Maskiner blir programmert/lært til å gjøre noe, og noe mer enn hva det er blitt «fortalt» til å gjøre, gjør den ikke. Så om maskiner forstår, det gjør de nok. De forstår oppgaven(e) de har og utfører de. Men de forstår ikke noe mer ut over det, slik som oss mennesker.

1.4 Retningslinje for menneske-AI-interaksjon

Nr 7. «*Support efficient invaction*»
Denne retningslinjen handler om at systemet/ produktet kan vise/gi informasjon som er relevant i den konteksten den er i. Et eksempel på dette kan være at jeg spør Google Nest om hvordan været skal bli i morgen.

HCI tar høyde for hvordan man skal designe og implementere datasystemer som tilfredsstiller menneskelige brukere. Det tar sikte på å forbedre interaksjon mellom mennesker og datamaskiner ved å forbedre hvordan et grensesnitt brukes og forstås av oss mennesker. Tar vi for eksempel Schneiderman´s åtte gylne regler innenfor HCI handler det om hvordan man skal utforme brukergrensesnitt. Human-AI-interaksjon handler om interaksjonen mellom mennesker og AI. Microsoft har delt sine retningslinjer inn i fire faser (før og under interaksjon, når systemet gjør en feil, og over tid), mens Schneiderman´s regler er mer generelle og statiske for hvordan utformingen bør være. De likhetene man ser er eksempelvis at begge har fokus på at brukeren for informativ respons for interaksjonen sin eller at brukeren kan unngå feil eller angre handlingene sine. Flere av prinsippene går inn på de samme aspektene, men de kommer med ulike mål.

Del 2

2.1 Kjennetegn på AI-baserte systemer

Nøkkelegenskaper ved et AI-basert system er læring, forbedring, blackbox og drevet av store datasett (Følstad A. , 2022, slide 28). Den første, læring, er viktig for at systemet skal kunne klare å ta imot informasjonen, bearbeide den og med et utføre oppgaver. Forbedring handler om at systemet skal kunne forbedre seg. Når systemet forbedrer seg, så utfører de også oppgaver bedre over tid. Dette impliserer også at systemer kan gjøre feil, og at systemet derfor bør være designet slik at de effektivt korrigerer dette. Den tredje egenskapen er blackbox. Det er mange prosesser som skjer i et AI, og som bruker forstår man ikke alltid alt. Det kan derfor være vanskelig å forstå hvorfor man får en output, ettersom man ikke nødvendigvis alltid vet hvilken informasjon AIet sitter på. Dette er noe av det som skjer i en blackbox. Den siste egenskapen er store datasett, som oppnås gjennom interaksjonen med brukere og systemer (Følstad A. , 2022, slide 28-47).

AI basert system kan for eksempel være YouTube. YouTube er verdens største nettsted for opplastning, visning og deling av videoklipp, og er bygger opp av algoritmer som lærer seg hvilket innhold du som bruker liker og foretrekker. Baseert på dette gir YouTube deg anbefalinger. Etter hvert som man trykker seg inn på foreslåtte forslag, blir den mer og mer nøyaktig på å finne innhold som er interessant. Dette kan være positivt og negativet. For eksempel kan det være at du har en hjertesak for abort og at denne saken forsterker seg i større grad nettopp fordi innholdet man interesserer seg for, forsterkes ved hjelp av algoritmen. Med mindre man selv oppsøker motstridende meninger, vil algoritmen fortsette å gi brukeren det samme budskapet. Et negativt aspekt ved dette er at tjenesten kan gi deg en følelse av at dette synet, er det eneste fornuftig alternativet. Det er ikke bare YouTube som benytter seg av disse, både Instagram og TikTok er gode på nettopp dette.

For eksempel har Instagram bare det siste året, spesialisert seg på å gi brukeren det brukeren liker og ser på mest. Personlig bruker jeg Instagram til å få inspirasjon og lærdom om mote og makeup, og det første jeg ser når jeg åpner Instagram er nettopp dette. Det samme ser jeg på TikTok. Der bruker jeg mye av tiden på å se treningsvideoer/øvelser, og hele min feed på TikTok blir fylt av dette. Trykker jeg tilfeldigvis på videoer om mat, og liker noen av disse, blir feeden min fylt med mat-videoer, og det er akkurat dette som er effekten av algoritmer. Spiralen fungerer ved at systemet lærer seg hva du liker, og forbedrer seg på bakgrunn av forslagene dine. Brukeren gir fra seg data som systemet tar med seg videre, og forer inn i AI-et. Dette skjer i en black box hvor brukeren ikke nødvendigvis har noen innsyn i hva som lagres, hvor mye som lagres og når det lagres.

2.2 Human-AI interaksjonsdesign

Amershi et al. (2019) viser til 18 designretningslinjer for menneske-AI-interaksjon. Retningslinjer er ment som en ressurs for å arbeide med systemer og tjenester som inneholder AI. Artikkelen tar for seg metoder for hvordan de har kommet fram til dette. Fase 1. var en innsamlingsfase hvor designanbefalinger for kunstig intelligens ble samlet inn fra 1. anmeldelse av produkter fra industrien, 2. nylig publiserte artikler om AI-design og 3. relevante forskningsartikler om AI-design. Fase 2 tok for seg en intern evaluering av funnene funnet i fase 1. Fase 3 tok for seg en brukerstudie der de samlet personer som arbeidet med HCI til vanlig for å se om de forstod retningslinjene i praksis og for å få tilbakemelding på hva de betyr og tydeligheten av dem. Fase 4 var siste fasen, der de hadde en ekspertevaluering av endringene de hadde gjort under brukerstudien.

Kocielnik et al. (2019) snakker om tre teknikker for forventingsjustering i et AI-drevet system og tar for seg systemet Scheduling Assistant. Scheduling Assistant er et AI-system som automatisk oppfatter møteforespørsler i eposter. Studien viser hvordan tre teknikker bevarer bruketilfredsheten og aksepten for at et system ikke er perfekt. Studiet gikk ut på å sette de rette forventingene til et AI-system for at brukeren skal kunne ta det imot på riktig måte. De utforsker to versjoner av samme system med samme treffsikkerhet og aksept, og foreslo tre teknikker for å legge grunnlaget for forventninger.

1. En *accuracy indicator* som eksplisitt sier hvor nøyaktig systemet er
2. *Eksempelbasert forklaring*, som prøver å øke forståelsen til bruker
3. *Performance control* som tillater brukeren til å direkte tilpasse presentasjonen

Studien viste også at AI-systemer som unngår å vise falske positive feil, fører til at brukeren føler at systemet er mindre treffsikkert og har lavere aksept enn de systemene som prøvde å unngå falske negativer, til tross for at de var like treffsikre.

Når det kommer til eksempelet over om YouTube, kan vi se på Amershi et al. (2019) sine 18 designretningslinjer for menneske-AI-interaksjon. Retningslinje 13 handler om læring fra brukeratferd, som er relevant når det kommer til mitt eksempel. YouTube satser stort på å personalisere brukeropplevelsen med deres tjenester ved at de analyserer alle brukerinput og brukerdata for å lære mer om hvordan systemet kan gi brukeren er bedre og tilpasset brukeropplevelse. Retningslinje 8 er også relevant og omhandler å støtte effektiv avfeiing. Det vil si at hvis du er bruker som får opp en innholdsskaper eller for eksempel en video som ikke er interessant, så kan man enkelt gi beskjed om at du ikke liker dette. YouTube har et bevisst forhold til hvordan deres AI funker ovenfor oss forbrukere og det ser ut til at de også prøver å forbedre disse punktene. Det de dog ikke kan, er å kontrollere hvordan systemet lærer brukeren å kjenne og hvordan brukeren tar imot/ reagerer på at systemet gjør dette. Det kan virke ubehagelig at system vet hva man liker, eller at de lagrer

informasjonen du mater den. Det de kunne gjort som en løsning, er å la brukeren velge om systemet skal kunne lagre informasjon om deg. På den måten kan man selv bestemme om man skal ha en profil som blir analysert, eller om han ønsker å være «anonym» selv om man har en konto på nettstedet. Denne muligheten bruker eksempelvis Apple på sine enheter. Her får man valget om man ønsker å bli analysert, eller om det er ønskelig å ha dette avslått. I tillegg kan man gå inn på innstilinger å endre dette etter behov.

2.3 Chatbots / samtalebrukergrensesnitt

Følstad og Brandtzæg (2017) beskriver tre implikasjoner for HCI; samtaler som designobjekt, behovet for å endre fokus fra brukergrensesnitt til tjenestedesign og behovet for å designe nettverk for både mennesker og bots. Nedenfor vil jeg gå gjennom disse nærmere.

Samtaler som designobjekt:

Når man skal designe chatbotter, er dette ulikt fra visuelt design ved at man her fokuserer på interaksjonen mellom bruker og chatbot, og mekanismene samt. innholdet for selve samtaleflyten. Det er viktig å designe for brukbarhet ettersom mye av funksjonaliteten i en chatbot er skjult for brukeren. Dette er ulikt fra eksempelvis grafisk design alene, nettopp fordi man designer for selve interaksjonen og gi brukeren forslag. Samtidig som man skal forklare hva forslagene kan brukes til, og med det legge til rette for å utnytte de funksjonene og mulighetene som er der. Følstad og Brandtzæg (2017) argumenterer også for at design av samtaler ikke bør sees på som en forklarende oppgave – men til en fortolkende oppgave. Det betyr at man går fra å forklare for brukeren hvilket innhold som faktisk er tilgjengelig, til å faktisk forstå hva brukeren behøver og med det forstå hvordan dette best kan støtte deres behov.

Brukergrensesnitt til tjenestedesign:

Her argumenterer de for at interaksjonen med forskjellige nettsider og tjenester vil smelte sammen i fremtiden, og at HCI i dag er mer rettet mot enkelbrukeren. Følstad og Brandtzæg (2017) mener det vil være fornuftig å ha mer fokus på hva tjenestene gjør i sin helhet, slik at overgangen ikke blir for brå.

Designe nettverk for både mennesker og bots:

Her argumenter Følstad og Brandtzæg (2017) for utfordringen ved å designe for flere agenter. I dag tar HCI stort sett for seg en bruker og en tjeneste. De studerer om det finnes større sosiotekninske systemer som krevet at man designer for flere agenter enn bare en-til-en interaksjon. Her trekker de blant annet inn Microsoft sin chatbot Tay som ble manipulert og lært av andre til å bli drøy, eller radikal om du vil.

Amershi et al. (2019) presenterer som nevnt 18 retningslinjer, og den første kategorien omhandler mennesker og kunstig intelligens som går på de tingene man kan designe før selve interaksjonen skjer.

Retningslinje 1: Gjør klart hva systemet kan. Denne retningslinjen handler om å hjelpe brukeren med å forstå hva AI-systemet er i stand til å gjøre. Det vil si at man må designe for samtalen og forklare brukeren hva den faktisk kan gjøre. Dette gjør man for å sette en hvis standard til brukeren, slik at brukeren vet i sin helhet hva systemet kan og ikke kan.

Retningslinje 2: Gjør det klart hvor gode systemet kan gjøre det det kan. Denne retningslinjen handler om å hjelpe brukeren å forstå hvor ofte AI-systemer kan gjøre feil. Med det mener jeg at systemet må gi en indikasjon på hvor godt det kan utføre de tingene den kan gjøre, slik at brukeren vet hva den er kapabel til og ikke.

Bibliografi

Amershi, S. W. (2019). Guidelines for human-AI interaction. . *In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (paper no. 3). ACM.* .

Bratteteig, T., & Verne, G. (2018). *Gjør AI PD foreldet? Utforsker utfordringer fra kunstig intelligens til deltakende design.* August: ACM.

Følstad, A. &. (2017). Chatbots and the new world of HCI. . *interactions, 24(4), 38-42.* .

Følstad, A. (2022). Interaction with AI and autonomous systems - Module 2. Session 4. . *SINTEF* .

Kocielnik, R. A. (2019). Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. . *In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (paper no. 411). ACM.* .

Regjeringen. (2022, september 6). *Nasjonal strategi for kunstig intelligens* . Retrieved from regeringen.no : https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/?ch=3

Sagdahl, M. S. (2019, 25 juni). *Autonomi - filosofi*. Retrieved from Store norske leksikon: https://snl.no/autonomi_-_filosofi

Søraa, R. A. (2022, mai 12). *Robot*. Retrieved from Store norske leksikon: https://snl.no/robot

Thrun, S. (2011, Juni 22). Human-Computer Interaction. *Toward a Framework for Human-Robot Interaction, Human-Computer Interaction*, p. 17.

Tidemann, A. (2022, september 6). *Kunstig inteliggens*. Retrieved from Store norske leksikon : https://snl.no/kunstig_intelligens

Vedlegg

Tilbakemelding iterasjon 1:

Jeg fikk tilbakemelding på at jeg hadde en fin presentasjon av historien til AI og at jeg tok opp viktige aspekter og store gjennombrudd. Jeg hadde en god egen komponert definisjon av AI, noe som viser at jeg har satt meg godt inn i andre definisjoner. Var et ønske om å reflektere litt mer, noe jeg har fikset opp i.

Tilbakemelding iterasjon 2:

Begrepene i teksten blir beskrevet konkret og er lett forståelig, noe som viser at jeg har god forståelse ovenfor pensum. Jeg reflekterer godt rundt «algoritmespiralen» på sosiale medier, noe som er et smertepunkt i AI-baserte systemer som mange kjenner seg i igjen i. Var et ønske om å bruke enda flere kilder i iterasjon 2, da dette ville fått oppgaven til å virke mer reflektert og innehold flere synspunkt på samme sak. Kilder er oppdatert.