Universitetet i Oslo
IN5620
Høst 2024

Modul 2: Individuell innlevering

KUNSTIG INTELLIGENS: EN INNFØRING

Av: Mari Reisjå
Epost: marireis@uio.no
Antall tegn: 24 431

Innholdsfortegnelse

1 Introduksjon **1**
2 Kunstig intelligens: Hva er det? **1**
2.1 Tre definisjoner og et forsøk 1
2.2 Autonome maskiner 2
2.3 KI-integrerte systemer 3
3 KI i forskning, praksis og på film **5**
4 KI i studenthverdagen: En øvelse **6**
5 Roboter og KI **7**
6 Design av KI **8**
6.1 Universell utforming 8
6.2 Menneske-KI interaksjoner og bruk av store språkmodeller 9
7 Konklusjon **11**
Referanser **13**
Vedlegg **15**
Vedlegg 1: Endringer etter tilbakemelding 15

1 Introduksjon

Begrepet kunstig intelligens (KI) popper opp overalt om dagen (Zewe, 2023), men hva innebærer det egentlig? Denne teksten gir en innføring i teknologien og design av interaksjoner med KI-systemer.

Først plasserer jeg KI i historisk kontekst. Jeg trekker så frem tre definisjoner, før jeg gjør et forsøk på å skape min egen definisjon. Deretter ser jeg på begrepene automatisk og autonomi, og hva vi egentlig legger i et KI-integrert system. Jeg drøfter fremstillingen av KI i forskning, praksis og populærkultur, og reflekterer roboter i sammenheng med KI. Videre diskuterer jeg prinsipper for universell utforming og retningslinjer for brukervennlig design av KI-systemer, og vurderer hvordan disse kan påvirke designet av interaksjoner – spesielt for avanserte språkmodeller. Avslutningsvis samler jeg trådene: Har refleksjonene ført meg nærmere et svar på hva KI egentlig er? Og hvilke implikasjoner har dette for design av interaksjon med slike systemer?

2 Kunstig intelligens: Hva er det?

"Can machines think?" Slik startet Alan Turing sin berømte artikkel "Computing Machinery and Intelligence" fra 1950. Han undersøkte om maskiner kunne gi svar som var umulige å skille fra de som mennesker ville gi. Med andre ord: *Kunstig intelligens er ikke noe nytt.*

Feltet ble etablert som en akademisk disiplin i 1950-årene (Bratteteig og Verne, 2018, s. 1), og selve begrepet *kunstig intelligens* ble lansert av John McCarthy i 1956 (Haenlein og Kaplan, 2019, s. 7). Siden har feltet gått gjennom opp- og nedturer, med betydelige investeringer som har gitt varierende resultater (Grudin, 2009). Det har tatt tid, men de siste årenes utvikling blir av flere beskrevet som en revolusjon (Følstad og Brandtzæg, 2017, s. 1; Khogali og Mekid, 2023, s. 1).

2.1 Tre definisjoner og et forsøk

Det finnes ingen standard definisjon av KI. McCarthy definerte i 1956 KI som *"the science and engineering of making intelligent machines"* (Manning, 2020). Grudin (2009, s. 51) siterer McCarthy på å ha skrevet at KI var en gren av matematikken, ikke psykologien. Han mente ikke nødvendigvis at KI skulle etterligne en menneskelig måte å løse problemer på, men løse problemer for oss basert på matematikk og logikk.

Wang (2019) på sin side tar tak i ordet "intelligens" og hva det egentlig betyr i KI-sammenheng. Han foreslår å definere det som *"adaptation with insufficient knowledge and resources"*. Denne definisjonen fremhever at teknologien ikke er altvitende, men at beslutninger tas basert på tilgjengelig informasjon og ressurser.

En tredje definisjon er nylig presentert av EU. I disse dager utformer de et KI-regelverk, og i et utkast står det at KI er *"a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments"* (Digdir, u.d.). Denne definisjonen legger vekt på hvordan KI-systemer fungerer i praksis, og systemenes evne til å lære og tilpasse seg mens de utfører oppgaver som kan påvirke både den fysiske og digitale verden.

I et forsøk på å slå elementer fra disse tre sammen til én helhetlig definisjon, kan KI beskrives som *"teknologi som kan påvirke verden, både digitalt og fysisk, gjennom å lære, tilpasse seg og løse problemer mer eller mindre autonomt med varierende tilgang på ressurser og informasjon."*

Denne definisjonen inkluderer EUs fokus på hvordan teknologien fungerer og påvirker omverdenen. Ved å si at teknologien gjør dette "mer eller mindre autonomt", understrekes McCarthys syn på at KI ikke nødvendigvis handler om å bygge en menneskelig hjerne – men å bygge systemer som kan løse problemer på egenhånd. Wangs poeng om begrenset tilgang på informasjon reflekteres i siste setning.

2.2 Autonome maskiner

I mange KI-definisjoner står begrepet *autonomi* sentralt. Autonomi refererer til "en persons evne til selv å bestemme sine handlinger og grunnlaget for sine handlinger." (Sagdahl, 2019). Det er relevant å diskutere forskjellen på menneskelig autonomi og en maskins autonomi, ikke minst hvor autonomt en maskin egentlig kan handle. Bestemmer virkelig maskinen over sine egne handlinger hvis systemet er programmert av et menneske? Her vil jeg trekke inn differansen mellom begrepene "automatisk" og "autonom". Når vi snakker om maskiner, kan vi si at skillet ligger i graden av menneskelig kontroll: Et system som gjør noe automatisk er mer menneskestyrt enn et system som kan gjøre noe autonomt – altså med en form for selvbestemmelse.

Yang et al. (2020) kategoriserer KI-systemer i fire nivåer, hvor det fjerde og mest avanserte nivået kjennetegnes av komplekse utfall og usikkerhet rundt systemets kapasiteter. Disse systemene er i kontinuerlig utvikling, adaptive og i stand til å lære fra nye data etter implementering. Jeg vil argumentere for at disse systemene nærmest kan betraktes som autonome maskiner, fordi deres handlinger kan komme ut av skapernes kontroll. Akkurat som et menneskes autonomi kan begrenses, for eksempel ved fengsling, kan maskiners autonomi begrenses ved å bli tatt ut av drift. Et viktig spørsmål for fremtiden er derfor hvordan, og av hvem, denne overvåkningen skal skje.

2.3 KI-integrerte systemer

Selv om KI i 2024 fortsatt føles nytt for mange, har de fleste av oss brukt teknologien en hel del allerede – kanskje uten å være helt klar over det selv. Hverdagslige ting som anbefalinger gitt av strømmetjenester, treff i Google-søk, GPS-er i biler eller ansiktgjenkjenning for å låse opp mobilen er gode eksempler. Jeg tør å tippe at få tenker over at dette er KI-drevet.

Amerishi et al. (2019, s. 1) definerer AI-integrerte systemer som "systems that have features harnessing AI capabilities that are directly exposed to the end user". I og med at det er ulike definisjoner på hva KI er, er det også diskusjoner om hva som kan betegnes som et KI-integrert system. Amershi et al. (2019), Kocielnik, Amershi og Bennett (2019) og nevnte Yang et al. (2020) er tre forskningsartikler på feltet som har hatt stor gjennomslagskraft de siste årene. Ut fra disse tre artiklene, vil jeg trekke frem fire sentrale kjennetegn disse forskerne til en viss grad er enige om at systemene innehar.

1: Tilpasningsdyktighet og mulighet til å lære fra data

AI-integrerte systemer skiller seg fra andre systemer med nettopp dette at de er designet for å utvikle seg. Amershi et al. (2019, s. 3) presenterer 18 retningslinjer for design av AI-integrerte systemer. Flere av disse omhandler systemenes evne til å tilpasse seg basert på kontekst og dataene de forholder seg til, noe som åpner for dynamiske resultater og bruk. Yang et al. (2020) er opptatt av hvordan AI-systemene varierer fra enkel til svært kompleks tilpasning til omgivelsene.

2: Evne til å samarbeide med mennesker

Kocielnik et al. (2019) undersøker hvorvidt brukere vil akseptere uperfekte AI-systemer, og hvordan man kan designe for at brukernes forventninger stemmer med hva systemet faktisk kan gjøre. Forfatterne understreker viktigheten av å balansere hva systemet gjør automatisk, og hva brukeren har kontroll over. De fremhever det som viktig at systemet oppleves å være til hjelp for å bli akseptert. Nettopp det *å være til hjelp* kan karakteriseres som et kjennetegn for KI-integrerte systemer. I alle de tre artiklene trekkes det frem at for at systemene skal lykkes, bør de designes for å bistå menneskelige prosesser, ikke nødvendigvis for å erstatte menneskene helt. Amershi et al. (2019, s. 1) er opptatt av at systemene må designes slik at brukerne forstår dem for å ikke bryte med tradisjonelle retningslinjer for interaksjonsdesign, og flere av de 18 retningslinjene handler konkret om dette.

3: Dyktige på personalisering

Evnen til å tilpasse seg brukeren er ofte god i AI-integrerte systemer. Systemene lærer fortløpende av brukeren, og skreddersyr tjenestene sine til brukerens behov og vaner. Amershi et al. (2019, s. 1) understreker at AI-teknologiens personaliseringsevne også er en mulig designfelle. Det kan oppleves som positivt at systemet kjenner deg som bruker godt, men hva er egentlig kostnaden? Hvilken informasjon har brukeren delt for å oppnå dette, og er brukeren klar over det selv?

4: Evne til å rette sine feil

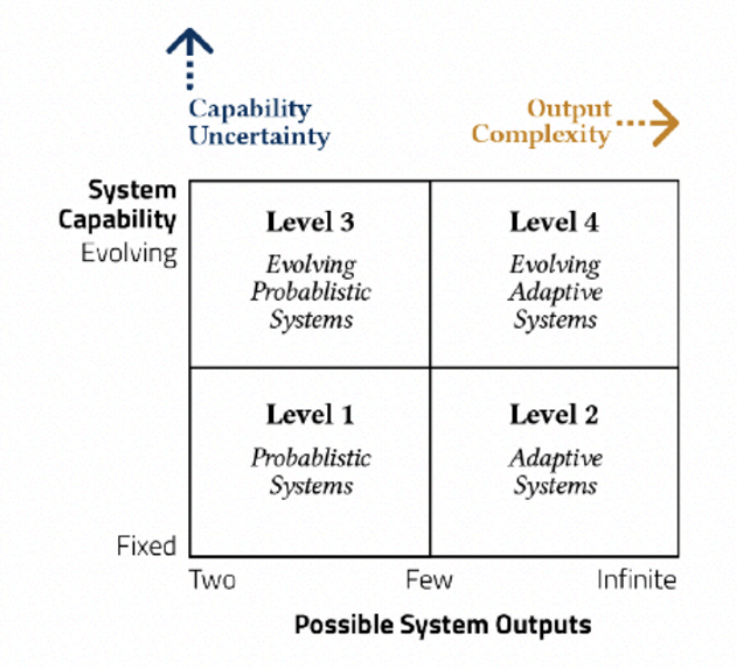
Forfatterne bak de tre artiklene tar opp KI-systemers evne til å rette egne feil. Kocielnik et al. (2019) er opptatt av hvilke feil brukerne vil akseptere, mens Amershi et al. (2019) har retningslinjer for at systemene bør være åpne om at de gjør feil og klare å oppdatere seg om brukeren varsler om feil. Yang et al. (2020, s. 7) spør seg hvordan systemene kan designes for at brukerne kan komme seg etter feil på en smertefri måte.

Generative KI-systemer som ChatGPT har blitt utbredt den siste tiden. Dette er et eksempel på et tydelig KI-drevet system som har alle disse kjennetegnene. Den nyeste versjonen av ChatGPT kan lære forløpende fra dokumenter og informasjon på internett. Den samarbeider med brukeren, og kan komme med innspill underveis i en dialog. Den er ekstremt god på personalisering, og brukeren kan konkret be den om å endre på måten den svarer på. Systemet kan til en viss grad rette sine feil, og vil alltid gjøre et nytt forsøk om det blir gjort klar over feil. Et stort problem med alle store språkmodeller, som er teknologien ChatGPT bygger på, er faren for at systemet finner på ting og hallusinerer – gjerne formulert svært

overbevisende (McIntosh et al., 2023, Rawte et al., 2023). Dette med å rette sine feil innebærer derfor store muligheter for forbedring.

3 KI i forskning, praksis og på film

Yang et al. (2020) har satt en bransjestandard med sin kategorisering av KI-systemer. Artikkelen handler om hva som gjør systemene spesielle, og hvorfor og hvordan designere og forskere må tenke helt nytt i møte med dem. Forfatterne presenterer et rammeverk som kan brukes til å analysere og adressere utfordringene knyttet til design av menneske-KI interaksjoner. Klassifisering i de fire nivåene utgjør essensen i dette rammeverket, og baseres på systemenes evne til å generere utfall og til å lære av nye brukerdata.



Figur 1: De fire nivåene presentert av Yang et al. (2020, s. 8)

Sett i lys av KI-systemene vi ser lanseres i dag, vil jeg argumentere for at det er på tide å diskutere om disse kategoriene bør redefineres eller om det bør legges til ytterligere nivåer. Et eksempel er at systemer som den første versjonen av ChatGPT har blitt plassert i nivå fire, det mest komplekse nivået, selv om den var begrenset til å være trent på "gammel" data publisert før 2021. Den nyeste versjonen av ChatGPT har løpende tilgang til alle åpne kilder online og har funksjoner på et høyere nivå både når det gjelder kompleksitet i utfall og å lære av brukerdata underveis.

Uansett hvor mye forskere forsøker å klassifisere, forklare og forstå KI-teknologien, er det sjelden slik systemene fremstilles for brukerne. Kommersielle aktører har en tendens til å

reklamere med hva systemene kan gjøre for brukerne, ikke hvordan de gjør det. ChatPDF er et godt eksempel på et slikt system. Når en bruker først ankommer nettsiden, er den teksten som kommer opp: “Chat with any PDF. Join millions of students, researchers and professionals to instantly answer questions and understand research with AI” (ChatPDF, u.d.). Så kan brukeren bare sette i gang med å laste opp en PDF-fil. Man må virkelig lete på nettsiden for å lære mer om hvordan dette faktisk fungerer.

En tredje type fremstilling av KI finner man i populærkulturen, spesielt science fiction. Opp gjennom tidene har kunstige skapninger med en slags menneskelig intelligens spredt frykt og dystopi. I romanen Frankenstein, skrevet av Mary Shelly i 1818, skaper en vitenskapsmann et monster som kommer ut av hans kontroll og etterhvert hevner seg på sin skaper (Korslund, 2024). Samtidig finnes det en rekke eksempler på snille systemer, som R2-D2 i Star Wars. R2-D2 er en snill og modig robot som er med på å redde galaksen fra de ondes klør. Dette kan ha påvirket hvordan folk flest anser KI i dag.

4 KI i studenthverdagen: En øvelse

Som tidligere nevnt, har mange av oss trolig brukt KI-teknologi uten å være helt klar over det selv. Med inntoget av stadig flere nye systemer som åpenbart er drevet av KI, vil mange i fremtiden ta teknologien aktivt i bruk, vel vitende om at man får hjelp fra KI. Dette gjelder også for studenter. I denne oppgaven er jeg som student oppfordret til å teste og reflektere over ulike verktøy for læring – med og uten bruk av KI.

Jeg har notert med penn og papir i tre minutter i to dager, med tastatur de to neste dagene og ved hjelp av lydopptak de siste to. Å notere med penn og papir involverte ikke KI, mens jeg brukte KI de neste dagene i form av automatisk retningsvingprogram og transkribering via programmet My Good Tape. Det jeg har reflektert mest over i etterkant av øvelsen, er hvilken aktivitet jeg husker best. Jeg har kommet frem til at det er det jeg skrev på papir. Jeg tegnet noen små tegninger og brukte ulike farger, noe som hjelper meg å huske. Øvelsen jeg husker minst, er lydopptaket som jeg transkriberte. Jeg tror grunnen er at jeg ikke selv jobbet aktivt med teksten. Samtidig er det trolig de digitale tekstene det er lettest å dele med andre, og jeg kan ta i bruk andre KI-systemer som UiO-GPT for å få hjelp til å forbedre tekstene og tydeliggjøre budskapet. Mine erfaringer antyder at selv om KI kan effektivisere prosesser, kan det også redusere den kognitive involveringen i selve skriveprosessen, som igjen påvirker læringsutbyttet.

5 Roboter og KI

Jeg skal være helt ærlig: Når noen sier ordet "robot", ser jeg for meg R2-D2. Dette er imidlertid ofte langt fra virkeligheten.

Selve ordet *robot* kommer fra tsjekkisk og betyr arbeider eller slave. I 2018 beskrev Dautenhahn hvordan vår oppfatning av roboter har endret seg betraktelig de siste 20 årene, og at det er vanskelig å forutsi hva vi vil kalle roboter om 50 år. Thrun (1997) understreker at definisjonene av roboter varierer, og trekker frem Det amerikanske robotinstituttets definisjon fra 1979: *"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"*. Dagens definisjon fra Store Norske Leksikon (Søraa, 2022) fokuserer på roboter som maskiner som bruker sensorer og algoritmer til å bearbeide sanseinntrykk fra omgivelsene for å utføre oppgaver. Forskjellen mellom disse to definisjonene viser at roboter har gått fra å være rene verktøy for fysisk arbeid til systemer som kan tilpasse seg omgivelsene, noe som knytter dem nærmere KI. Definisjonene legger fortsatt vekt på fysiske maskiner, noe som kan føre til at systemer som chatbots faller utenfor. Jeg vil derfor driste meg til å foreslå en bredere definisjon av roboter som inkluderer både fysiske og digitale systemer: *"En robot er en maskin eller en tjeneste, fysisk eller digital, som ved hjelp av kunstig intelligens eller annen forhåndsprogrammert logikk kan utføre oppgaver mer eller mindre autonomt, basert på sanseinntrykk eller data fra omgivelsene. Moderne roboter er i stand til å lære, tilpasse seg og reagere ytre faktorer."*

Denne definisjonen er dekkende for tidlige, mekaniske roboter, men også moderne systemer som chatbots og selvkjørende biler. Et problem med definisjonsforslaget, er at den kan viske ut grensene mellom hva som er en robot og hva som er et KI-system. Hvor plasserer vi for eksempel en chatbot?

Mitt forsøk på å definere KI	**Mitt forsøk på å definere roboter**
Kunstig intelligens (KI) er teknologi som kan påvirke verden, både digitalt og fysisk, gjennom å lære, tilpasse seg og løse problemer mer eller mindre autonomt med varierende tilgang på ressurser og informasjon.	En robot er en maskin eller en tjeneste fysisk eller digital, som ved hjelp av kunstig intelligens eller annen forhåndsprogrammert logikk kan utføre oppgaver mer eller mindre autonomt, basert på sanseinntrykk eller data fra

	omgivelsene. Moderne roboter er i stand til å lære og tilpasse seg og reagere på ytre faktorer.

Figur 2: Likheter mellom mine to definisjoner.

Jeg vil argumentere for at en robot kan defineres som selve maskinen eller tjenesten, som kan være basert på KI, men ikke nødvendigvis. En chatbot er for eksempel et system som er basert på en KI-teknologi i form av en språkmodell. Et praktisk eksempel er at brukeren selv måtte programmere ruten til de første robotstøvsugerne, og om støvsugeren møtte et hinder, så stoppet den. Den siste robotstøvsugeren fra produsenten Ecovacs bruker KI for å unngå hindringer og legge en plan for å gjøre huset rent (Elkjøp, u.d.).

6 Design av KI

6.1 Universell utforming

Bufdir definerer universell utforming som design som er tilgjengelig for alle uten behov for spesialtilpasninger (Bufdir, u.d). Selv om dette er idealet, er det utfordrende i praksis. Her kan KI spille en viktig rolle med sin evne til å sømløst tilpasse seg brukere (Hyatt og Owenz, 2024; Pisoni et al., 2021). Eksempelvis kan systemer for tekst til tale hjelpe personer med nedsatt syn. Programmer som Adobe Acrobat har denne funksjonen tilgjengelig for alle, synshemmede eller ei. Teknologien bak "Seeing AI" er et hjelpemiddel som potensielt kan integreres i en rekke systemer for å gjøre dem mer universelle. Her brukes kameraet på telefonen til å beskrive omgivelsene. Det finnes også mange eksempler på KI-systemer som kan bistå personer med bevegelses- eller kognitive utfordringer, som for eksempel stemmestyrte assistenter og automatiserte hjelpemidler.

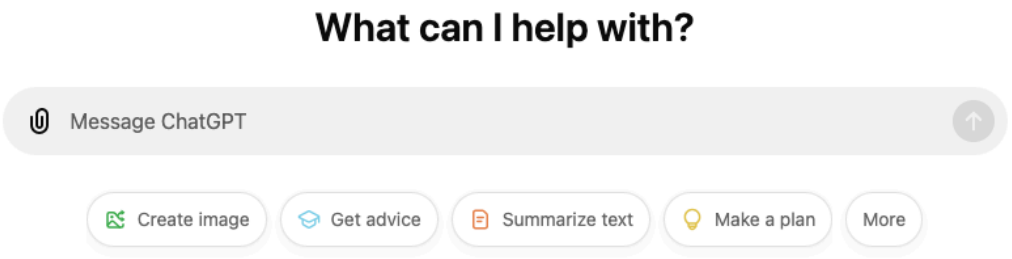
Selv om KI åpner opp for sømløs inkludering, er eksemplene på at det fører til ekskludering mange. Forskning har blant annet vist at KI har problemer med å gjenkjenne mørkhudede mennesker, siden systemene ofte er trent på bilder av hvite personer (Holmes-Iverson, 2023). Dette problemet fører diskusjonen inn på begrepene "forstå" og "forståelse", og spørsmålet om hva en maskin egentlig kan forstå. Som Wang (2019) fremhever i sin definisjon av KI, er intelligensen til disse systemene begrenset til informasjonen systemet har til rådighet. Når systemet ikke er trent opp på bilder av mørhudede mennesker, mangler

det også evnen til å tolke slik informasjon. Dette illustrerer hvorfor det er viktig å utvikle gode retningslinjer for design av KI. De seneste årene har flere prøvd på nettopp dette.

6.2 Menneske-KI interaksjoner og bruk av store språkmodeller

Det har vært og er et behov for å utvikle gode retningslinjer for design av menneske-KI-interaksjoner. Jeg har allerede skrevet om Amershi et al., som presenterte sine 18 retningslinjer i 2019. Disse kan i stor grad sies å være en utvidelse av Jakob Nielsen sine 10 "Usability Heuristics for User Interface Design" fra 1994. Nielsens 10 prinsipper er mye brukt til å vurdere hva som er brukervennlig design innenfor HCI-feltet. Amershi et al. (2019) legger opp til den samme metoden med å vurdere brukervennlighet utfra en rekke standardiserte prinsipper. I likhet med Nielsen sine prinsipper fra 1994, er Amershi et al. opptatt av å holde brukeren informert underveis. En interessant distinksjon finnes likevel allerede i Amershi et al. sin første retningslinje, nemlig at det skal gjøres tydelig fra start hva systemet kan gjøre. Den andre er å gjøre det tydelig hvor godt systemet kan gjøre dette, og hvor systemet kan gjøre feil. Dette illustrerer mye av det som skiller KI-design fra annen HCI-design, nemlig at KI-systemene kan være så komplekse at brukerne ikke forstår hva de skal brukes til eller hvordan det løser oppgavene sine – og langt mindre risikoen for at systemet gjør feil uten å informere om dette.

Et system som forsøker å følge den andre retningslinjen er ChatGPT, som konstant har denne setningen under chattene til brukerne: "ChatGPT can make mistakes. Check important info." (OpenAI, u.d.). Ettersom systemet er svært komplekst og har en hang til å hallusinere, må det diskuteres om denne retningslinjen egentlig oppfylles bare ved hjelp av denne korte teksten. Det er også utfordrende for systemer som ChatGPT å følge den første Amershi et al.-retningslinjen og informere fra start om hva systemet kan gjøre. Det gjøres et forslag ved at brukeren møtes av spørsmålet «What can I help with» og en rekke forslag under skrivefeltet til brukeren (fig. 3).



Figur 3: Skjermbilde av startskjermen til ChatGPT, tatt 23. oktober 2024.

Dette er eksempler på områder hvor ChatGPT og andre samtalegrensesnitt for store språkmodeller faller utenfor retningslinjene og kategoriseringene forskere og designere opererer med i dag. Dette underbygger behovet for mer forskning og løpende utvikling av disse retningslinjene for å sikre etisk, trygt og brukervennlig design av disse systemene.

Gitt at det ikke finnes garantier i dag for at slike systemer ikke vil gjøre feil, kan en løsning være å designe bedre for at brukeren skal forvente, akseptere og være klar over feilene som gjøres. Kocielnik et al. (2019) tester i sitt prosjekt metoder for å sørge for dette. Forfatterne ser konkret på hvordan planleggingssystemer integrert i mailprogrammer kan oppdage og lagre avtaler for brukeren. De foreslår tre konkrete teknikker for å øke brukernes aksept av feilbarlige AI-systemer. Disse innebærer å:

1) informere om feilene gjennom en nøyaktighetsindikator
2) begrunne og forklare feil gjennom eksempler
3) gi brukeren førstehåndserfaringer med feil og mulighet til å kontrollere risikoen for feil selv

Det kan diskuteres i hvor stor grad disse løsningsforslagene kan adapteres til design av samtalegrensesnitt for store språkmodeller. Noen som helt spesifikt har diskutert utfordringene med disse samtalegrensesnittene, er Bender et al. (2021). Det må legges merke til at artikkelen er publisert før lanseringen av ChatGPT. Bender et al. (2021, s. 618) er spesielt bekymret for at språkmodeller lærer seg og sprer fordommer og stereotyper og at det er stor usikkerhet knyttet til hva systemene husker fra dataen de er trent på. De viser til det beskrivende sitatet "Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy." De diskuterer også hvem som har ansvaret for hva disse språkmodellene lirer av seg.

Bender at al. (2021) argumenterer for at løsningen er å vise hensyn og prioritere mangfold og aksept når fremtidige modeller trenes. De mener at nøye utvelgelse av hvilke data systemene trenes på er avgjørende å si for hvilke stereotyper som videreføres, og at også mengden data og størrelsen på modellene må vurderes opp mot klimahensyn. Generelt oppfordrer til at utviklere av slike systemer må ta mer ansvar, også etter lansering, og mener at slik kan man sørge for at det gjøres bedre risikoanalyser underveis i utviklingsprosessene.

Bekymringene til Bender at al. (2021) illustrerer noen av farene disse nye, generative KI-systemene kan utgjøre når mennesker skal interagere med dem. I tillegg til at det trengs gode retningslinjer i fremtiden, kan det også argumenteres for at teknologien må begrenses og kontrolleres med sterkere midler for å sørge for etisk bruk. Det kommende KI-regelverket til EU er et eksempel på at dette er noe vi kan vente oss i fremtiden.

Til tross for farene, er det liten tvil om at KI-systemer som ChatGPT har kommet for å bli. Skjuve, Brandtzaeg og Følstad undersøkte i 2024 hvorfor folk valgte å bruke ChatGPT, og kom frem til disse svarene: økt effektivitet, forbedring av læring og kreativitet, støtte i kommunikasjon med venner og i beslutningstakning og hjelp med rutineoppgaver. De konkluderer til og med at det å ikke bruke ChatGPT kan bli ansett som en kostnad.

7 Konklusjon

KI er på "alles" lepper om dagen, men det er langt fra noe nytt. Feltet ble etablert for over et halvt århundre siden, men den siste tidens utvikling har gjort det brennaktuelt. I dag lever vi i en verden hvor mange mennesker bruker KI hver eneste dag, om det er i støvsugeren, Netflix' anbefalingssystem eller GPS-en i bilen. Derfor er det også viktig å forsøke å skape en felles forståelse for hva KI er, og å etablere gode retningslinjer som sørger for etisk og inkluderende design av systemene.

For å bidra til en felles forståelse, har jeg satt sammen tre etablerte definisjoner av KI i et forsøk på å formulere en enhetlig definisjon. Den lyder slik: *"Kunstig intelligens (KI) er teknologi som kan påvirke verden, både digitalt og fysisk, gjennom å lære, tilpasse seg og løse problemer mer eller mindre autonomt med varierende tilgang på ressurser og informasjon."*

Så, har denne teksten brakt oss nærmere et svar på hva KI egentlig er? Definisjonen indikerer ja, men teksten viser også kompleksiteten i feltet. Teknologiens kompleksitet har

også store implikasjoner for design av KI-integrerte systemer. Teknologien kan ha uheldige konsekvenser, og listen over mulige fallgruver er mange. I forskningen er det spesielt knyttet bekymringer til design av store språkmodeller som trenes opp på stereotypiske fordommer og systemer som kan hallusinere og gjøre alvorlige feil uten at brukeren merker det. Det kreves derfor kontinuerlig innsats for å holde teknologien innenfor etiske rammer – ikke minst for å unngå dystopiske science fiction-scenarioer.

Referanser

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bratteteig, T., & Verne, G. (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) (Article 8, pp. 1–5). Association for Computing Machinery. https://doi.org/10.1145/3210604.3210646

Bufdir. (u.d.). Universal design - terms and terminology. Hentet 15.09.24 fra https://www.bufdir.no/en/likestilling/terms-and-terminology

ChatPDF. (u.d.). Hentet 15.09.24 fra https://www.chatpdf.com/

Dautenhahn, K. (2018). Some brief thoughts on the past and future of human-robot interaction. *ACM Transactions on Human-Robot Interaction,* 7(1), 1-3. https://doi.org/10.1145/3209769

Digdir. (u.d.). Ny forordning for kunstig intelligens. Hentet 17.09.24 fra https://www.digdir.no/kunstig-intelligens/ny-forordning-kunstig-intelligens/4271

Elkjøp. (u.d.). En robot for enhver familie - over hele verden. Hentet 18.09.24 fra https://www.elkjop.no/brand/ecovacs

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

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

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, 61(4), 5-14. https://doi.org/10.1177/0008125619864925

Holmes-Iverson, K. (2023). Your computer might be racist. *Howard Magazine*, Spring/Summer 2023. Hentet 18.09.24 fra https://magazine.howard.edu/stories/your-computer-might-be-racist

Hyatt, S. E., & Owenz, M. B. (2024). Using universal design for learning and artificial intelligence to support students with disabilities. *College Teaching*, 1–8. https://doi.org/10.1080/87567555.2024.2313468

Khogali, H., & Mekid, S. (2023). The blended future of automation and AI: Examining some long-term societal and ethical impact features. *Technology in Society*, 73, 102232. https://doi.org/10.1016/j.techsoc.2023.102232

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). Association for Computing Machinery. https://www.microsoft.com/en-us/research/uploads/prod/2019/01/chi19_kocielnik_et_al.pdf

Korslund, K. (2024). Frankenstein. *Store norske leksikon*. Hentet 15.09.24 fra https://snl.no/Frankenstein

Manning, C. (2020). Artificial intelligence definitions. Hentet 17.09.24 fra https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf

McIntosh, T. R., Liu, T., Susnjak, T., Watters, P., Ng, A., & Halgamuge, M. N. (2023). A culturally sensitive test to evaluate nuanced GPT hallucination. *IEEE Transactions on Artificial Intelligence*. https://doi.org/10.1109/TAI.2023.3332837

Nielsen, J. (1994). 10 usability heuristics for user interface design. Hentet 18.09.24 fra https://www.nngroup.com/articles/ten-usability-heuristics/

OpenAI. (u.d.). ChatGPT. Hentet 18.09.24 fra https://www.openai.com/chatgpt

Pisoni, G., Díaz-Rodríguez, N., Gijlers, H., & Tonolli, L. (2021). Human-centered artificial

intelligence for designing accessible cultural heritage. *Applied Sciences*, 11(2), 870. https://doi.org/10.3390/app11020870

Rawte, V., Sheth, A., & Das, A. (2023). A Survey of Hallucination in "Large" Foundation Models. AI Institute, University of South Carolina, USA. https://doi.org/10.48550/arXiv.2309.05922

Sagdahl, M. S. (2019). Autonomi - filosofi. *Store norske leksikon*. Hentet 17. september 2024 fra https://snl.no/autonomi_-_filosofi

Skjuve, M., Brandtzaeg, P. B., & Følstad, A. (2024). Why do people use ChatGPT? Exploring user motivations for generative conversational AI. *First Monday*, 29(1). https://doi.org/10.5210/fm.v29i1.13541

Søraa, R. A. (2022). Robot. *Store norske leksikon*. Hentet 18.09.24 fra https://snl.no/robot

Thrun, S. (2004). Toward a framework for human-robot interaction. *Human-Computer Interaction*, 19(1-2), 9-24. https://doi.org/10.1207/s15327051hci1901&2_2

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. Oxford University Press on behalf of the Mind Association. http://www.jstor.org/stable/2251299

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2), 1–37. https://doi.org/10.2478/jagi-2019-0002

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* (pp. 1-13). https://doi.org/10.1145/3313831.3376301

Zewe, A. (2023). Explained: Generative AI. *MIT News*. Hentet 10.09.24 fra https://news.mit.edu/2023/explained-generative-ai-1109

Vedlegg

Vedlegg 1: Endringer etter tilbakemelding

Etter første iterasjon av denne oppgaven, mottok jeg denne tilbakemeldingen fra en medstudent:

Hei Mari!
Kjempebra og spennende refleksjon om KI og roboter.

- ⭐Utrolig godt skrevet, både i språk og struktur gjennom hele oppgaven!
- ⭐Kjempegod bruk av pensum, spesielt godt bruk av Yang(2020)
- 💡Eneste jeg savner er oppgave 1.5 Journal - everyday engagement, interaction and use.

Dette var en svært nyttig tilbakemelding, ettersom jeg hadde misforstått oppgave 1.5 til å være en metode vi skulle bruke for å reflektere over hele oppgaven, ikke en egen del som skulle skrives. Jeg har lagt til «4 KI i studenthverdagen: En øvelse» i denne iterasjonen, i tillegg til å svare på de nye oppgavene vi mottok i modul 2.
Ettersom de nye oppgavene i modul 2 overlappet flere steder med tekst jeg allerede hadde skrevet, har jeg bygget ut med ny tekst og refleksjoner flere steder i oppgaven, fremfor å legge dette som kronologiske punkter nederst. Jeg håper dette ikke er forvirrende for neste medstudent som skal gi tilbakemelding.