IN5620 Individuell oppgave høst 2022
Daniel Gloppestad Bajer

Modul 1

1.1 Konsepter, definisjoner og historien til AI og autonome systemer
Det var John McCarthy som først brukte begrepet AI om kunstig intelligens (Artificial Intelligence, AI). Han var en del av en teknologioptimistisk bevegelse i etterkrigstiden, men hadde et annet syn enn mange fra sin samtid. Mange så for seg at AI om ikke lenge ville kunne handle intellektuelt på linje med mennesker. Med et slikt standpunkt så man naturligvis også for seg at AI etter hvert ville utføre oppgaver tidligere ansett som intellektuelle, mye bedre enn det mennesker var i stand til, og på den måten utkonkurrere mennesket. Alan Turing var en viktig person i den tidlige populariseringen av AI, og med sin velkjente Turing -test bidro han til ideen om at AI skulle kunne måle seg med mennesket. McCarthy mente på sin side at AI ikke nødvendigvis skulle sammenlignes med menneskelig intelligens og at AI var en del av fagfeltet Computer Science, ikke psykologi.

Science fiction litteraturen spilte også sin rolle med å gjøre kunstig intelligens til et kjent konsept og AI -relatert forskning og arbeid fikk mye økonomisk støtte på 40 og 50 -tallet. Man hadde sett datamaskinens potensial i praksis under andre verdenskrig, så det var store økonomiske krefter i sving for å støtte utviklingen av AI -teknologi. *(Grudin, Jonathan. 2009)*

Det skulle etter hvert vise seg at AI ikke utviklet seg så fort som først spådd og selv i dag, hvor datamaskiner er i stand til å gjøre mange oppgaver bedre enn mennesker, er det vanskelig å argumentere for at AI skal kunne erstatte menneskelig intelligens (Montemayor et al. 2022)

Schuett, Jonas. (2019) presenterer tre definisjoner av AI.

1. AI er en datamaskin som består Turing -testen.
    - Alan Turing, 1950

2. AI betyr vitenskapen og oppbyggingen av intelligente maskiner. Intelligens refererer her til regningen som ligger bak evnen til å oppnå mål i den virkelige verden.
    - John McCarthy, 2007

3. AI er en intelligent agent, det vil si en maskin som har evnen til å ta inn sensorinput fra verden rundt seg og prosessere disse for å utføre handlinger som følger en gitt heuristikk.
    - Russel og Norvig, 2009

Av disse definisjonene er det McCarthy sin jeg ønsker å basere meg på i størst grad. I denne oppgaven vil jeg gi følgende definisjon av AI:

AI er en entitet som er stand til å samle inn og prosessere data for å oppnå forhåndsdefinerte mål.

Som mennesker er vi i stand til å ta egne valg. Eller vi lever hvert fall i troen om at vi er det. En måte å se på menneskelig autonomi på er å tenke seg at vi er bundet til vår biologiske eksistens og at våre lyster og begjær ikke tillater oss å handle ut ifra egen autonomi. Mennesker kan likevel føle at de handler autonomisk og i denne oppgaven vil jeg gå ut fra denne tankegangen for å definere menneskelig autonomi. Min definisjon av menneskelig autonomi er:

En følelse av å kunne gjøre egne valg og å handle på en måte som gir mening for enkeltmennesket og dets omgivelser.

Vi sammenligner denne definisjonen med hvordan autonomi defineres av *Antsaklis, & Rahnama, A. (2018)*

A system is autonomous with respect to a set of goals and
under a set of (subject to certain classes of) uncertainties,
if it achieves these goals under those uncertainties without
external intervention.

Her ser vi at følelser ikke er inkludert i *Antsaklis, & Rahnama, A. (2018)*, og det er vanskelig å tenke på maskiner som følende vesener. Man kan også se en forskjell mellom definisjonene der den første bruker ordet *mening*, bruker *Antsaklis, & Rahnama, A. (2018)* måloppnåelse. Man kan her argumentere for at mennesker trenger mening for å føle autonomi, mens maskiner ikke har noen følelse for mening, bare måloppnåelse. I tillegg definerer Antsaklis, & Rahnama (2018) et avgrenset system for maskinautonomi, nettopp for å kunne kontrollere måloppnåelse og eventuell menneskelig kontroll utenfra. Menneskelig autonomi på den annen side er ikke adskilt fra påvirkning utenfra. Det er fullt mulig, og ganske naturlig, for et menneske å føle autonomi selv ved påvirkning fra omgivelsene. Dette er ikke en hindring for menneskelig autonomi fordi vi alltid er knyttet sammen med omgivelsene våre.

På *dictionary.com* finner vi følgende definisjon av hva *automatisk betyr:*

automatic
having the capability of starting, operating, moving, etc., independently:
an automatic sprinkler system; an automatic car wash.

Vi kan se at i likhet med definisjonen for *autonomi* finner vi også i denne definisjonen et fokus på uavhengighet, altså noe som er stengt fra påvirkning utenfra. Det kan virke som om den største forskjellen mellom en autonom og en automatisk maskin, ut ifra definisjonene ovenfor, er at en autonom maskin er i stand til å håndtere en større grad av usikkerhet. Hvis vi sammenligner en vaskemaskin med en robotgressklipper, kan vi tenke oss at vaskemaskinen er automatisk mens robotgressklipperen er autonom. Vaskemaskinen utfører et bestemt program, mens robotgressklipperen er nødt til å gjøre vurderinger og valg underveis i jobben basert på skiftende omstendigheter. Hvis man ser nærmere på saken, kan man imidlertid se at det ikke er så enkelt å klassifisere inn i automatisk og autonom. En vaskemaskin må for eksempel tilpasse ytelsen sin basert på hvor mye klær som vaskes om gangen. Dersom det er mye klær kreves det mer motorkraft for å oppnå et bestemt antall

omdreininger i minuttet. Dette kan sees på som en form for autonomi. Man kan dermed argumentere for at forskjellen mellom en automatisk og autonom kun er basert på grad av kompleksitet, og det er vanskelig å sette en konkret grense for hva som er komplekst nok til å kunne regnes som autonomt, eller omvendt for automatisk.

Tone Bratteteig and Guri Verne. 2018 diskuterer om AI gjør Participatory Design (PD) overflødig. De peker på at moderne AI -teknologi deduserer sine egne regler basert på store mengder data. Dette skaper en utfordring for designere fordi det blir vanskelig å forstå valgene som gjøres av AI -teknologien som former de AI -skapte resultatene og produktene. Det er også vanskelig å forutse hvordan fremtidig teknologi skapt av AI vil fungere, ettersom det vil endre seg over tid, og dermed vanskelig å ta det med i betraktningen i en designprosess. Det er viktig innen PD at konteksten og situasjonen skal være sentrale i designprosessen. Dette blir også en utfordring med tanke på at maskinlæring baserer seg på store mengder data som bidrar til generalisering og større avstand til den aktuelle konteksten. Hvis man imidlertid er bevisst hvilke data AI -systemet trenes på og sørger for at disse kommer fra en relevant kontekst har man gjort et viktig steg. De konkluderer med at AI står for en stor utfordring for PD, både for designere, deltakere, men at PD også kan være nyttig for å møte utfordringene som oppstår sammen med økende bruk av AI

Jeg er enig med synet som presenteres i denne artikkelen. Jeg mener det er viktig å sette fokus på brukeren i en tid der mer og mer automatiseres.

På sine hjemmesider presenterer *Meta* sin språksatsing på følgende måte:
No Language Left Behind (NLLB) is a first-of-its-kind, AI breakthrough project that open-sources models capable of delivering evaluated, high-quality translations directly between 200 languages—including low-resource languages like Asturian, Luganda, Urdu and more. It aims to give people the opportunity to access and share web content in their native language, and communicate with anyone, anywhere, regardless of their language preferences.

Meta presenterer her AI som kan oversette mellom 200 språk. De presenterer det som en tjeneste, som i tillegg er open source. Det kommer likevel også frem at de ser på denne teknologien som sentral i sin metaverse -ide, som også vil fungere som en verdensomspennende økologi. Hvis NLLB betraktes som en tjeneste, har det potensial til å være et nyttig verktøy for kommunikasjon og samarbeid. Det er imidlertid flere ulemper som kan dukke opp dersom det blir en integrert del av sosiale medier. Vi ser allerede i dag at tekster og innlegg oversettes automatisk på sosiale medier, dette kan i visse tilfeller føre til at tekster feiloversettes og misforstås. Sosiale nettverk kan la brukeren velge selv om han eller hun ønsker å oversette en bestemt tekst eller ikke (Dette gjør f.eks. Facebook, se referanser). Det vil være viktig å gi brukeren denne valgfriheten for at ikke feil og misforståelser skal oppstå og for å ta hensyn til kulturelle forskjeller som påvirker kommunikasjonen. På den måten kan vi unngå en utvikling der vi legger mer og mer ansvar for mellommenneskelig kommunikasjon på AI -baserte systemers.

Filmen Her av Spike Jonze fra 2013 handler om en mann som forelsker seg i et operativsystem. Filmens handling utspiller seg i en virkelighet der det er veldig vanlig å til enhver tid være tilkoblet en digital personlig assistent, og kommunisere med denne verbalt.

AI presenteres som en allvitende entitet, en entitet som kan hente opp en enorm mengde informasjon. Denne interaksjonen med AI minner mye om hvordan vi forholder oss til smartassistenter som for eksempel Apples Siri i dag, men bruken er mye mer omfattende i filmen.

Utover i filmen utvikler hovedpersonens forhold med sin smartassistent seg til å bli romantisk og det ender opp med at smartassistenten dumper ham. Smartassistenten uttrykker at *hun* trenger mer sofistikert stimulering og at hun ikke kan finne dette hos menneskelige partnere og derfor vil forsvinne inn i internett. I denne avsluttende delen av filmen møter vi på former for AI som foreløpig ikke eksisterer i virkeligheten. Den kunstige intelligensen viser menneskelige trekk som følelser, ambisjoner og behov for tilhørighet.

1.2 Roboter og AI -baserte systemer

I følge dictionary.com har ordet *robot* følgende opprinnelse.

Czech, coined by Karel Čapek in the play R.U.R. (1920) from the base robot-, as in robota compulsory labor, robotník peasant owing such labor

I Sebastian Thrun (2004) finner vi to definisjoner av hva en robot er.
In 1979, the Robot Institute of America
defined a robot as "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" (Russell & Norvig,
1995). In contrast, the Merriam Webster's collegiate dictionary (1993) defines
a robot as "An automatic device that performs functions normally ascribed to
humans or a machine in the form of a human."

Vi ser her at Russell og Norvig ser på konseptet robot som en maskin som kan utføre fysiske oppgaver. På den annen side ser vi at Webster definerer en robot som et stykke utstyr som kan utføre oppgaver som vanligvis utføres av mennesker. Hun sier ikke eksplisitt at en robot må gjøre fysiske oppgaver, men hun er nærmere en moderne definisjon av AI ved å peke på at en robot kan gjøre oppgaver som vanligvis utføres av mennesker. En smartassistent vil for eksempel være en robot etter Websters definisjon, men ikke etter Russell og Norvigs, siden den ikke utfører fysiske oppgaver.

I denne oppgaven vil jeg definere en robot som *en AI -maskin som er i stand til å utføre fysiske oppgaver*, i henhold til definisjonen av AI gitt i del 1.1.

Vi ser på en Tesla -bil som et eksempel, for å se hvordan denne passer inn i definisjonen av en robot, og for å se hvordan mennesker interagerer med den. Som en bil er en Tesla i stand til å utføre fysiske oppgaver. Hvis vi tar for oss en selvkjørende versjon av Teslaen, kan vi se at den også er et AI -system og dermed oppfyller begge kravene fra denne oppgavens definisjon av en robot. Grunnlaget for at vi kan kalle en Tesla for et AI -system er at den samler inn og prosesserer data, både gjennom egne sensorer og fra databaser for veitrafikkregler. I tillegg handler den i henhold til forhåndsdefinerte mål som for eksempel, å ikke krasje og å kjøre fra A til B. Den eneste inputen en Tesla krever av sin menneskelige bruker er A og B og så er den i stand til å gjøre resten selv. Den har imidlertid flere feedback

og informasjonssystemer for at brukeren skal føle seg trygg. I tillegg finnes det også mulighet for at brukeren kan overstyre Teslaens handlinger.

1.3 Universell utforming og AI -baserte systemer

I Iwarsson, Susanne, and Agneta Ståhl finner vi følgende definisjon av Universell utforming, skrevet av Mace (1985)
Universal design is an approach to design that incorporates products as well as building features which, to the greatest extent possible, can be used by everyone

Denne definisjonen understreker at universelt design skal skape produkter som kan brukes av alle. Sentralt i fagfeltet Universell utforming finner vi tanken om at det skal designes for alle fra starten av designprosessen. Dette skiller seg fra eldre tankeganger om at systemer kan tilpasses brukere med spesielle behov i en avsluttende fase av systemutviklingen.

AI gir store muligheter når det kommer til å inkludere mennesker med spesielle behov. Hvis vi tar for oss produkter som kan hjelpe hørselhemmede og synshemmede, finner vi mange løsninger som baserer seg på AI -teknologi. AI kan blant annet hjelpe til med følgende oppgaver:

- Hjelp ved navigasjon og rutevalg
    - For eksempel Google Maps
- Hjelp med å identifisere fysiske objekter
    - For eksempel Seeing AI fra Microsoft
- Hjelp med å omforme verbal informasjon
    - Tale til tekst -teknologi

Disse teknologiene kan brukes for at flere mennesker skal føle seg inkludert i samfunnet, og også for å designe systemer med tanke på universell utforming.

Hvis vi ser på AI -teknologi for såkalt Augmented Reality (AR), er dette en bruk av AI som er med på å ekskludere mennesker som av en eller annen grunn ikke har tilgang eller mulighet til å bruke AR. Det kan være eldre mennesker som ikke har nok kunnskap eller ferdigheter til å bruke teknologien, eller for eksempel synshemmede som ikke har mulighet til å ta til seg visuell informasjon.

Hva innebærer det å forstå, hva er forståelse?
Det er vanskelig å teste om en annen person eller maskin forstår et konsept. Derfor ønsker jeg i denne oppgaven å gå ut fra følgende definisjon av *forståelse*:

Forståelse er en følelse av at et konsept gir mening, eller at det passer med andre konsepter man har kjennskap til.

Ved denne definisjonen kan man argumentere for at maskiner kan forstå. Maskiner kan registrere data og sammenligne det med data som er registrert fra før og eventuelt se om det passer inn i et system eller trend. Dette minner om hvordan mennesker tar til seg ny informasjon og forstår den.

Det er likevel forskjeller mellom hva mennesker og maskiner kan forstå. Maskiner kan for eksempel ikke forstå en tanke slik som mennesker, siden en tanke ikke kan representeres i sin helhet med diskrete data. Det er også en forskjell i hvordan mennesker og maskiner forstår, og vi kan finne noe av det i begrepene *følelser* og *mening.* Mennesker sammenligner på en annen måte enn maskiner for å finne ut om noe gir mening. Vi sammenligner ikke nødvendigvis konsepter mot andre konsepter for å finne mening. Følelsen av forståelse åpner også opp for grader av forståelse hos mennesker.

1.4 Retningslinjer for menneske - AI -interaksjon

I Microsofts 18 retningslinjer for interaksjon med AI finner vi blant andre følgende retningslinje:

> *09 Support efficient correction*
> *Make it easy to edit, refine, or recover when the AI system is wrong.*

Her kan vi igjen se på Tesla som et eksempel. Tesla tillater brukeren å ta kontroll over bilen når brukeren selv ønsker det. Dette virker som en god løsning ettersom det ikke er sikkert at Teslaen selv registrerer at det har oppstått en feil. Det gir også brukeren trygghet og hindrer at man havner i en situasjon der brukeren ikke forstår hva bilen gjør, uten muligheten til å gjøre noe med det. Det må ikke ha blitt begått en trafikksikkerhetsmessig feil for at brukeren skal kunne bryte inn. Det er nok at brukeren har lyst til å bryte inn.

Abascal, & Nicolle, C. (2005) presenterer HCI -retningslinjer for inkluderende design. En av retningslinjene sier at man skal minimere maskinautomatikk der det er mulig, slik at brukeren ikke mister følelsen av kontroll. Dette går veldig imot utviklingen i AI -teknologien i dagens samfunn og viser til noe av det samme problemet som *Tone Bratteteig and Guri Verne (2018)* peker på, nemlig at det er vanskelig å holde brukeren i fokus i møte med AI -baserte teknologier. I tillegg har *Abascal, & Nicolle, C. (2005)* et fokus på etikk, personvern og menneskelig autonomi, mens Microsofts retningslinjer fokuserer mer på at brukeren skal holdes informert og forstå hvordan den kan interagere med systemer. Dette finner vi også igjen i *Abascal, & Nicolle, C. (2005)* sine retningslinjer, men de har et bredere syn, og tar også med flere etiske hensyn.

Modul 2

AI -baserte systemer

AI -infused systems, AI -baserte systemer på norsk, defineres av *Amershi et al. (2019)* som:

"Systems that have features harnessing AI capabilities that are directly exposed to the end user."

Her kan det virke som om det er et lite semantisk skille mellom det engelske og norske begrepet. Ordet infused på engelsk kan forstås som at det har eksistert et system der AI

-teknologi har blitt introdusert på et senere tidspunkt. Det norske begrepet *AI-basert system* åpner på en annen side for en forståelse om at AI -teknologien er en av grunnsteinene i systemet. Systemet er basert på det. Hvis vi ser på definisjonen til *Amershi et al. (2019)*, ser vi imidlertid at dette tilsynelatende skillet i semantikk ikke legges vekt på. Det er med andre ord ikke viktig på hvilken måte AI blir integrert i systemet. Det essensielle i definisjonen er at brukeren opplever kontakt med AI -teknologi.

Yang et al. (2020) anser *Amershi et al. (2019)* sin definisjon, blant flere andre, for å være dårlig formulert. Det som mangler i denne definisjonen er en definisjon av hva AI, og AI -evner, er. *Yang et al. (2020)* peker også på at det i forskningen på fagfeltet ofte ikke kommer klart frem om man snakker om teknologien som ligger bak AI, eller om opplevelsen en bruker kan ha om å interagere med en form for kunstig intelligens. For å utforme god interaksjon med AI -baserte systemer er det viktig å vite om når man snakker om den tekniske siden av AI, eller om brukeropplevelsen. Som interaksjonsdesignere er vi mest interessert i hvordan AI -oppleves, men det er viktig å ha en forståelse for teknologien som ligger bak og gjør at man kan snakke om en AI -brukeropplevelse.

Yang et al. (2020) peker på to egenskaper ved AI-teknologi som gjør det utfordrende å designe AI -baserte systemer.

- Capability uncertainty
- Output complexity

Capability uncertainty går ut på at det er vanskelig å vite nøyaktig hva et AI-basert system er i stand til å gjøre, hvilke oppgaver det er i stand til å utføre. Grunnen til dette er at AI-teknologi baserer seg på store datasett, og det lar seg ikke alltid gjøre å danne et overblikk over mulighetene som finnes basert på denne informasjonen. I tillegg kan et AI -basert system endre seg når det får tilgang på nye datasett.

Output complexity handler om at enkelte AI -baserte systemer kan gi veldig mange forskjellige resultater og dermed også mange forskjellige feil. I og med at systemet er komplekst og vanskelig å ha full oversikt over blir det også vanskelig for brukeren å forstå hvorfor man får et gitt resultat, spesielt når resultatet er noe man kanskje ikke har sett før eller ikke forventet.

Noe som også bidrar til at det kan være vanskelig for brukere å forholde seg til AI -baserte systemer er deres probabilistiske natur. Som *Amershi et al. (2019)* forklarer, kan AI- baserte systemer gi forskjellig output ved lik input, noe som kan være vanskelig for brukere å forholde seg til.

Google Translate
Basert på definisjonen over kan man anse *Google Translate* (GT) som et AI -basert system. GT tar input i form av tekst eller tale på et gitt språk og gir output i form av en oversettelse av inputen til et annet språk. For å se nærmere på GT som et AI-basert system kan vi analysere det i henhold til egenskapene som er beskrevet over.

- GT baserer seg på **store datamengder og probabilistiske metoder** for å gi en oversettelse som er sannsynlig basert på inputen.

- GT gir brukeren en opplevelse av sine AI-evner ved at den oversetter på en måte som er mer **komplisert og sammensatt** enn det man kan oppleve ved å benytte seg av et tradisjonelt oppslagsverk.

- GT **utvikler seg over tid**. Med et stadig voksende datagrunnlag er GT i stand til å oversette mer presist enn det var for noen år siden.

- GT kan gi veldig **variert output**. Man finner variasjon både i setningsoppbygging, ordvalg og GT kan også gi output som både tekst og lyd.

Det at GT bruker probabilistiske metoder og gir variert output kan virke forvirrende for brukeren. Når man oversetter noe har man en viss forventning om hva man skal få tilbake, i de tilfellene dette ikke skjer er det vanskelig for brukeren å forholde seg til. En oversettelse bør for eksempel stemme med grammatikken i språket, så det er veldig lett for et menneske som kjenner språket å vurdere om en oversettelse gir mening eller ikke. Et annet eksempel er at man som bruker kanskje vil endre på detaljer i teksten, hvor man da forventer at endringen i sluttresultatet også skal være liten. Dette er ikke alltid tilfelle med GT som benytter en probabilistisk algoritme.

GT har hatt et dårlig rykte på seg og blitt latterliggjort for feil i oversettelsene sine. Det er derfor mange som ikke anser GT og andre automatiserte oversettelsestjenester som en gode og fraråder å bruke dem *(Groves, Mundt (2015))*. Dette er et eksempel på at menneskelige brukere skaper seg et førsteinntrykk og holder seg ved det, selv om GT som et AI-basert system er i stadig utvikling. Det samme gjelder forsåvidt til en viss grad også når det kommer til menneskers møter seg imellom. Førsteinntrykk er viktige for oss men det er ikke alltid hensiktsmessig å basere seg på et førsteinntrykk, verken i møte med mennesker eller AI -basert teknologi, nettopp fordi begge forandrer seg over tid.

En annen ting som kan gjøre det vanskelig for brukere å forholde seg til GT er at det er vanskelig å forstå hvordan google kommer frem til en gitt oversettelse. I oversetting av naturlig språk vil det ofte være varianter av oversettelser i og med at strukturen i ulike språk ikke er den samme. GT viser ikke disse variantene og det er da vanskelig å forstå hvorfor oversettelsen ser ut akkurat som den gjør. Det finnes funksjoner i GT som hjelper på dette problemet. Det finnes for eksempel et ikon som viser at en gitt oversettelse er kontrollert av menneskelige bidragsytere. Det virker likevel som om dette er gjort med tanke på å samle inn mer input til GT og ikke av hensyn til brukervennlighet. Det samme kan sies om GT sin tilbakemeldingsfunksjon, som lar brukeren si hva han eller hun synes om en gitt oversettelse. Den bidrar til at kvaliteten på GT sine oversettelser kan forbedres, men den gjør ikke GT mer brukervennlig.

Amershi et al. (2019) presenterer 18 retningslinjer for utforming av AI -baserte systemer. Disse retningslinjene baserer seg på en gjennomgang av forskning på området, i tillegg til ekspertvurdering og en brukerstudie hvor designere har fått gjøre sin egen vurdering av retningslinjene etter å ha fått kjennskap til retningslinjene i sammenheng med en rekke forskjellige AI -baserte systemer. *Amershi et al. (2019)* har utformet retningslinjene med fokus på generalitet, men presiserer at det til en viss grad går på bekostning av spesialitet og at retningslinjene ikke uten videre kan brukes på absolutt alle AI -baserte systemer.

Retningslinjene er med andre ord ment som et utgangspunkt og noe man kan basere mer spesialiserte retningslinjer på.

Kocielnik et al. (2019) undersøker forventninger til AI -baserte systemer og hvordan det AI -baserte systemets ytelse påvirker forholdet brukerne får til det. De peker på at mange AI -baserte systemer er utformet for høy presisjon. Det vil si at de forsøker å unngå å gi brukeren feilaktige resultater og heller vil holde tilbake resultater som potensielt kunne vært nyttige for brukeren. Dette gir brukeren inntrykk av at det AI -baserte systemet gjør få feil, fordi de resultatene brukeren ser har større sjanse for å være korrekte.

I studien vises det at et fokus på høy *recall* gir større grad av akseptanse og oppfylling av forventningene til brukerne. Et slikt fokus inkluderer flere resultater, men eksponerer dermed også brukeren for flere feil. Det argumenteres for at det i mange tilfeller er mindre strevsomt for en bruker å vurdere resultatene og luke ut disse feilene, sammenlignet med å finne frem til resultatene manuelt dersom de har blitt utelukket av det AI -baserte systemet.

Dette gjenspeiles også i retningslinje G10 fra *Amershi et al. (2019)*. Den sier at et AI -basert system skal ha en metode for å løse opp i uklarheter dersom det er tvil rundt brukerens intensjon. Dette finner vi ikke hos GT, som kun gir en enkelt oversettelse for hver input brukeren gir. Det anbefales også at systemet skal innrømme feil dersom det ikke er i stand til å gi et tilfredsstillende resultat. GT gjør heller ikke dette. Dersom man skriver inn tekst i GT som ikke gir mening, eller refererer til et ekte ord, vil GT likevel bare oppgi et resultat uten videre forklaring. Det ville tilsynelatende vært enkelt å implementere dette ved at brukeren får tilbakemelding om at systemet ikke forstår den gitte inputen. Det ville heller ikke vært vanskelig å la brukeren velge mellom flere alternativ i tilfeller der flere resultater dukker opp som sannsynlige. Man kan spekulere i hvorfor en slik funksjonalitet ikke allerede finnes i GT. Kanskje det er på grunn av at utviklerne ønsker at GT skal fremstå som et system uten feil eller svakheter, eller kanskje det er et resultat av prioritering av større grad av automatisering fremfor brukervennlighet.

Retningslinje G2 gjør oss oppmerksom på en annen svakhet ved GT. Den sier at det skal komme klart frem hvor godt systemet er i stand til å gjøre det det gjør. Som nevnt over utvikler GT seg over tid og var tidligere kjent for mange feil. Selv om GT har forbedret seg finnes det fortsatt store muligheter for å få feilaktig resultat. Dette informeres imidlertid ikke om og dette gjør det vanskelig for brukeren å stole på GT og vite hvordan systemet kan brukes best mulig. En slik funksjonalitet virker vanskeligere å implementere i GT ettersom det er vanskelig også for designere å forstå hvordan GT kommer frem til en gitt oversettelse. Så dette kan være en utfordring og mulighet for utvikling hos GT. Hvordan kan man veilede brukerne til å forstå mulighetene og begrensningene ved systemet?

Bender et al. (2021) skriver om AI -baserte systemer som baserer seg på store språkmodeller. I forskningsartikkelen påpekes det at slike systemer bringer med seg en rekke utfordringer og etiske problemstillinger. For det første så er dette en ressurskrevende teknologi som krever store databaser for å lagre disse språkmodellene. Dette bringer med seg utfordringer både når det kommer til miljø og økonomi. Driften av serverne til disse databasene krever mye energi, og man kan spørre seg om det er moralsk å sette av så mye ressurser til en slik oppgave. I tillegg er tilgangen på store nok servere begrenset av

økonomi, og det er dermed ikke alle som har tilgang til å forske på denne teknologien, noe som igjen kan føre til prioritering av språk og folkegrupper og diskriminering av andre.

For det andre kan det være problematisk etisk å benytte teknologi som etterligner mennesker. I forskningen sin påpeker *Bender et al. (2021)* at dette kan føre til problemer rundt hvem som er ansvarlig for informasjon og utsagn. Når AI -generert språk og beskjeder kan blandes inn og forveksles med menneskers ytringer blir det fort veldig vanskelig å ha kontroll over hvilken informasjon man sprer og hvor den kommer fra. Dette kan igjen bidra til en polarisering av meninger og av samfunnet. Vi ser allerede eksempler på dette i sosiale medier der heuristikken virker å fremme et budskap som brukeren allerede er mottakelig for, og dermed potensielt lede brukeren videre i retning mer ekstreme synspunkter. *(Tucker et al. 2018)*

Følstad og Brandtzæg (2017) diskuterer muligheter og utfordringer rundt utforming av chatboter. De forklarer at når man designer en chatbot, designer man i hovedsak en samtale. Dette endrer kravene til verktøykassen til interaksjonsdesignere. Tidligere har man fokusert på visuelle brukergrensesnitt og mulighetene som ligger i tilpasning av disse. I design av chatboter mister man noen av disse mulighetene, og blir nødt til å tenke på en ny måte. Nå får verbal kommunikasjon en enda større rolle. Man er nødt til å designe for hvordan brukerens ønsker kan bli hørt og forstått. Dette står i kontrast til hvordan man i lang tid har designet systemer som man har måttet introdusere og forklare for brukeren. Med chatboter designer man for at brukeren skal kunne ytre sine ønsker gjennom naturlig språk. Dette byr naturligvis også på utfordringer. En chatbot bør for eksempel være i stand til å lede en samtale tilbake på rett spor dersom det oppstår misforståelser.

I tillegg kan det oppstå en del etiske problemstillinger som man må ta høyde for når man designer en chatbot. Fokus på personvern er ikke unikt for design av chatboter, men er absolutt en del av bildet også her. Man kan tenke seg at brukere gjennom naturlig språk vil dele personopplysninger i visse tilfeller uten å være bevisst det selv. Der man tidligere var nødt til å fylle ut skjemaer med informasjon vil man som bruker kunne dele denne informasjonen ved bruk av naturlig språk med en chatbot. Dette kan potensielt føre til at brukeren får et mer avslappet forhold til å dele sensitiv informasjon, og dette er noe man som designer bør være bevisst på i utformingen av en chatbot. I tillegg oppstår også utfordringen som *Bender et al. (2021)* påpekte også med chatbots, nemlig at de etterligner menneskelig samtale. Og at det dermed blir vanskelig for brukeren å vite hvem som står bak informasjonen. Det kan virke som om et viktig element for å adressere disse etiske utfordringene er opplysning og bevisstgjøring av brukeren.

Hvis vi ser på retningslinjene G1 og G2 fra *Amershi et al. (2019)* ser vi at disse er aktuelle også for chatboter. G1 sier at det skal komme tydelig frem for brukeren hva systemet kan gjøre. Dette vil bidra til å redusere antall misforståelser og hjelpe brukeren til å ytre sine ønsker i henhold til begrensningene til systemet. En chatbot kan gi inntrykk av å kunne kommunisere på lik linje som et menneske, så det er viktig at man hjelper brukeren med å danne mer presise forventninger til hva systemet kan gjøre. G2 presiserer at det også er viktig å informere om hvor godt systemet gjør det det gjør. Dette er også med på å sette fornuftige forventninger til systemet hos brukeren.

Det kan imidlertid se ut som det hadde vært bruk for en annen retningslinje for bevisstgjøring rundt hvordan et AI -basert system deler og lagrer informasjon. Dette kommer ikke klart frem i retningslinjene til *Amershi et al. (2019)*, men det kan være gunstig å informere om hvem som har skrevet meldingene til chatboten, hva slags informasjon man bør være forsiktig med å dele, og hvordan dataene lagres. På denne måten kan brukeren vurdere informasjonen som blir gitt selv og være mer bevisst hva slags informasjon den selv deler.

Referanser:

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Teevan, J. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (paper no. 3). ACM. (https://www.microsoft.com/en-us/research/uploads/prod/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf)

Kocielnik, R., Amershi, S., & Bennett, P. N. (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. (https://www.microsoft.com/en-us/research/uploads/prod/2019/01/chi19_kocielnik_et_al.pdf)

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In Proceedings of the 2020 CHI conference on human factors in computing systems (Paper no. 164). (https://dl.acm.org/doi/abs/10.1145/3313831.3376301)

Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. interactions 24, 4 (July-August 2017), 38–42. https://doi-org.ezproxy.uio.no/10.1145/3085558

Michael Groves, Klaus Mundt,
Friend or foe? Google Translate in language for academic purposes,
English for Specific Purposes,
Volume 37,
2015,
Pages 112-121,

Tucker, Joshua Aaron and Guess, Andrew and Barbera, Pablo and Vaccari, Cristian and Siegel, Alexandra and Siegel, Alexandra and Sanovich, Sergey and Stukal, Denis and Nyhan, Brendan, Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature (March 19, 2018). Available at SSRN: https://ssrn.com/abstract=3144139 or http://dx.doi.org/10.2139/ssrn.3144139

Kilder:

Grudin, Jonathan. "AI and HCI: two fields divided by a common focus." *AI Magazine*, vol. 30, no. 4, winter 2009, pp. 48+. *Gale Academic OneFile*, link.gale.com/apps/doc/A216267417/AONE?u=oslo&sid=bookmark-AONE&xid=d85837d1. Accessed 13 Sept. 2022.

Schuett, Jonas. (2019). A Legal Definition of AI. SSRN Electronic Journal. 10.2139/ssrn.3453632.

Antsaklis, & Rahnama, A. (2018). Control and Machine Intelligence for System Autonomy. Journal of Intelligent & Robotic Systems, 91(1), 23–34. https://doi.org/10.1007/s10846-018-0832-6

https://www.dictionary.com/browse/automatic

Tone Bratteteig and Guri Verne. 2018. Does AI make PD obsolete? exploring challenges from artificial intelligence to participatory design. In Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2 (PDC '18). Association for Computing Machinery, New York, NY, USA, Article 8, 1–5. https://doi-org.ezproxy.uio.no/10.1145/3210604.3210646

https://ai.facebook.com/research/no-language-left-behind/

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

Mace R. Universal Design, Barrier-Free Environments for Everyone. Los Angeles: Designers West, 1985.

Iwarsson, Susanne, and Agneta Ståhl. "Accessibility, usability and universal design—positioning and definition of concepts describing person-environment relationships." *Disability and rehabilitation* 25.2 (2003): 57-66.

Abascal, & Nicolle, C. (2005). Moving towards inclusive design guidelines for socially and ethically aware HCI. Interacting with Computers : the Interdisciplinary Journal of Human-Computer Interaction, 17(5), 484–505. https://doi.org/10.1016/j.intcom.2005.03.002

Montemayor, C., Halpern, J. & Fairweather, A. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. AI & Soc 37, 1353–1359 (2022). https://doi-org.ezproxy.uio.no/10.1007/s00146-021-01230-z

Vedlegg 1. Tilbakemeldinger

Tilbakemeldingene jeg har fått i løpet av arbeidet med de individuelle oppgavene har satt lys på enkelte punkter i teksten min med potensiale for forbedring. Tilbakemeldingene har ikke vært kritiske til utformingen eller innholdet i sin helhet, og jeg har derfor ikke gjort store endringer på dette underveis. Det har imidlertid vært nyttig å bli gjort oppmerksom på konkrete mangler i enkelte argumenter og eksempler.