Individuell assignment

1.1 – Concepts, definition and history with interaction with AI an autonomous systems

Begrepet "kunstig intelligens" ble først introdusert i 1956 av den amerikanske professoren John McCarthy under en konferanse på Dartmouth College. Ideen om at maskiner kunne simulere menneskelig intelligens hadde imidlertid røtter langt tidligere, med forløpere som Alan Turing. Forskere begynte å eksperimentere med algoritmer og logiske systemer som kunne utføre oppgaver som normalt krevde menneskelig intelligens, som problemløsning, beslutningstaking, læring og språkforståelse.

Tre definisjoner av kunstig intelligens

McCarthy, som opprinnelig skapte begrepet, definerte AI som "vitenskapen og ingeniørkunsten ved å lage intelligente maskiner". Denne definisjonen er bred og inkluderer ethvert forsøk på å simulere menneskelig intelligens i maskiner. Dette perspektivet er vanlig i det tekniske miljøet og fokuserer på AI som et verktøy for problemløsning og automatisering.

Russell og Norvig, i sin bok Artificial Intelligence: A Modern Approach definerer AI som "studiet av agenter som mottar oppfatninger fra sitt miljø og utfører handlinger". Her blir AI sett som et system som oppfatter, resonerer og handler. Denne definisjonen er mye brukt innen akademia og forskningsmiljøer og legger vekt på intelligens som evnen til å ta handling basert på data.

IBM definerer AI som "et felt som kombinerer datavitenskap og robuste datasett for å muliggjøre problemløsning". Denne definisjonen understreker de teknologiske og datadrevne aspektene ved AI, og hvordan maskiner lærer og forbedrer seg over tid gjennom data. Dette perspektivet er utbredt i industrien, spesielt blant teknologiselskaper som fokuserer på maskinlæring og dataanalyse.

Basert på de tre definisjonene, vil jeg definere kunstig intelligens som "utviklingen av systemer som kan utføre oppgaver som krever intelligens, ved å bruke data for å forstå,

tilpasse seg og forbedre sin ytelse". Denne definisjonen vektlegger tre viktige aspekter: evnen til å utføre intelligente oppgaver, bruken av data for læring, og den iterative forbedringen over tid. Intelligens er her sett som mer enn bare problemløsning – det handler også om å kunne lære av erfaringer og forbedre ytelsen. "Kunstig" refererer til det faktum at intelligensen ikke er biologisk, men skapt av mennesker.

Likheter mellom menneskelig autonomi og maskinell autonomi

Menneskelig autonomi innebærer evnen til å ta beslutninger basert på fri vilje, erfaring og etikk, mens maskinell autonomi er basert på forhåndsdefinerte algoritmer og data. Selv om begge former for autonomi innebærer evnen til å handle uten direkte innblanding, er menneskelig autonomi forankret i bevissthet og moral, mens maskinell autonomi er begrenset til tekniske begrensninger.

Forskjell mellom begrepene automatisk og autonom

Begrepene "automatisk" og "autonom" begge refererer til systemer som fungerer uten mennesker. Mens "automatisk" ofte refererer til programmerte prosesser. "Autonom" betyr evnen til å ta beslutninger og tilpasse seg på egenhånd. Autonome systemer er dermed mer komplekse enn automatiske, da de innehar en form for evne til selvstyring.

Anmeldelse av pensumartikkel

I artikkelen "The 'Problem' with Automation: Inappropriate Feedback and Interaction, Not Over-Automation" utfordrer Donald A. Norman den tradisjonelle ideen om at over-automatisering er hovedproblemet med automatiserte systemer. I stedet hevder han at den virkelige utfordringen ligger i dårlig utformet feedback og interaksjon mellom mennesker og maskiner. Norman mener at manglende forståelighet og forutsigbarhet i systemenes handlinger fører til frustrasjon og ineffektivitet. Han argumenterer for at bedre design – med klarere informasjon og mer passende interaksjonsmuligheter – kan løse mange av de problemene vi i dag forbinder med automatisering.

Selskap som bruker AI

Et eksempel på et moderne selskap som jobber med kunstig intelligens er Meta. Meta beskriver AI på sine nettsider som en integrert del av flere av sine tjenester, inkludert

Facebook, Instagram, WhatsApp og Messenger. AI brukes til å forbedre brukeropplevelsen på disse plattformene, for eksempel ved å gi mer presise svar, lage bilder fra tekstbaserte forespørsler, generere forslag til kommentarer og gi inspirasjon til kreativt innhold. Denne sømløse integrasjonen av AI i tjenester er noe Holmquist argumenterer for i artikkelen «Intelligence on Tap: Artificial Intelligence as a New Design Material»

Film om AI

Filmen Her (2013) av Spike Jonze utforsker en fremtid der mennesker har nære og intime relasjoner med avanserte AI-systemer. Hovedpersonen, Theodore, utvikler et romantisk forhold til en AI-assistent, Samantha, som er designet for å tilpasse seg brukerens følelser og behov, og som kontinuerlig lærer og utvikler seg. AI i filmen blir fremstilt som svært sofistikert, ikke bare i evnen til å forstå og svare på spørsmål, men også i sin evne til å utvikle emosjonelle bånd.

Det som fascinerer meg med denne fremstillingen er hvordan AI her blir portrettert som en partner, men også som en selvstendig entitet med egen personlighet, tanker og utvikling. Denne formen for AI går langt utover det vi i dag forbinder med smarte assistenter som Siri eller Alexa. Her utforsker ideen om at AI kan bli så avansert at de blir mer enn bare verktøy – de blir følgesvenner med evnen til å ha dype, meningsfulle relasjoner med mennesker.

Personlig synes jeg det er både spennende og litt skremmende å tenke på hvor grensene mellom menneskelige relasjoner og teknologi kan viskes ut. Filmen reiser viktige spørsmål om hva det betyr å være menneske, og om teknologi kan fylle de samme emosjonelle og sosiale behovene som ekte menneskelige interaksjoner. Selv om jeg tror AI kan bidra til å forbedre livet på mange måter, er jeg skeptisk til å erstatte menneskelige bånd med teknologi, da det emosjonelle aspektet av slike relasjoner, slik som det vises i filmen, kanskje aldri helt kan matche det menneskelige samspillet.

1.2 – Robots and AI systems

Opprinnelse av ordet robot

Ordet "robot" ble først introdusert av den tsjekkiske forfatteren Karel Čapek i hans skuespill R.U.R. (Rossum's Universal Robots) fra 1920. Čapek brukte ordet til å beskrive

menneskelignende maskiner som ble produsert for å utføre arbeid for mennesker. Ordet stammer fra det tsjekkiske ordet "robota", som betyr tvangsarbeid eller slaveri.

Tre definisjoner av ordet robot

Ifølge Oxford Dictionary er en robot "en maskin som kan utføre komplekse serier av handlinger automatisk, spesielt en som kan programmeres av en datamaskin." Denne definisjonen legger vekt på robotens evne til å utføre oppgaver og understreker at den kan programmeres og automatiseres.

Ifølge Robot Institute of America (RIA) defineres en robot som en "reprogrammable, multifunctional manipulator designed to move material, parts, tools, or specialized devices through various programmed motions for the performance of a variety of tasks." Denne definisjonen fokuserer på robotens evne til å bli omprogrammert for ulike oppgaver, og den beskriver roboter som verktøy som kan manipuleres til å utføre spesifikke arbeidsoppgaver.

Basert på disse to definisjonene, kan en robot defineres som "en programmérbar maskin som kan utføre fysiske eller kognitive oppgaver autonomt eller semi-autonomt, og som kan tilpasses ulike miljøer og oppgaver gjennom omprogrammering." Denne definisjonen legger vekt på robotens fleksibilitet og tilpasningsevne, og inkluderer både fysiske og kognitive oppgaver. Når det gjelder chatbots, kan de i denne definisjonen regnes som en form for robot. Selv om de ikke utfører fysiske oppgaver, utfører de kognitive handlinger som å forstå og svare på tekstbasert kommunikasjon.

Forskjell mellom robot og AI

En robot og AI er nært knyttet, men de er ikke det samme. En robot er en fysisk maskin som kan utføre automatiserte oppgaver, ofte styrt av et kontrollsystem. AI, på den andre siden, refererer til teknologier som simulerer menneskelig intelligens gjennom læring, problemløsning og beslutningstaking. Begge kan utføre oppgaver autonomt og kan tilpasses til ulike formål. Dautenhahn argumenterer for i hennes artikkel «Some Brief Thoughts on the Past and Future of Human-Robot Interaction» at roboter må tilpasse seg menneskelige behov, som må forhåndsprogrammeres. AI kan selv lære og utvikle egenskaper innen menneskelig interaksjon. En robot er primært en fysisk enhet, mens AI er en teknologi som

kan være ren programvare. Roboter kan fungere med forhåndsdefinerte instrukser uten AI, mens AI er mer fokusert på å etterligne eller utvikle kognitive evner.

Ekte moderne robot

Spot er en firbent robot fra Boston Dynamics som kan navigere gjennom komplekse terreng ved hjelp av sensorer og kameraer. Spot beveger seg ved å bruke en kombinasjon av motorer og hydraulikk slik at den kan gå, løpe, klatre trapper og reise seg etter å ha falt. Når det gjelder interaksjon, kan Spot styres manuelt av mennesker via en fjernkontroll, men den kan også fungere autonomt. Spot brukes for eksempel i inspeksjon, der den kan bevege seg gjennom farlige områder og utføre oppgaver uten å sette mennesker i fare.

1.3 – Universal Design and AI systems

Regjeringen definerer universell utforming (Universal Design) som "utforming eller tilrettelegging av hovedløsningen i de fysiske forholdene, inkludert informasjons- og kommunikasjonsteknologi (IKT), slik at virksomhetens alminnelige funksjoner kan benyttes av flest mulig, uavhengig av funksjonsnedsettelse". Jeg forstår universell utforming som en designfilosofi som tar sikte på å inkludere alle brukere ved å skape løsninger som er fleksible, enkle å forstå og tilgjengelige for alle. Det handler om å redusere barrierer og skape inkluderende omgivelser der alle, uansett funksjonsnivå, kan delta fullt ut.

AI potensial for menneskelig persepsjon, bevegelse og kognisjon/emojonsforståelse

AI kan hjelpe mennesker med syns- og hørselsutfordringer gjennom løsninger som AI-baserte visuelle gjenkjenningssystemer (for eksempel seende apper for blinde) og taleteknologi som kan oversette lyd til tekst i sanntid. For mennesker med fysiske funksjonsnedsettelser kan AI-styrte eksoskjeletter forbedre bevegelsesfriheten. Eksoskjeletter kan hjelpe personer med nedsatt gangfunksjon med å gå igjen ved å bruke AI-algoritmer som tilpasser seg personens bevegelser. AI kan støtte mennesker med kognitive utfordringer ved å gi hjelpemidler som tilpasser seg deres behov, som digitale assistenter som kan minne om oppgaver eller hjelpe med organisering. AI kan også gi emosjonell støtte, for eksempel gjennom samtaler med ensomme mennesker.

Eksempler på AI-systemer som inkluderer brukere på nye måter

Seeing AI er en app fra Microsoft som bruker AI til å beskrive omgivelsene for blinde eller svaksynte. Den kan lese opp tekst, gjenkjenne ansikter, og identifisere objekter rundt brukeren, og gir dermed blinde en form for visuell persepsjon. Project Euphonia fra Google er et AI-prosjekt som forbedrer stemmegjenkjenning for personer med talevansker. Ved å trene AI til å forstå utydelig tale, kan personer med taleforstyrrelser enklere bruke stemmeaktiverte enheter.

Potensial for både inkludering og ekskludering

Som vist med eksemplene ovenfor kan AI gjøre teknologier mer tilgjengelige for mennesker med funksjonsnedsettelser, og dermed åpne opp for nye måter å delta i samfunnet på. Dette kan inkludere muligheten til å kommunisere, bevege seg fritt eller delta i arbeidslivet. Samtidig kan AI også ekskludere brukere hvis teknologien ikke er designet med universell tilgjengelighet i tankene. For eksempel kan AI-baserte ansiktsgjenkjenningssystemer være mindre presise for personer med mørkere hudtoner eller uvanlige ansiktstrekk, noe som kan føre til diskriminering.

WCAG og menneske-AI-retjningslinjer

I WCAG 2.1 og menneske-AI-retningslinjer refererer "forstå" og "forståelse" brukes for å sikre at teknologi er lett å bruke og tilgjengelig for alle brukere, uavhengig av deres ferdigheter. "Forstå" innebærer her at brukeren skal kunne navigere og bruke teknologien på en intuitiv og effektiv måte. Det innebærer klar kommunikasjon, enkel tilgang til informasjon, og at systemet er designet på en måte som er lett forståelig.

Kan maskiner «forstå»?

Forståelse i menneskelig forstand involverer ofte kontekst, erfaring og bevissthet – noe maskiner mangler. Maskiner kan simulere forståelse ved å følge algoritmer og data, men de har ikke subjektiv opplevelse eller bevissthet på samme måte som mennesker. Derfor kan vi si at maskiner kan etterligne forståelse, men de har ikke en dyp forståelse i menneskelig forstand.

1.4 – Human-AI Interaction Guideline – "Make Clear Why the System Did What It Did"

Denne retningslinjen handler om at AI-systemer bør forklare sine handlinger på en forståelig måte for brukerne, slik at de kan forstå hvorfor en bestemt beslutning ble tatt. Tenk på et AI-system som anbefaler filmer på en strømmetjeneste. For å oppfylle retningslinjen bør systemet forklare årsaken til anbefalingen, for eksempel "Denne filmen er valgt basert på dine tidligere valg av actionfilmer og filmer regissert av samme regissør."

HCI-designretningsliner

En velkjent HCI-designretningslinje er Nielsen's 10 Heuristics for User Interface Design. Noen av heuristikkene inkluderer: Tilbakemelding til brukeren (visibility of system status), brukerkontroll og frihet, samsvar mellom system og virkelighet. Begge sett med retningslinjer fokuserer på brukeropplevelse. Forskjellen mellom dem er at HCI-retningslinjene fokuserer mer på generelle brukergrensesnitt, mens Human-AI-retningslinjene er spesifikke for samspillet med AI. HCI-retningslinjene legger større vekt på brukerkontroll og enkelhet, mens Human-AI-retningslinjene tar opp kompleksiteten med AI-beslutninger og hvordan man kan presentere disse for brukerne. HCI og AI har generelt to ulike målsetninger, noe artikkelen «AI and HCI: Two Fields Divided by a Common Focus» vektlegger.

1.5 – Journal - Everyday engagement, interaction and use

Dag 1 og 2: Penn og papir: Å skrive med penn og papir var langsommere, og jeg måtte være mer bevisst på hva jeg skrev, ettersom det ikke var mulig å redigere like lett som på en datamaskin. Håndskriften gjorde også teksten mer personlig, men det tok lengre tid å få ordene ned på papiret. Jeg la merke til at jeg også organiserte tankene mine annerledes – mer lineært og strukturert – og jeg måtte være tålmodig med både tankene og pennen.

Dag 3 og 4: QWERTY-tastatur og skjerm: Med tastatur og skjerm var skrivingen mye raskere og mer effektiv. Det å kunne redigere underveis gjorde prosessen mer flytende og mindre stressende. Likevel merket jeg at jeg hadde mindre bevissthet om selve innholdet i det jeg skrev, fordi jeg var mer fokusert på hastigheten. Det føltes også mer upersonlig og distansert

fra tankene mine. Fordelen var at jeg enkelt kunne strukturere, lagre, og finne tilbake til tekstene. Hele prosessen var mer teknisk og praktisk.

Dag 5 og 6: Stemme-til-tekst: Å bruke stemmen til å skrive var en helt ny opplevelse. Det var raskt og effektivt når det gjaldt å få tanker ned, men det krevde mer etterarbeid i form av redigering, da tale ofte er mer ustrukturert enn tekst skrevet for hånd eller på tastatur. Å snakke fremfor å skrive var mer utfordrende å tilpasse seg måten å strukturere tankene på. Jeg følte meg mindre i kontroll over språket og mer avhengig av teknologien for å tolke mine ord korrekt.

1.6 – Characteristics of AI-infused systems

1. Usikkerhet og inkonsekvens

AI-systemer, spesielt de basert på maskinlæring, er probabilistiske. Deres oppførsel er basert på sannsynlighetsberegninger og kan derfor variere avhengig av input og kontekst (Amershi, 2019). Dette kan føre til uforutsigbare handlinger og resultater, noe som kan forvirre og frustrere brukere. AI-systemer kan avvike fra tradisjonelle brukervennlighetsprinsipper, som konsistens. De kan reagere forskjellig på samme input over tid, avhengig av oppdateringer til AI-modellen eller personlig tilpasning. AI-systemer er ikke perfekte og kan gjøre feil, noe som kan ha betydelige konsekvenser for brukerne. Dette krever strategier for å håndtere feil, for eksempel forklaringer og muligheter for korrigering.

2. Behov for forventningsstyring

Brukere har ofte urealistiske forventninger til AI-systemers evner (Kocielnik, 2020). Dette kan føre til skuffelse og redusert aksept av teknologien. Det er avgjørende å kommunisere tydelig hva systemet kan og ikke kan gjøre, samt hvor nøyaktig det er. Dette kan oppnås gjennom for eksempel nøyaktighetsindikatorer, eksempler og forklaringer. Brukere bør forberedes på at systemet kan gjøre feil, og at det er strategier for å håndtere disse. Dette kan bidra til økt toleranse for AI-systemers ufullkommenheter.

3. Dynamisk og adaptiv oppførsel

Mange AI-systemer lærer og tilpasser seg kontinuerlig basert på nye data (Yang, 2020). Dette kan føre til endringer i systemets oppførsel over tid, noe som krever strategier for å holde brukerne informert om disse endringene. AI-systemer kan tilpasses individuelle brukere basert på deres preferanser og oppførsel. Dette kan føre til en mer skreddersydd opplevelse, men krever også åpenhet om hvordan data brukes til personlig tilpasning. AI-systemer kan arve og forsterke skjevheter fra treningsdataene, noe som kan ha negative konsekvenser for brukere. Dette krever strategier for å identifisere og redusere skjevhet i AI-systemer.

4. Behov for brukerkontroll og transparens

Brukere må ha tillit til AI-systemer for å akseptere og bruke dem. Dette krever åpenhet om systemets funksjoner, begrensninger og potensielle skjevheter. Brukere bør ha kontroll over AI-systemets oppførsel og kunne korrigere feil. Dette kan oppnås gjennom for eksempel konfigurerbare innstillinger, tilbakemeldingsmekanismer og muligheter for å deaktivere AI-funksjoner. Brukere bør kunne få tilgang til forklaringer om hvorfor AI-systemet handlet som det gjorde. Dette kan bidra til økt forståelse og tillit til systemets beslutningsprosesser.
AI-infunderte systemer presenterer både spennende muligheter og unike utfordringer for designere. Ved å forstå og håndtere disse egenskapene kan vi skape brukervennlige og effektive AI-systemer som gagner brukerne.

Eksempel på AI-verktøy basert på retningslinjene

ChatGPT er et godt eksempel på et AI-infundert system som tydelig demonstrerer de nevnte egenskapene. Her er en diskusjon om hvordan disse karakteristikkene påvirker brukere av systemet:

Usikkerhet og inkonsekvens: ChatGPTs svar er basert på sannsynlighetsberegninger og kan derfor variere selv med lik input over tid. Dette skyldes AI-modellens kompleksitet og treningsdataene den er basert på. Brukere kan oppleve uforutsigbare resultater, og dette kan være forvirrende, særlig når systemet plutselig gir ulike svar på lignende spørsmål. For å håndtere dette, er det viktig at systemet kan gi forklaringer eller alternativer for brukerne, slik at de kan forstå hvorfor det oppstår variasjoner og korrigere feil om nødvendig.

Behov for forventningsstyring: Mange brukere har ofte høye forventninger til ChatGPT, noe som kan føre til skuffelse når systemet ikke oppfyller disse forventningene. Brukerne kan forvente at AI-modellen har perfekt kunnskap, selv om den er begrenset av treningsdataene. For å redusere denne typen skuffelse, er det viktig å kommunisere klart hva ChatGPT kan og ikke kan gjøre, og vise brukerne at selv om systemet er kraftig, er det fortsatt sannsynlig at det vil gjøre feil.

Dynamisk og adaptiv oppførsel: ChatGPT tilpasser seg kontinuerlig basert på tilbakemeldinger fra brukere og forbedringer til modellens treningsdata. Dette kan føre til endringer i oppførselen over tid, noe som kan være forvirrende for brukere som forventer konsistente svar. Det er derfor viktig å informere brukerne om eventuelle oppdateringer som kan påvirke hvordan modellen fungerer, slik at de vet hva de kan forvente.

Behov for brukerkontroll og transparens: For å opprettholde tillit, er det viktig at brukerne har en viss grad av kontroll over AI-systemet. For eksempel kan ChatGPT tillate brukerne å rapportere feil eller gi tilbakemelding som påvirker hvordan systemet oppfører seg fremover. Åpenhet om hvordan modellen tar beslutninger, hvordan treningsdataene kan påvirke svarene, og hvilke begrensninger modellen har, er avgjørende for å bygge brukernes tillit til systemet.

1.7 – Human-AI interaction design

Oppsummering fra artiklene av Amershi og Kocielnik

Både Amershi et al. (2019) og Kocielnik et al. (2019) tar for seg utfordringer og løsninger relatert til design av brukergrensesnitt for systemer som benytter kunstig intelligens. Begge artiklene understreker viktigheten av å skape brukervennlige og effektive AI-systemer som brukerne kan forstå og stole på.

Amershi fokuserer på å utvikle et sett med generelle designretningslinjer for interaksjon mellom mennesker og AI. Artikkelen presenterer 18 retningslinjer som er destillert fra over 150 AI-relaterte designanbefalinger fra både akademia og industri. Disse retningslinjene dekker et bredt spekter av temaer, inkludert; tydeliggjøre systemets evner og begrensninger, sette riktige forventninger, tillat brukerne å tilpasse AI-systemets oppførsel

og korrigere eventuelle feil, forklare AI-systemets beslutninger, og Bygg tillit ved å være åpen om systemets funksjoner og begrensninger.

Kocielnik et al. (2019) fokuserer på hvordan man kan justere brukerens forventninger til AI-systemers ytelse før de bruker systemet. Studien viser at brukere kan ha høyere aksept for systemer som fokuserer på å fange opp flest mulig relevante tilfeller, selv om det medfører noen falske positiver. Dette er i motsetning til systemer som fokuserer på høy presisjon, som kun identifiserer tilfeller de er svært sikre på, men som kan gå glipp av noen relevante tilfeller (falske negativer). Studien konkluderer med at forventningsjustering er en effektiv måte å øke brukeraksept for AI-teknologi.

Felles for begge artiklene er at de fremhever behovet for å ta hensyn til AI-systemers unike egenskaper, som usikkerhet, inkonsekvens og potensiell skjevhet, når man designer brukergrensesnitt. De understreker også viktigheten av åpenhet, brukerkontroll og forventningsstyring for å skape tillit og aksept for AI-teknologi.

LLM-er kan bli evaluert ved å bruke retningslinjene til Amershi et al. (2019): G6: "Vis kontekstuelt relevant informasjon". Denne retningslinjen er spesielt relevant for LLM-er, som ofte opererer med begrenset kontekst. Systemets evne til å vise relevant informasjon fra tidligere interaksjoner eller eksterne kilder kan forbedre samtaleflyten og brukerens forståelse. For eksempel referanse til tidligere utsagn. LLM-en kan bruke pronomen eller referanser som tydelig knytter seg til tidligere utsagn fra brukeren eller modellen selv. Dette bidrar til å opprettholde en sammenhengende dialog.

Ved å koble LLM-en til en kunnskapsgraf, kan systemet hente frem fakta og relasjoner som er relevante for samtalen. Dette kan berike svarene og gi brukeren mer dybdeinformasjon. For å øke transparens og tillit, kan systemet vise kildematerialet som ligger til grunn for genererte svar. Dette gir brukeren mulighet til å verifisere informasjonen og forstå modellens resonnement.

Den andre retningslinjen, G10: "Tillat brukere å omdirigere samtalen"

I en samtale med en LLM kan brukeren ønske å endre tema, stille et nytt spørsmål eller korrigere et tidligere utsagn. G10 understreker viktigheten av å gi brukeren fleksibilitet og kontroll over samtalens retning. Systemet bør gjenkjenne brukerens intensjon om å endre tema og tilpasse seg sømløst til den nye konteksten. Hvis brukerens input er tvetydig, kan LLM-en be om avklaring for å sikre at den forstår intensjonen. Systemet bør tillate brukeren å gå tilbake til tidligere svar for å korrigere misforståelser eller utforske alternative svarmuligheter.

Implementering av G6 og G10 kan forbedre eksisterende LLM-systemer på flere måter. G6 kan bidra til å redusere risikoen for hallusinasjoner ved å gi modellen mer kontekst og tillate brukeren å verifisere informasjonen. G10 kan gi en mer naturlig og dynamisk samtaleflyt ved å la brukeren styre retningen på samtalen.

Oppsummering av Bender's argumenter

Internett-data, som ofte brukes til å trene store språkmodeller, overrepresenterer yngre brukere og de fra utviklede land. Moderatorpraksis på nettsteder som Reddit, Twitter og Wikipedia kan gjøre dem mindre imøtekommende for marginaliserte befolkninger, noe som resulterer i at dataene reflekterer et begrenset sett med delpopulasjoner. Filtrering av datasett kan ytterligere dempe stemmene til personer fra marginaliserte identiteter.

Språkmodeller trent på statiske datasett risikerer å bli "verdi-låst", og reflekterer utdaterte og mindre inkluderende forståelser. Sosiale bevegelser, som ofte driver språklig endring og nye måter å kommunisere på, kan være underrepresentert eller feilrepresentert i treningsdata. Store språkmodeller har vist seg å ha ulike typer skjevheter, inkludert stereotypiske assosiasjoner og negativt sentiment mot spesifikke grupper. Disse skjevhetene gjenspeiler skjevhetene som finnes i treningsdataene.

I stedet for å bare fokusere på størrelse, bør betydelige ressurser brukes på å kurere og dokumentere treningsdata for språkmodeller. Dette inkluderer å lage datasett gjennom en gjennomtenkt prosess og aktivt søke å inkludere underrepresenterte samfunn. Størrelsen på språkmodeller øker deres miljømessige og økonomiske kostnader. Forskere bør prioritere

energieffektivitet og kostnader for å redusere negativ miljøpåvirkning og urettferdig tilgang til ressurser.

I stedet for å bare fokusere på å forbedre ytelsen, bør forskere fokusere på å forstå hvordan maskiner utfører oppgavene og hvordan de vil inngå i sosiotekniske systemer. Dette inkluderer å bruke metoder som pre-mortem-analyser og verdifølsomt design for å identifisere og håndtere potensielle risikoer og skader.
Forfatterne oppfordrer til et skifte i tankesett innen NLP-forskning, som prioriterer nøye planlegging, inkludering og ansvarlighet fremfor et ensidig fokus på størrelse og ytelse.

1.8 – Conversational user interfaces to large language models

Brukere har ofte høye forventninger til hva AI kan gjøre, spesielt når de interagerer med avanserte språkmodeller. Utfordringen er å balansere disse forventningene ved å tydelig kommunisere modellens begrensninger. Når brukere forventer feilfri kommunikasjon, kan de bli skuffet over modellens feil og misforståelser, noe som kan redusere tilliten til systemet.

Som probabilistiske systemer er språkmodeller som ChatGPT ikke perfekte og kan produsere både falske positive og falske negative resultater. Dette kan være problematisk i kritiske applikasjoner der nøyaktighet er avgjørende. Det å håndtere denne usikkerheten krever mekanismer for å forklare hvorfor feil oppstår, og i noen tilfeller kan brukerne ha behov for å rette opp eller overstyre modellens svar.

Språkmodeller etterligner menneskelig kommunikasjon, men mangler ekte forståelse. Dette kan føre til problemer som overtilpasning til konteksten, der modellen gir svar som kan virke overbevisende, men som egentlig er feil. Brukere kan bli forledet til å tro at modellen har en dypere forståelse enn den faktisk har, noe som skaper etisk problematikk rundt informasjonsformidling og beslutningsstøtte.

En annen utfordring er modellens evne til å forstå og håndtere kontekst gjennom samtalen. ChatGPT og lignende systemer kan ha problemer med å huske eller tolke lange eller komplekse samtaler korrekt, noe som kan føre til inkonsistente eller irrelevante svar.

Designere må finne løsninger for å håndtere hvordan kontekst blir lagret, forstått og brukt i sanntid, for å sikre en mer sømløs og sammenhengende brukeropplevelse.

For å øke tilliten til slike systemer, er det viktig at brukere har innsikt i hvordan AI-modellen fungerer. Dette kan oppnås ved å tilby forklaringer for modellens beslutninger, samt gi brukerne kontroll over visse aspekter av interaksjonen, som å justere tonen eller stilen i svarene. Mangel på transparens kan føre til mistillit, spesielt hvis brukerne ikke forstår hvorfor modellen gir visse svar.

Retningslinjene G1 og G2 fra Amershi et al. (2019)

Retningslinjene G1 og G2 kan bidra til å adressere flere av de sentrale utfordringene som i dag eksisterer i samtalegrensesnitt for store språkmodeller (LLM). G1, som omhandler å gjøre det klart hva systemet kan gjøre, handler om å gi brukeren en tydelig forståelse av modellens evner og begrensninger. Dette er viktig, ettersom LLM-er, som påpekt av Skjuve et al. (2024), ofte lider av fenomenet hallusinasjoner, der modellen genererer feilaktig eller oppdiktet informasjon. Ved å implementere G1, kan utviklere tydelig definere hvilke oppgaver modellen er egnet for, samt gi eksempler på vellykkede interaksjoner. Dette vil gi brukerne bedre innsikt i modellens faktiske kapasitet og redusere risikoen for at de stoler blindt på feilaktige svar. Gjennom å forklare hvilke typer input som fungerer best, kan også brukernes forventninger justeres, slik at interaksjonen blir mer målrettet og effektiv.

G2, som fokuserer på å gjøre det klart hvor godt systemet kan utføre de oppgavene det er designet for, kan bidra til å sette realistiske forventninger til modellens ytelse. I tråd med Bommasani et al. (2021), som fremhever utfordringene med skjevheter i LLM-er, kan transparens om modellens usikkerhet og potensielle feilkilder gjøre brukerne mer bevisste på når og hvorfor feil kan oppstå. Ved å visualisere usikkerhet eller gi et konfidensnivå for svarene, kan brukerne lettere vurdere informasjonen kritisk. Dette kan også redusere frustrasjonen når feil skjer, ettersom brukerne er forberedt på usikkerheten og forstår at LLM-er er probabilistiske systemer.

Disse retningslinjene kan også hjelpe med å bygge tillit mellom brukerne og systemet, som Noy & Zhang (2023) understreker er avgjørende for å overkomme brukernes mistillit til LLM-

er. Ved å være åpne om modellens begrensninger og tilby forklaringer på hvorfor enkelte feil oppstår, kan brukerne bedre forstå modellens beslutningsprosesser og utvikle en mer nyansert oppfatning av hvordan LLM-er fungerer.

Sammenfattende kan G1 og G2 bidra til å redusere problemer som hallusinasjoner, skjevhet og mangel på transparens i LLM-baserte samtalegrensesnitt. Ved å tydelig definere hva modellen kan og ikke kan gjøre, samt kommunisere dens usikkerhet og feilkilder, kan disse retningslinjene føre til en mer effektiv, pålitelig og brukervennlig interaksjon med store språkmodeller.

Appendix

Endringer fra tilbakemelding

Et forbedringspotensial med første iterasjon av denne oppgaven var å inkludere mitt personlige synspunkt på filmen «Her». I den opprinnelige teksten ble det kun beskrevet hvordan filmen presenterer en fremtid der mennesker utvikler intime relasjoner med AI-assistenter. Etter endringen reflekterer teksten nå også over personlige tanker og reaksjoner på filmens fremstilling av AI som emosjonelle partnere og selvstendige entiteter, samt hvilke etiske og samfunnsmessige implikasjoner dette kan ha.

Litteraturliste

- Amershi, S. (2019, 1. februar). Guidelines for human-AI interaction design. Microsoft. https://www.microsoft.com/en-us/research/blog/guidelines-forhuman-ai-interaction-design/
- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019).

Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19) (pp. 1-13). ACM. https://doi.org/10.1145/3290605.3300233

- 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)
- Bommasani, R. et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. (https://arxiv.org/abs/2108.07258)
- Dautenhahn, K. (2018). Some brief thoughts on the past and future of humanrobot interaction. Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI), 1-3. https://doi.org/10.1145/1228716.1228719
- 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)
- 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
- Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381. 187-192 (https://www.science.org/doi/full/10.1126/science.adh2586)
- Regjeringen. *Veileder for universell utforming* https://www.regjeringen.no/no/dokumenter/veileder-i-universell-utforming/id2850026/?ch=3
- Skjuve, M., Brandtzaeg, P. B., Følstad, A. (2024) Why do people use ChatGPT. First Monday. (https://firstmonday.org/ojs/index.php/fm/article/view/13541)
- Jordan, J. M. (2019, 29. juli). The Czech Play That Gave Us the Word 'Robot'. The MIT Press. https://thereader.mitpress.mit.edu/origin-word-robot-rur/
- 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)