Obligatorisk oppgave modul 1

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

Kunstig intelligens (AI) er ifølge Bratteteig og Verne (2018) et underfelt av informatikk som har som mål å utvikle datasystemer som etterligner menneskelig intelligens. I starten var AI regelbasert, med algoritmer som etterlignet menneskelig problemløsning. Rundt 1950-tallet bidro Alan Turing med betydelige ideer om maskiners evne til å imitere menneskelig tenkning, og under 2. verdenskrig ble regelbaserte systemer brukt til å knekke koder (Grudin, 2009). Etter krigen ble forskning på datamaskiner finansiert, og begrepet «kunstig intelligens» ble introdusert i 1956 av John McCarthy. Siden da har AI utviklet seg raskt, med maskinlæring som et viktig underfelt. Maskinlæring lar datamaskiner lære og forbedre seg ved å gjenkjenne mønstre i store datasett, noe som har revolusjonert AI fra tidligere faste, regelbaserte systemer til mer fleksible og adaptive modeller (Bratteteig & Verne, 2018).

Jeg vil kort redegjøre for tre definisjoner av AI. Den første er fra Bratteteig og Verne (2018), som beskriver AI som «a subfield of computer science aimed at specifying and making computer systems that mimic human intelligence or express rational behaviour, in the sense that the task would require intelligence if executed by a human» (s. 2). Denne definisjonen kommer fra Participatory Design-feltet og sier at AI er datasystemer som kan etterligne menneskelig intelligens. Den andre er fra Jonathan Grudin (2009), som i artikkelen *«AI and HCI: Two Fields Divided by a Common Focus»* skriver at «AI focused on devising better algorithms, HCI on how to improve the use of existing algorithms. AI originated in mathematics and engineering, HCI in psychology» (s. 48). Grudin fremhever hvordan AI og HCI har utviklet seg i forskjellige retninger, til tross for deres felles mål. Til slutt har vi definisjonen fra Jo Herstad (2024), som sier: «The words Interaction, Artificial Intelligence and Autonomous are key concepts that are used both to describe «humans, like you and me» and «things, like a computer or a service» (s. 4). Herstad understreker hvordan begreper som tidligere beskrev menneskelige egenskaper, nå brukes for å karakterisere teknologi. Jeg definerer AI som datamaskiners evne til å simulere menneskelig logikk og beslutningstaking. AI kan overgå menneskelige evner ved å utføre oppgaver raskere, med færre feil, og gjøre ting mennesker ikke kan. Likevel har AI begrensninger, som mangel på menneskelig kreativitet, følelsesmessig innsikt, og nyansert kontekstforståelse.

Forskjellen mellom menneske autonomi og maskinell autonomi

Ifølge Thrun (2004) handler maskinell autonomi om roboter som kan ta egne beslutninger i forskjellige situasjoner, inkludert sosiale interaksjoner. Mennesker tar beslutninger basert på følelser, etikk, og personlig erfaring, mens maskiner følger algoritmer og dataanalyse. Dette reiser etiske spørsmål om ansvar når AI-systemer, som selvkjørende biler, forårsaker skade eller tap av liv (Holmquist, 2017). Maskiner kan ha varierende grader av autonomi, fra fullstendig AI-kontroll til delt kontroll med brukeren, i kontrast til menneskelig autonomi, som er knyttet til personlige verdier og kontekst.

Likheter mellom begrepet automatisk og begrepet autonom

I artikkelen *«Ironies of Automation»* beskriver Bainbridge (1982) automatisering som en erstatning for menneskelig kontroll, planlegging og problemløsning gjennom automatiske enheter og datamaskiner. Begrepene «automatisk» og «autonom» overlapper ved at begge handler om systemer som opererer uten direkte menneskelig innblanding. Automatiske systemer følger forhåndsprogrammerte regler, mens autonome systemer har en høyere grad av selvstyring og kan tilpasse seg endringer og ta egne beslutninger basert på sanntidsinformasjon.

Kort gjennomgang av artikkelen av «Classifying Human and Robot Movement at Home and Implementing Robot Movement Using the Slow In, Slow Out Animation Principle» av Schulz, Herstad & Torresen (2018).

Artikkelen tar for seg Human-Robot Interaction HRI, og hvordan man kan benytte animasjonsteknikker for å tilpasse robotbevegelser mer i forhold til mennesker. Forfatterne fokuserer spesielt på prinsippet om «Slow In, Slow Out», som er en teknikk innen animasjon der bevegelsen starter og slutter sakte, noe som gir en mer naturlig og menneskelignende bevegelse. Studien analyserer menneskelig og robotbevegelse i hjemmemiljøer og diskuterer hvordan roboter kan tilpasse sine bevegelser for å bedre samhandle med mennesker.

En bedrift som jobber med AI - Telenor

Telenor presenterer kunstig intelligens (AI) som både en plattform og en tjeneste som driver innovasjon internt og eksternt, samtidig som de framhever AI som en transformerende idé med betydelig potensial for samfunnsutvikling. De beskriver AI som en kjernekomponent i deres overgang til å bli en datadrevet virksomhet, med fokus på optimalisering av operasjoner, automatisering av prosesser og forbedring av

kundeinteraksjoner. På nettsiden finnes det en spesifikk side om AI, samtidig som det finnes flere nyhetsartikler relatert til AI.

Kunstig intelligens i Black Mirror

I episoden «Joan is Awful» fra Black Mirror utforskes et dystopisk scenario hvor strømmetjenester bruker en kraftig AI til å lage hyperrealistiske TV-serier basert på vanlige menneskers liv, uten deres samtykke. Hovedpersonen Joan oppdager at hennes daglige liv blir omgjort til en TV-serie, der en populær skuespillerinne spiller henne. Episoden tar for seg hvordan AI-teknologi kan bli brukt til å overvåke, analysere, og deretter produsere innhold uten noen form for etisk vurdering eller hensyn til individets rettigheter.

Robots and AI systems

I 1979 definerte Robot Institute of America en robot som en «omprogrammerbar, multifunksjonell manipulator» som flytter materialer, deler, verktøy eller spesialiserte enheter gjennom forskjellige programmerte bevegelser for å utføre varierte oppgaver (Thrun, 2004). Merriam Webster's collegiate dictionary definerte robot som «An automatic device that performs functions normally ascribed to humans or a machine in the form of a human.» (Thrun, 2004, s.11). Denne definisjonen beskriver robot som en enhet som kan utføre oppgaver som vanligvis krever menneskelig innsats, samt en maskin som etterligner menneskelig form eller handlinger.

Jeg vil selv definere robot som en fysisk maskin eller plattform som kan programmeres for å gjøre spesifikke oppgaver som simulerer eller utkonkurrerer mennesker i forhold til presisjon, effektivitet og kapasitet.

FN klassifiserer roboter i tre hovedkategorier: industrielle roboter, profesjonelle serviceroboter, og personlige serviceroboter, basert på deres bruksområder. Industrielle roboter, som er den eldste og mest utbredte typen, brukes til oppgaver som montering, sveising, pakking og transport, og har hatt en betydelig vekst siden 1960-tallet.

Relasjonen mellom AI og roboter

Basert på min definisjon av robot, kan ikke en chatbot sees som en form for robot ettersom den ikke har en fysisk kropp eller plattform. For å bli betraktet som en robot, måtte den ha en fysisk komponent, for eksempel en fysisk enhet som kan interagere med den virkelige verden. AI gir roboter intelligens og beslutningstakingsevne, mens roboter gir en fysisk plattform for å manifestere disse funksjonene. AI kan

eksistere uavhengig av fysisk maskinvare, som i tilfelle av chatbots og andre programvarebaserte systemer, som beskrevet av Bratteteig & Verne (2018) som «a subfield of computer science aimed at specifying and making computer systems that mimic human intelligence.». Dette betyr at AI kan utføre oppgaver som krever menneskelig intelligens, men uten behov for fysisk kropp.

Eksempel på en moderne fysisk robot

I artikkelen *«Classifying Human and Robot Movement at Home and Implementing Robot Movement Using the Slow In, Slow Out Animation Principle»* av Schulz, Herstad & Torresen (2018) blir TurtleBot3 brukt som eksempel på en moderne robot som anvender animasjonsteknikken «slow in, slow out» for å forbedre bevegelsene sine. Denne teknikken gjør at roboten akselererer og bremser gradvis, noe som resulterer i mer naturlige og forutsigbare bevegelser. Når TurtleBot3 navigerer i et hjemmemiljø, gir denne myke bevegelsen brukeren en bedre opplevelse ved å forhindre plutselige bevegelser og kollisjoner, og gjør interaksjonen med roboten både mer intuitiv og komfortabel.

Universal Design and AI systems

Ifølge Herstad (2024) handler universell utforming om «questions concerning inclusion». Dette innebærer å skape løsninger som sikrer at alle mennesker, uavhengig av deres individuelle behov og evner, kan delta på like vilkår. Målet er å utvikle systemer og produkter som er tilgjengelige og brukbare for et mangfold av brukere, og dermed unngå å ekskludere noen. Det dreier seg om vilkår og muligheter for å inkludere alle. Målet er at nye systemer ikke skal skape barrierer for ulike mennesker. I en video fra The Walrus Talks Inclusion understreker Jutta Treviranus (2019) viktigheten av å anerkjenne menneskelig unikhet. Dette prinsippet står i kontrast til masseproduksjon, massemarkedsføring, massekommunikasjon og enkle offentlige retningslinjer. Ofte designes systemer for «flertallet», der begrensede ressurser brukes for å nå mange. 80/20-prinsippet, hvor 20 % av ressursene gir 80 % effekt, er ikke inkluderende, da opptil 20 % av befolkningen kan bli utelatt. Derfor er det viktig i universell utforming å kontinuerlig spørre: «Blir noen utelatt?».

Hvordan AI kan forbedre persepsjon, bevegelser og kognisjon/følelser

For eksempel kan AI forbedre sanseinntrykk ved å hjelpe synshemmede med å tolke visuelle data, noe som kan gi dem bedre navigasjonsmuligheter. Innen bevegelse kan kunstig intelligens og maskinlæring tilpasse proteser etter individuelle bevegelsesmønstre, og dermed skape mer naturlige og flytende bevegelser. Når det gjelder kognisjon, kan AI støtte personifisert læring ved å tilpasse

undervisningsopplegg til den enkeltes læringsstil og tempo, noe som kan forbedre læringsopplevelsen og effektiviteten.

Potensialet til AI for både å inkludere og ekskludere mennesker

Selv om AI har stort potensial for å inkludere nye brukergrupper, er det også risikoer for ekskludering. For eksempel kan AI-teknologier oppleves som en trussel mot arbeidsplasser, da automasjon og maskinlæring kan overta oppgaver som tidligere ble utført av mennesker. Dette kan spesielt ramme grupper som er avhengige av enklere arbeidsplasser, som personer med funksjonsnedsettelser som har tilpasset seg bestemte jobber.

WCAG 2.1 «understand» og «understanding»

I WCAG 2.1 (2023) står det: «Understandable - Information and the operation of user interface must be understandable». Dette betyr at brukere skal være i stand til å forstå både informasjonen som presenteres og hvordan brukergrensesnittet fungerer. Utfordringer knyttet til kunstig intelligens i denne sammenhengen kan oppstå fordi menneskelig kognisjon og handlinger skjer i en kompleks kontekst som ikke alltid kan fullt ut representeres eller forstås av algoritmer og maskiner (Bratteteig & Verne, 2018, s. 2). AI-systemer baserer seg på data og mønstre, men mangler den dype kontekstuelle forståelsen som mennesker har. Dette kan føre til situasjoner der AI gir svar eller utfører handlinger som ikke helt matcher brukerens faktiske behov eller situasjon, da maskiner ikke kan forstå den subtile konteksten som er viktig for menneskelig forståelse.

Retningslinjer for Human-AI Interaction av Microsoft

«02 Make clear how well the system can do what it can do» handler om å sikre at brukeren tydelig forstår hvilke oppgaver AI-systemet kan utføre (Microsoft, 2019). Eksempler på dette kan være å vise hva en navigasjonsapp kan søke etter, forklare hvilke data en aktivitetsmonitor registrerer, eller informere om hvordan et sosialt nettverk anbefaler personer å følge. Brudd på retningslinjen skjer når systemets funksjoner ikke er tydelig kommunisert, som når en nyhetsfeed eller bildeorganisator ikke opplyser om hva som vises eller hvilke kategorier som er tilgjengelige (Amershi et al., 2019).

Forskjeller mellom HCI design retningslinjer og Human-AI Interaction retningslinjer

Jeg ønsker å sammenligne retningslinje 02 fra Microsofts «18 Guidelines for Human-AI Interaction» (2019) med designprinsippet om «feedback» som beskrevet av Don Norman i *The Design of Everyday Things* (2013, s. 23). Microsofts retningslinje «02 Make clear how well the system can do what it can do» fokuserer på å gi brukerne en klar forståelse av systemets kapabiliteter og begrensninger på forhånd. Dette innebærer å sette klare forventninger til hva systemet kan oppnå før eller under bruken. På den annen side dreier Don Normans konsept om tilbakemelding seg om å gi brukeren umiddelbar informasjon om resultatet av en spesifikk handling, som bekrefter om handlingen var vellykket eller mislykket. Mens «feedback» gir informasjon etter at en handling er utført, søker Microsofts retningslinje å informere brukeren om systemets evner og begrensninger på forhånd. Generelt fokuserer HCI-designretningslinjer på universelle prinsipper som tilbakemelding og brukervennlighet for effektiv interaksjon mellom mennesker og datamaskiner. I kontrast er retningslinjene for Human-AI Interaction spesifikke for hvordan brukere interagerer med AI-systemer, og tar for seg utfordringer som kommunikasjon av AI-kapabiliteter og håndtering av systemets begrensninger.

Dagbok – Refleksjon over daglig bruk av ulike skriveverktøy

Ved å bruke penn og papir la jeg stor vekt på å få håndskriften til å se pen ut. Den taktile følelsen av papir og blekk ga en tilfredsstillende opplevelse, men effektiviteten var begrenset, særlig når det gjaldt redigering og omskrivning. Med QWERTY-tastatur og skjerm økte hastigheten og effektiviteten. Jeg slapp å bekymre meg for håndskriftens utseende, og redigering og navigering ble mye lettere. Men skjermen kunne av og til virke distraherende. Stemmeskrivning introduserte en fin måte å raskt å fange ideer. Utfordringer med nøyaktighet og feilkorrigering ble tydelige, og selv om denne metoden frigjorde meg fra fysisk skriving, krevde den kontinuerlig overvåking for å rette feil og effektiv organisering og formatering av teksten. Alt i alt foretrakk jeg å skrive med penn og papir når det kommer til å skrive dagbok.

Vanligvis skriver jeg med tastatur, ettersom jeg setter pris på skriveapplikasjoner som Notion. Disse gir meg en enkel oversikt over tidligere notater innen et emne og lar meg søke etter spesifikke ord. Selv om jeg foretrekker å skrive for hånd, gjør digitaliseringen det verdt å bruke PC.

Modul 2

Kjennetegn på AI-baserte systemer

I artikkelen *«Guidelines for Human-AI Interaction»* av Amershi et al. (2019, s. 1) defineres AI-baserte systemer som «Systems that have features harnessing AI capabilities that are directly exposed to the end user». Dette innebærer at brukerne interagerer direkte med funksjoner drevet av kunstig intelligens. Ifølge Amershi mfl. (2019, s. 1), er AI-baserte systemer preget av *usikkerhet*, *kontekstfølsomhet* og *personalisering bak kulissene*. Usikkerhet oppstår ved at AI-drevne systemer ofte kan ha uforutsigbar atferd, noe som kan resultere i forvirring, forstyrrelser eller til og med mulig støtende eller farlige situasjoner. Kontekstfølsomhet skyldes at AI-systemer er iboende inkonsekvente, da de baserer seg på sannsynlighetsberegninger og læring over tid. Videre, kjennetegnes AI-baserte systemer av personalisering bak kulissene; AI-funksjoner tilpasser ofte brukeropplevelsen automatisk uten direkte interaksjon. Dette skaper en mangel på kontroll og transparens, hvor brukerne kan føle usikkerhet rundt grunnlaget for personaliseringen, og de har begrensede muligheter til å påvirke eller korrigere systemets valg (Amershi et al., 2019, s. 1).

Kocielnik et al. (2019) skriver i artikkelen *«Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems»,* hvordan AI-baserte systemer ofte har preg av usikkerhet tilsvarende som Amershi et al., med ved at de er *probabilistiske* noe som betyr at AI-baserte systemer er basert på sannsynlighetsregninger og ikke alltid gir perfekte resultater (Kocielnik et al., 2019, s. 2). Som en konsekvens, operer AI-systemer nesten alltid med en viss grad av unøyaktighet. Kocielnik et al. (2019, s. 3) fremhever også, i likhet med Amershi et al., hvordan AI-baserte systemer kontinuerlig tilpasser og endrer seg basert på brukerinteraksjon. De legger imidlertid til at slike systemer ofte sliter med *transparens*, noe som gjør det essensielt å kommunisere systemets virkemåte tydelig til brukeren. Forfatterne argumenterer for at brukerne er mer tilbøyelige til å akseptere feil i AI-systemer dersom de forstår hvordan systemene fungerer og hvorfor de tar de beslutningene de gjør (Kocielnik et al., 2019, s. 9).

Yang et al. (2020, s. 8) utdyper, på lignende måte som Amershi et al. og Kocielnik et al., hvordan usikkerheten knyttet til AI-baserte systemer gjør det utfordrende å forutsi nøyaktig hvordan disse systemene vil prestere i praksis. Ytelsen kan variere basert på data, treningsprosesser og kontekst, noe som gjør det vanskelig å designe pålitelige og effektive interaksjoner. Artikkelen argumenterer for

viktigheten av videre forskning innen HCI for å utvikle nye metoder som kan møte utfordringene knyttet til design for AI-systemer.

Et eksempel på en AI-basert chatbot som passer til beskrivelsene av Amershi et al. (2019), Kocielnik et al. (2019), og Yang et al. (2020) er NAVs chatbot «NAV Hjelp», beskrevet av Verne et al. (2022). NAV Hjelp bistår brukere med spørsmål om velferdsytelser ved å gi svar basert på naturlig språkbehandling. Likevel kan chatbotten, som mange AI-systemer, skape utfordringer når svarene er irrelevante eller ufullstendige. Dette samsvarer med Amershi et al. (2019) påpekning av at AI-systemer ofte har uforutsigbar oppførsel, noe som kan føre til usikkerhet for brukerne. Når NAV Hjelp gir feilaktige eller mangelfulle svar, kan det få alvorlige konsekvenser for brukere som er avhengige av korrekt informasjon for å ta avgjørende beslutninger.

Videre er NAV Hjelp preget av både kontekstfølsomhet og personalisering bak kulissene, noe Amershi et al. (2019) og Kocielnik et al. (2019) beskriver som sentrale aspekter ved AI-systemer. Chatbotten tilpasser seg brukernes input, men denne kontekstfølsomheten kan føre til feil hvis systemet ikke klarer å fange opp alle relevante nyanser. I tillegg kan personalisering uten tilstrekkelig transparens skape en følelse av mangel på kontroll for brukerne, spesielt når de ikke forstår hvorfor bestemte svar blir gitt. For å opprettholde tillit og brukertilfredshet bør NAV Hjelp derfor fokusere på større åpenhet og forklarbarhet rundt hvordan systemet fungerer og tilpasser seg brukernes behov.

Human-AI interaction design

Amershi et al. (2019) og Kocielnik et al. (2019) adresserer begge utfordringer ved design av AI-drevne systemer, men de nærmer seg temaet fra ulike vinkler.

Amershi et al. (2019) fremhever behovet for videre forskning og innovasjon innen AI-teknologier, og påpeker at AI-systemer ofte bryter med tradisjonelle brukervennlighetsretningslinjer og kan vise uforutsigbar oppførsel som kan være forvirrende, forstyrrende eller til og med farlig. For å adressere disse problemene foreslår Amershi et al. 18 designretningslinjer for menneske-AI-interaksjon, utviklet gjennom flere evalueringsrunder, inkludert en brukerstudie med 49 designpraktikere som testet retningslinjene mot 20 AI-drevne systemer. Deres funn antyder at det å følge disse retningslinjene kan føre til utvikling av mer menneskesentrerte AI-teknologier.

I motsetning til dette fokuserer Kocielnik et al. (2019) på å håndtere sluttbrukernes forventninger til AI-drevne systemer for å forbedre brukeropplevelsen. De påpeker et gap i forskningen når det gjelder metoder for å sette forventninger før bruk av AI-systemer. Deres studie av en planlegger-assistent førte til utviklingen av tre teknikker for å forberede brukerne på AI-ufullkommenheter: en *nøyaktighetsindikator* for å vise systemets nøyaktighet, *eksempelbaserte forklaringer* for å klargjøre hvordan systemet fungerer, og en *kontrollskyveknapp* som lar brukerne justere systemets ytelse. Målet deres er å øke brukernes aksept for AI-drevne systemer, noe de viser noe effektivitet for i sine funn.

Amershi et al. sine 18 designretningslinjer for AI-baserte systemer

Amershi et al. (2019, s. 2) sine 18 designretningslinjer bygger på prinsipper som åpenhet, forklarbarhet og forutsigbarhet i AI-drevne systemer. De har som mål å forbedre brukeropplevelsen ved å tydeliggjøre systemets funksjoner og begrensninger, samtidig som interaksjonene tilpasses den spesifikke konteksten. Videre legger retningslinjene vekt på at systemet skal oppføre seg i tråd med brukernes forventninger og sosiale normer. I tillegg understrekes behovet for å gi brukerne mulighet til effektiv tilbakemelding og korreksjon, samt å sikre at viktig informasjon bevares over tid for å opprettholde systemets pålitelighet og brukernes tillit.

Den første retningslinjen, *G1: Gjør det klart hva systemet kan gjøre*, setter søkelys på å hjelpe brukeren med å forstå hva AI-systemet er i stand til. For eksempel kan en aktivitetsmonitor vise brukeren alle de måledataene den sporer, samt hvordan dette gjøres (Amershi et al., 2019, s.3). NAVs chatbot gjør en god jobb med å informere brukerne om hva den kan gjøre ved å tilby svar på spørsmål knyttet til velferdsytelser og tjenester. Chatbotten lar brukerne stille spørsmål og gir dem umiddelbar tilgang til informasjon om velferdsregler. Dette gjør det enkelt for brukerne å forstå hvilke tjenester de kan få hjelp med. Selv om chatbotten gir god informasjon, kan det være nyttig å inkludere en oversikt over spesifikke temaer og spørsmål som chatbotten kan håndtere. En mer visuell representasjon av chatbotens funksjoner, som en FAQ-seksjon eller en liste over emner, kan forbedre brukeropplevelsen ved å gjøre det enklere for brukerne å navigere i systemet.

Den andre retningslinjen, *G2: Gjør det klart hvor godt systemet kan utføre det det kan gjøre*, fokuserer på å informere brukeren om systemets nøyaktighet og mulige feil. For eksempel kan et musikkanbefalingssystem bruke formuleringer som «vi tror du vil like» for å indikere at anbefalingene ikke alltid er feilfrie. Bruken av *hedging-språk* bidrar til at brukerne forstår at selv om systemet har evnen

til å gi gode anbefalinger, er det fortsatt en risiko for feil. Ved å kommunisere denne usikkerheten kan utviklere redusere mulige skuffelser og misforståelser, noe som kan øke brukerens tillit til systemet (Amershi et al., 2019, s. 3). NAVs chatbot demonstrerer en god evne til å forklare hva den kan utføre. Da jeg spurte chatboten om vilkårene for dagpenger, opplyste den om at det finnes mange betingelser som må være oppfylt, i stedet for å ramse opp alle vilkårene, noe som kunne ha ført til feil informasjon. I stedet henviste den meg til nettsiden for mer informasjon om dagpenger. På denne måten unngikk chatboten å spre unøyaktig informasjon. Den benyttet også hedging-språk som: «Jeg kan svare deg på spørsmål om NAVs regler og hjelpe deg med å navigere på nav.no, men jeg kan ikke sjekke saken din eller ta imot beskjeder,» samt, «Jeg lærer stadig nye ting, og blir bedre til å forstå hva du spør om».

Kritiske perspektiver på AI-infiserte systemer: Innsikter fra Bender et al. (2021)

Bender et al. (2021) peker på flere problematiske aspekter ved tekstinnholdet fra store språkmodeller. *Bias og stereotyper* utgjør en betydelig bekymring, ettersom disse modellene ofte reflekterer og forsterker skjevheter fra treningsdataene, noe som kan føre til tekst som opprettholder fordommer eller diskriminerende språk, med alvorlige samfunnsmessige konsekvenser for marginaliserte grupper (Bender et al., 2021, s. 617). *Desinformasjon* er en annen utfordring; store språkmodeller kan generere overbevisende, men faktisk uriktig informasjon, noe som bidrar til spredning av feilaktige opplysninger, der brukere feilaktig kan anta at teksten er pålitelig. I tillegg skaper *ansvarlighet* etiske og juridiske spørsmål, ettersom det blir vanskelig å holde enkeltpersoner eller organisasjoner ansvarlige for innholdet som skapes uten forståelse eller intensjon (Bender et al., 2021, s. 610). Det er også *miljømessige og økonomiske kostnader* knyttet til trening og drift av store språkmodeller, noe som reiser spørsmål om bærekraft og rettferdig fordeling av teknologiske fordeler (Bender et al., 2021, s. 612). For å håndtere de etiske utfordringene knyttet til store språkmodeller (LLMs) foreslår Bender et al. (2021, s. 619) *value sensetive design* som går ut på å identifisere og integrere interessentenes verdier gjennom hele utviklingsprosessen for teknologi. Videre anbefaler de regulerings- og policytiltak for å etablere klare retningslinjer for bruken av disse modellene, noe som kan bidra til å dempe mulige skader og fremme ansvarlig utvikling som prioriterer velvære for enkeltpersoner og samfunn som påvirkes av teknologiene (2021, s. 619).

Samtalebrukergrensesnitt til store språkmodeller

Samtalegrensesnitt for store språkmodeller som ChatGPT introduserer flere utfordringer i designprosessen. Som tidligere nevnt, er usikkerhet et sentralt aspekt ved AI-baserte systemer (Amershi et

al., 2019, s. 1; Kocielnik et al., 2019, s. 2). Siden disse systemene er basert på sannsynlighetsberegninger, kan de ikke alltid levere perfekte resultater. Dermed opererer AI-systemer ofte med en viss grad av unøyaktighet. Yang et al. (2020, s. 1) understreker dette ved: «... current HCI design methods meant to mitigate unintended consequences (i.e. sketching and prototyping) can seem ill-fitted for AI. HCI professionals cannot easily sketch the numerous ways an AI system might adapt to different users in different contexts [13, 58]. Nor can they easily prototype the types of inference errors a not-yet developed AI system might make». Denne sitatet illustrerer hvordan tradisjonelle designmetoder, som skissering og prototyping, ikke alltid er tilstrekkelige for å håndtere kompleksiteten som følger med AI-systemer. Videre nevner Yang et al. (2020, s. 2-3) at designere ofte står overfor betydelige utfordringer når det gjelder å forstå AI-systemers evner - hva de kan og ikke kan gjøre. Dette inkluderer vanskeligheter med å forestille seg og utvikle innovative, implementerbare AI-løsninger for spesifikke brukeropplevelsesproblemer. I tillegg er det utfordrende å gjennomføre iterativ prototyping og testing av samhandling mellom mennesker og AI.

I forhold til retningslinjene G1 og G2 fra Amershi et al. (2019) kan iverksetting av disse bidra til å løse noen av utfordringene som eksisterer i dagens samtalegrensesnitt for store språkmodeller. G1, som oppfordrer designere til å klargjøre hva systemet kan gjøre, kan hjelpe brukerne med å forstå de faktiske evnene til AI-en, og dermed redusere usikkerhet. Dette kan oppnås ved å gi tydelige beskrivelser av systemets funksjoner og eksempler på hva brukerne kan forvente. G2, som fokuserer på å gjøre det klart hvor godt systemet kan utføre oppgavene sine, bidrar til å sette realistiske forventninger ved å gi informasjon om nøyaktighet og pålitelighet. Ved å anvende disse retningslinjene kan designere skape en mer transparent og brukervennlig opplevelse som forbedrer tilliten til AI-systemene og reduserer frustrasjonen knyttet til uventede resultater.

Litteratur

Amershi, S., Inkpen, K., Teevan, J., & Kikin-Gil, R. (2019). Guidelines for human-AI interaction. *In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). https://www.researchgate.net/publication/332742200_Guidelines_for_Human-AI_Interaction

Bainbridge, L. (1982). Ironies of automation. *IFAC Proceedings Volumes*, 15(6), 129-135. https://doi.org/10.1016/S1474-6670(17)62897-0

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). ACM. https://dl.acm.org/doi/pdf/10.1145/3442188.3445922

Bratteteig, T., & Verne, G. (2018). Does AI make PD obsolete?: Exploring challenges from artificial intelligence to participatory design. *Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2, 1–5*. https://doi.org/10.1145/3210604.3210646

Grudin, J. (2009). AI and HCI: Two fields divided by a common focus. *The AI Magazine, 30(4), 48-57*. https://go-gale-com.ezproxy.uio.no/ps/i.do?p=AONE&u=oslo&id=GALE|A216267417&v=2.1&it=r

Herstad, J. (2024, August 22). IN5620 Interaction with AI and autonomous systems [Lecture notes]. University of Oslo. https://www.uio.no/studier/emner/matnat/ifi/IN5620/h24/fra-forelesning/in5620-lecture-notes-22-august-2024.pdf

Holmquist, L. E. (2017). Intelligence on tap: Artificial intelligence as a new design material. *Interactions*, 24(4), 28-33. https://doi.org/10.1145/3085571

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

Microsoft. (2019, February 1). Guidelines for human-AI interaction design. https://www.microsoft.com/en-us/research/blog/guidelines-for-human-ai-interaction-design/

Norman, D. A. (2013). *The design of everyday things*. MIT Press. https://www.uio.no/studier/emner/matnat/ifi/IN1050/h23/timeplan/norman-principles.pdf

Sharp, H., Rogers, Y., & Preece, J. (2023*). Interaction design: Beyond human-computer interaction* (Sixth edition). Wiley.

Thrun, S. (2004). Toward a framework for human-robot interaction. *Human–Computer Interaction*, 19(1-2), 9-24. https://doi.org/10.1080/07370024.2004.9667338

The Walrus. (2019, October 25). *We count!* | *Jutta Treviranus* [Video]. YouTube. https://www.youtube.com/watch?v=OAXmCAqZqRk

Verne, G. B., Steinstø, T., Simonsen, L., & Bratteteig, T. (2022). How can I help you? A chatbot's answers to citizens' information needs. *Scandinavian Journal of Information Systems*, 34(2), 7. https://www.duo.uio.no/handle/10852/110430

World Wide Web Consortium. (2023). Web content accessibility guidelines (WCAG) 2.1. W3C Recommendation. https://www.w3.org/TR/WCAG21/

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

Appendiks

Feedback jeg fikk fra modul 1:

«Hei, Ingeborg!

Under kommer feedback til den individuelle oppgaven din, iterasjon 1:

Two stars:

- Du har besvart alle oppgaver og gjort dette på en lettfattelig måte, bra! Du har også brukt gode kilder fra pensum for å underbygge argumentene dine, som gjør konklusjonene dine veldig troverdige.
- Jeg liker definisjonene du kommer med. De dekker det viktigste, og måten du bruker dem på viser at du har forstått stoffet.

One wish:

- Skulle gjerne lest mer om din opplevelse med bruk av ulike skriveverktøy (selv om du nok hadde dårlig med plass pga. maks 6 sider). Det hadde vært spennende å høre mer om hvilket skriveverktøy du vanligvis bruker, hvorfor du ikke bruker de andre, hvilke andre konklusjoner du kom frem til under eksperimentet osv.

Det er bare å sende e-post dersom noe var uklart»

Basert på tilbakemeldingen skrev jeg mer om hvilket skriveverktøy jeg vanligvis bruker, og konklusjonene jeg kom frem til.