IN5620
Individuell oppgave, høsten 2022

tinenor

Modul A

Konsepter, definisjoner og Interaksjon med AI og autonome systemer fra et historisk perspektiv

AI i et historisk perspektiv

Ifølge Jonathan Grudin (2009, s. 2) begynner historien om kunstig intelligens (heretter omtalt som AI) med det potensialet man så i datamaskiner i forbindelse med kodeknekking under andre verdenskrig. Den britiske matematikeren Alan Turing var tidlig ute med å spå at datamaskiner skulle kunne konkurrere med menneskelige intellektet på alle felt. Det var imidlertid amerikaneren John McCarthy som i 1955, sammen med et lite knippe andre akademikere, som først brukte begrepet «artificial intelligence» i innkallelsen til et forskningsprosjekt som gjerne omtales som «the Dorthmouth workshop» (McCarthy et al., 1955). Det tok mange år før Turings spådom skulle bli noe som likner på sann, men de seneste 25 årene har utviklingen innen AI revolusjonerende. Grudin peker på sjakksimulatoren Deep Blue som i 1997 verdensmesteren i sjakk, Gary Kasparov, som et mulig startskudd for en AI-æra som har gitt oss blant annet fjernstyrte rom-roboter, selvkjørende biler og anbefalingsalgoritmer (2009, s. 9).

Å definere AI

Det finnes utallige definisjoner på AI, og jeg vil her se nærmere på tre av dem. La oss først ta en titt på den originale definisjonen til McCarthy et al.: De beskriver det de kaller "the artificial intelligence problem" slik: "[..]how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (1955). Denne definisjonen må sees på i lys av at det ennå ikke fantes noen maskin som kunne løse dette problemet, men at McCarthy og hans kolleger hadde et ønske om å få det til. Definisjonen er derfor ikke veldig konkret, men peker på hvilke aspekter ved den menneskelige intelligens de ønsker å utforske med datamaskiner. Allikevel synes jeg at den tar for seg mange relevante egenskaper som også vil kunne kjennetegne mange av dagens AI-løsninger.

Bratteteig og Verne (2018, s.1-2) definerer 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". Det første man kan merke seg er at det har skjedd en signifikant endring i AI beskrevet som et *datasystem* fremfor en *maskin*. Denne definisjonen har allikevel flere fellestrekk med McCarthys. Begge to peker på det menneskelige, samt at maskinen skal løse et problem eller en oppgave. Men i motsetning til McCarthy befinner Bratteteig og Verne seg i en samtid der AI er et faktum. Og der McCarthy kunne sette skillet mellom menneske og maskin basert på at maskinen fortsatt ikke hadde inntatt oppgaver som krever intelligens, må Bratteteig og Verne sette skillet et annet sted. Sammen med deres ståsted som utøvere av det mer menneskesentrerte feltet participatory design er det kanskje derfor ordet *etterlikne* («mimic») blir brukt om maskinens (eller systemets) intelligens.

Eu-kommisjonens AI-ekspertgruppe (High Level Expert Group on AI) foreslo i 2019 følgende definisjon på AI: "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions" (European Commission, 2019). Som de selv beskriver er denne definisjonen ment å være en utvidet definisjon som skal tydeliggjøre visse aspekter ved AI som forskningsdisiplin, og den bærer også preg av dette. Ekspertgruppa er tydelig opptatt av å beskrive både AIs fysiske form, samt alle mulige måter et AI kan jobbe. I motsetning til de to foregående definisjonene nevnes ikke menneskelig intelligens eller problemløsning i denne definisjonen. Mennesket er kun angitt som designeren av systemet. Det er allikevel interessant at denne definisjonen, i sin detaljrike beskrivelse av hvordan AI fungerer, bruker ord som oppfatte,

tolke, resonnere og oppførsel. Dette kan sies å være ord som først og fremst brukes for å beskrive levende vesener, og som til en viss grad bidrar til å personifisere eller levendegjøre datasystemer. Det er også verdt å merke seg hvordan den siste setningen som handler om hvordan AI kan analysere konsekvensene av egne handlinger og basert på dette tilpasse sin oppførsel, kan ses på som en slags moderne speiling av McCarthys mer forenklede beskrivelse av maskiner som kan forbedre seg selv.

Basert på disse tre definisjonen gjør jeg følgende forsøk på å lage min egen definisjon av AI:

AI er et datasystem som, med utgangspunkt i et komplekst mål, samler data og basert på prosessering av denne utfører handlinger for å nå dette målet. Noen former for AI kan også tilpasse sine handlinger basert på konsekvensene av sine tidligere handlinger.

I min definisjon har jeg forsøkt å beholde McCarthys enkelthet, men jeg har valgt å holde mennesket helt utenfor. I stedet har jeg etterstrebet å ta med meg de viktigste aspektene fra ekspertgruppas beskrivelse av hva en datamaskin gjør, samt Bratteteig og Vernes beskrivelse av AI som et datasystem.

Autonomi

Menneskelig autonomi og maskinautonomi har mye til felles. Det handler gjerne om uavhengighet, men mens man i menneskelig sammenheng gjerne snakker om å handle uavhengig av andre eller bestemme over seg selv, så snakker man ofte om at autonome maskiner handler *uavhengig av mennesket*. Der menneskets autonomi ofte nevnes i samme setning som handlingsfrihet, slik Thilo Hagendorf også gjør det i artikkelen *The Ethics of AI Ethics: An Evaluation of Guidelines* (2020, s. 113, 114), beskrives imidlertid maskin- eller systemautonomi som «en funksjonalistisk orientering mot systemets egne koder» (Hagendorff, 2020, s. 109), altså om at maskinen følger gitte mønstre. Videre finnes det flere interessante problemstillinger knyttet både til hvor mye autonomi et system har i forhold til brukerens kontroll (Holmquist, 2017, s 33), men også hvordan menneskets autonomi begrenses av maskiner og systemer (Hagendorff, 2020, s. 110).

Autonomi og automatikk er to begreper som er beslektet, både etymologisk og i betydning. Når man snakker om automatiske maskiner snakker man gjerne om maskiner som fungerer av seg selv, uten direkte inngripen (Det norske akademis ordbok, 2022), argumenterer Sebastian Thrun for at autonomi i tillegg handler om maskinen (eller robotens) evne til å tilpasse seg variasjonene i sine omgivelser (2004, s. 14).

AI and HCI: two fields divided by a common focus

Artikkelen “AI and HCI: two fields divided by a common focus” av Jonathan Grudin (2009) er først og fremst et historisk perspektiv på hvordan HCI og AI har utviklet seg og sameksistert fra 1950-tallet og frem til artikkelen ble skrevet. Grudin argumenterer for hvordan de to fagfeltene har springer ut fra de samme teoriene, men har endt opp med å konkurrere om økonomiske midler og oppmerksomhet. Særlig peker han på hvordan entusiasmen og optimisme rundt AI har tatt mye oppmerksomhet i perioder, mens HCI har hatt sine glansdager når optimismen har bleknet og AI har vært nede for telling. Grudin skriver at dette har skapt et skille mellom de to fagfeltene som han allikevel gir uttrykk for at kunne ha tjent på å nærme seg hverandre. Avslutningsvis viser han også til tendenser på at dette er i ferd med å skje.

Jeg synes dette var et interessant perspektiv på de to disiplinenes historiske bakgrunn, som jeg ikke allerede var klar over. Ettersom artikkelen er over ti år gammel, blir jeg imidlertid nysgjerrig på hva som har skjedd siden. Den oppsvingen for AI som Grudin beskriver på begynnelsen av 200-tallet må jo kunne sies å ha fortsatt helt frem til i dag. Betyr dette at HCI igjen har kommet i skyggen av sin fascinerende bror eller har de to fagfeltene smeltet mere sammen og dermed latt HCI få være med på oppturen?

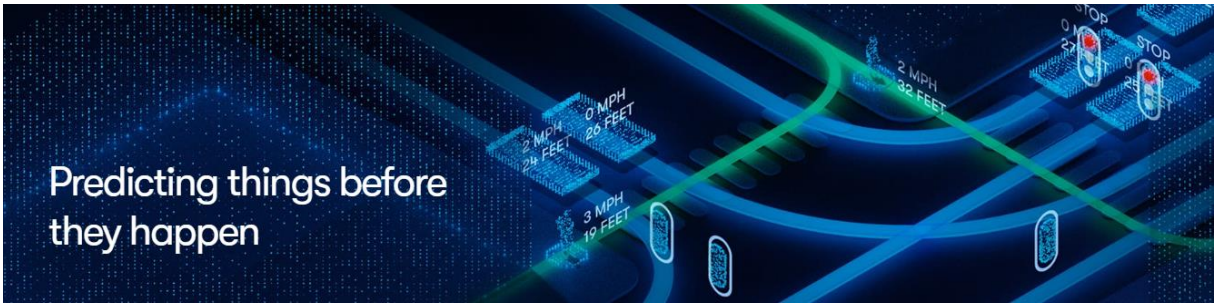
Waymo: “The world’s most experienced driver”

Waymo er kanskje et av de mest berømte og vellykkede prosjektene hva gjelder selvkjørende biler. Bedriften startet opp som et Google-prosjekt I 2009, og ble I 2016 et eget teknologiselskap. Siden 2020 har innbyggerne i Phoenix i Arizona kunnet booke en selvkjørende bil som tar dem dit de skal i

indre by. Waymo har også begynt å eksperimentere med selvkjørende trailere for frakting av varer (Waymo, u.å.).

På Waymos nettsider brukes aldri begrepet AI. De kaller det derimot konsekvent for «autonomous driving technology» og har gitt denne et navn; «Waymo Driver». Og Waymo Driver blir på førstesiden omtalt som verdens mest erfarne sjåfør («The world's most experienced driver»). Dette blir videre utdypet med at det er den samme sjåføren som kjører alle bilene («Same Driver. Different Vehicle.»).

Det er først og fremst en *tjeneste* Waymo Driver fremstilles som. Og i denne tjenestehistorien er det (naturlig nok) trygghet som står i sentrum. "Informed by unmatched experience and designed with safety at its heart, the Waymo Driver is our autonomous driving technology that never gets drunk, tired, or distracted" star det på nettsiden. Men for å sørge for å underbygge følelsen av trygghet hos både kunder og andre trafikanter, går de allikevel langt i å beskrive *produktet* Waymo Driver, og forklare hvordan teknologien fungerer. I segmentet som kalles «How it works» endrer den grafiske utformingen av nettsiden seg, og blir mer «teknologisk» med mørk bakgrunn og mengder av prikker som tydelig skal demonstrere mengden av data som ligger bak valgene Waymo Driver tar:



Figur 1: Skjermbilde fra Waymos "How does it work"-segment (Waymo, u.å)

Minority report: et 2002-perspektiv på kunstig intelligens

Minority report er en film fra 2002, regissert av Steven Spielberg. Handlingen er satt til år 2054, og det amerikanske politiet har gjennom en spesialavdeling ved navn Precrime utryddet kriminalitet. Ved å fange synene til tre mennesker med overnaturlige evner, kalt Pre-Cogs, kan Precrime stoppe kriminelle handlinger før de skjer. I løpet av filmen settes det imidlertid spørsmålstegn ved rettsikkerheten i det å arrestere noen som faktisk ikke har utført den kriminelle handlingen ennå, samt hvor sårbare noen blir i et samfunn der noen få har kontroll over et system som styrer over liv og død.

Selv om Precrime baserer seg på evnene til Pre-Cogene, og ikke AI, så er det lett å trekke linjer fra Minority Report og til det at man i noen land har tatt i bruk AI for å styrke forebyggende politiarbeid, ofte kalt «smart policing». Dette har skapt mange etiske diskusjoner som blant annet dreier seg om hvordan AI i noen tilfeller kan føre til at noen sårbare demografiske grupper blir ytterligere stigmatisert og mistenkeliggjort (Reese, 2022).

I tillegg til denne koblingen spiller AI også mange biroller i Minority Report. Eksempler på dette er blant annet biometrisk identifikasjon (iris-gjenkjenning), spesialtilpassede hologram-reklamer som kan navnet ditt (reklame-algoritmer) og edderkopp-roboter som går inn i utrygge lokasjoner (droner/feltroboter). Da jeg første gang så denne filmen i 2002, opplevde jeg disse som typiske science fiction-elementer som overdriver morgendagens teknologi, mens de i dag er mer gjenkjennelige enn utrolige.

Roboter og AI-systemer

Ordet

Fordi ideer om å lage menneskeliknende maskiner med autonomi har eksistert i menneskers fantasi (og dermed i kunsten) mye lenger enn teknologien som kunne virkeliggjøre ideen, var det også i kunsten ordet robot først oppsto. Ordet har sin opprinnelse i slaviske språk og på tsjekkisk betyr *robotnik* «arbeider» eller «slave». Det var den tsjekkiske kunstneren Josef Čapek som først brukte begrepet, og broren Karel Čapek populariserte det i 1920 med sitt teaterstykke *Rossumovi Univerzální Roboti* som handler om en fabrikk som lager menneskekopier kalt «roboti» (Store norske leksikon, 2022).

Å definere robot

Ettersom begrepet robot først oppsto i kunsten og mer spesifikt i sjangeren vi gjerne kaller science fiction, er det kanskje ikke så rart at det er stor spredning i definisjonene av begrepet. At fungerende roboter bare såvidt var ute av vuggen, er kanskje også grunnen til at the Robot Institute of America i 1979 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). På dette tidspunktet var roboter først og fremst maskiner som jobbet på fabrikk og som hadde tatt over komplekse, men repetative oppgaver som tidligere ble utført av mennesker. Definisjonen bærer tydelig preg av dette, i tillegg til at den rommer robotene som folk flest kanskje kjente best; nemlig de fra litteratur og film.

I dag definerer Store norske leksikon (SNL) en robot som “en maskin som kan utføre handlinger automatisk, og som ved hjelp av elektroniske «sanser» (sensorer) og «oppskrifter» på hva de skal gjøre (algoritmer) kan bearbeide inntrykk fra omgivelsene, som igjen påvirker hvordan disse oppgavene utføres. En robot består derfor både av programvare (software) som instruerer den i oppgavene, og maskinvare (hardware) som gir den en «kropp» å bevege” (Store norske leksikon, 2022). Denne definisjonen vitner om den utviklingen som har skjedd siden 1979, og er en mye mer konkret beskrivelse, som også utelukker mange maskiner som kunne falt inn under den forrige definisjonen. Der en enkel oppvaskmaskin kunne vært definert som en automatisk enhet som utfører en oppgave tidligere utført av mennesker, vil den ikke nødvendigvis møte alle kriteriene i SNLs definisjon.

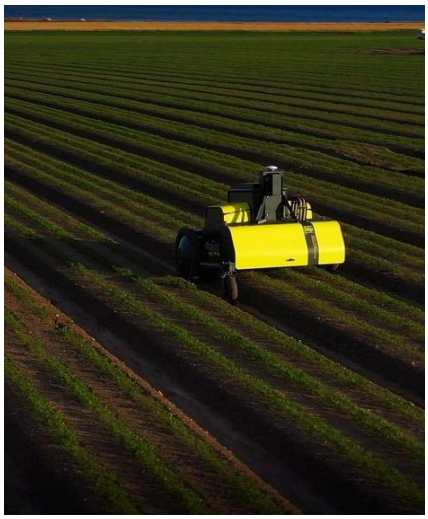
Videre vil jeg definere en robot som *en maskin som kan utføre fysiske oppgaver ved hjelp av sensorer og programvare uten direkte inngripen fra mennesker underveis*. Jeg tenker at en definisjon burde være mer konkret enn 1979-definisjonen, og at vi godt kan holde den menneskeformede maskinen utenfor dette. Jeg synes definisjonen til SNL gir mye mening, samtidig savner jeg en understreking av at oppgaven(e) som roboten utfører er fysisk(e). Jeg har også valgt å spesifisere at en robot utfører oppgaver uten menneskelig inngripen underveis for å utelukke fjernstyrte enheter, som jeg tenker at er et separat konsept.

AI vs. Robot

For å sammenlikne definisjonene mine på robot og AI, så ligger det en tydelig forskjell allerede i hvilket substantiv som er valgt; robot defineres som en «maskin» mens AI defineres som et «datasystem». Her tenker jeg at ulikheten ligger i den fysiske formen som en robot defineres av, men ikke AI. Det er allikevel også mange likhetstrekk, og det viktigste kan kanskje beskrives som autonomi. De utfører begge oppgaver eller handlinger, og de skal begge kunne gjøre det uten at mennesker må bidra underveis. For at noe skal kunne kalles AI kreves det allikevel kanskje en større grad av autonomi enn det en robot nødvendigvis må ha. Dette tydeliggjøres særlig med den siste setningen i AI-definisjonen: «Noen former for AI kan også tilpasse sine handlinger basert på konsekvensene av sine tidligere handlinger», som stemmer godt overens med Thruns definisjon som (en robots) autonomi som evnen til å tilpasse seg sine omgivelser (2004, s. 14). Robot og AI vil dermed på ingen måte være gjensidig utelukkende; en robot *kan* ha AI, men må ikke, og AI *kan* være en robot, men kan også være et datasystem helt uten en fysisk «kropp».

AX1: "The next generation of ultra-high precision weeding"

Jeg har valgt å se på den norske landbruksroboten AX1. Den bruker AI-teknologi og bildegjenkjenning til å sprøyte små doser ugressmidler direkte på ugresset og dermed ikke på avlingen. AX1 er det som ifølge Thrun kan defineres som en «profesjonell tjenesterobot», som vil si en robot som manipulerer og navigerer sine fysiske omgivelser og hjelper mennesker med å nå sine profesjonelle (2004, s. 12). Roboten bruker et GPS-spor som bonden lager når hun sår, og roboten beveger seg derfor akkurat slik andre gårdsmaskiner gjør. Bonden trenger dermed ikke å interagere så mye med roboten mens den jobber, og kan heller bruke tiden på noe annet. Det virker imidlertid som om det finnes en fjernstyringsfunksjon via programvare for nettbrett som kan brukes når AX1 skal flyttes, og jeg går ut ifra at man også bruker denne programvaren ved igangsettelse og avslutting (Kilter, u.å.; Kilter, 2021).



Figur 2: Landbruksroboten AX1. Bildet hentet fra https://www.kiltersystems.com/ax1

Universell Utforming og AI-systemer

Å definere Universell Utforming

§ 17 i lov om likestilling og forbud mot diskriminering: "Med universell utforming menes utforming eller tilrettelegging av hovedløsningen i de fysiske forholdene, inkludert informasjons- og www.kiltersystems.comkommunikasjonsteknologi (IKT), slik at virksomhetens alminnelige funksjoner kan benyttes av flest mulig, uavhengig av funksjonsnedsettelse" (Regjeringen, 2021). Jeg forstår denne definisjonen som at UU handler om at fysiske og digitale miljøer skal utformes på en slik måte at en størst mulig del av befolkningen skal kunne bruke dem, uansett hva slags fysiske eller psykiske forutsetninger de måtte ha. Klarer man å oppnå dette, vil man demokratisere disse miljøene og ikke ekskludere noen.

AI som inkluderer

Ettersom AI I stor grad handler om mønstergjenkjenning er det mange som ser potensiale i å bruke AI til å kjenne igjen både omgivelser, bevegelse og til og med menneskelige følelser, for så å gjenskape disse. For mennesker som har nedsatt syn, hørsel, bevegelse eller andre kognitive evner, kan teknologien da tilby muligheten til å støtte dem i hverdagen. Eksempler på dette er kameraer som ved hjelp av bildegjenkjenning kan fortelle hva som er foran linsa (Microsoft, 2022) eller proteser med innebygget AI for å forbedre brukerens kontroll over bevegelsene (Discover magazine, 2022).

Eksemplene over kan, hvis de fungerer godt nok, hjelpe til med å inkludere flere i større deler av samfunnet vårt. Samtidig finnes det også eksempler på bruk av AI som har vist seg å være ekskluderende. Et eksempel på dette er det tidligere nevnte fenomenet «smart policing» der AI bistår politiet med å forutse hvilke områder det er mest sannsynlig at kriminalitet vil oppstå basert på analyse av data. Men det har blitt pekt på at dersom politiet legger ekstra ressurser på å være til stede i visse områder, så vil de også finne mer kriminalitet der. Dette kan føre til at befolkningen i disse områdene (som gjerne i utgangspunktet er sårbare og til en viss grad ekskludert fra deler av samfunnets gode selskap) risikerer å bli stigmatisert og dermed ytterligere ekskludert (Reese, 2022).

Å forstå "forståelse"

Både i WCAG-rammeverket (W3C, 2018) og i Amershi et al. sine retningslinjer for menneske-AI-interaksjon (Amershi et al., 2019) brukes ordet «forstå» først og fremst om det mennesket som *bruker* systemet skal oppnå. Ut ifra disse kildene forstår jeg begrepet slik at det å *forstå* er mer enn å kun *oppfatte* noe. Å oppfatte innebærer å kunne sanse informasjon eller deler av grensesnittet, mens

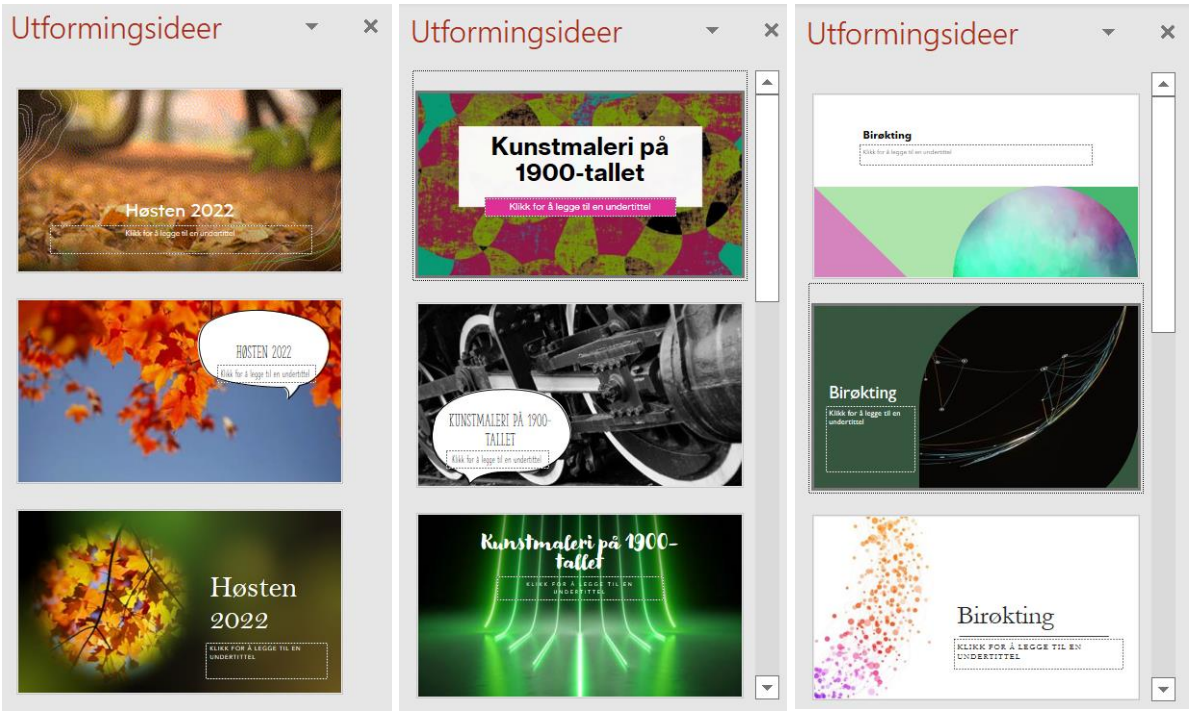
å forstå innebærer å kunne sette denne informasjonen i en sammenheng som igjen muliggjør at man kan bruke den til noe, som å forutsi noe eller unngå feil.

Videre vil jeg argumenter for at de fleste maskiner ikke har evnen til å forstå, da de først og fremst handler basert på sine forhåndsprogrammerte «mønstre». Flere og flere maskiner kan nå lagre informasjon, og kan dermed sies å være i stand til å «oppfatte», men denne informasjonen vil fortsatt ikke settes i noen annen sammenheng enn de tidligere nevnte mønstrene som maskinen handler etter. I svært forenklet forstand vil man kanskje allikevel kunne argumenter for at vi befinner oss i et vannskille av en tid, der AI er i ferd med å kunne sette ervervet informasjon i sammenheng og bruke den til å optimalisere sine handlinger, unngå å repetere sine feil og forutse hendelser. Dette kan altså betegnes som maskiner som kan forstå, men vi er nok fortsatt et langt stykke unna at maskiner kan forstå på samme måte som vi mennesker gjør det.

Guideline for Human-AI interaction

Jeg har valgt guideline nr 11: “Make clear why the system did what it did”, som er under kategorien “when wrong”. Jeg tenker at det er veldig viktig som designer å ha tenkt nøye igjennom hva som skal skje når et system feiler. Videre tenker jeg at særlig i arbeidet med AI er det viktig å vise en grad av transparens når systemet ikke oppnår det det skal. Dette vil kunne dempe den følelse av fremmedgjøring som AI i sine verste former kan skape.

Et eksempel er når presentasjonsverktøyet PowerPoint kommer med utformingsforslag med bilder. Det er lett å forstå hvorfor denne AI-programvaren har valgt bilder av høstløv når overskriften er «Høsten 2022» (se figur 3). Når man som bruker får et forslag med et bilde av en dampmaskin(?) til overskriften «Kunstmaleri på 1900-tallet» kunne det vært nyttig å vite på hvilket grunnlag dette bildet er valgt. Eksempelvis kunne forlaget vært merket med en beskjed i retning av «Dette forlaget er basert på følgende tekst: «1900-tallet». På samme måte kunne det vært nyttig å få hjelp til å forstå hvorfor overskriften «Birøkting» kun gir generiske bildeforslag. Er det fordi programvaren ikke kjenner igjen ordet? Eller fordi den ikke har tilgang på bilder som passer?



Figur 3: Utformingsformingsforslag fra PowerPoint

Jeg har tatt en titt på Don Normans seks designprinsipper (Design Principles, u.å.). Forskjellene mellom disse og Human-AI interaction-retningslinjene (Amershi et al, 2019) er mange. Eksempelvis

har sistnevnte et tidselement som man ikke kan finne i Normanns prinsipper. Dette med tid er både tydelig i inndelingen av prinsippene («initially», «during interaction», «over time»), men også flere av prinsippene til Amershi et al. referer til tidsaspektet. Prinsippet «Time services based on context» er kanskje det tydeligste eksempelet på dette. Videre er Amershi et al. sine prinsipper mye mer konkrete enn Don Normanns. Allikevel kan nok mange av Amershi et al. sine retningslinjer sies å bygge opp under de seks designprinsippene. Særlig vil mange av retningslinjene sørge for «feedback» som kan relateres til formuleringene «notify», «convey» og «make clear» i Amershi et al. sine retningslinjer. Også «consistency» har sine slektninger i retningslinjene i form av «match relevant social norms» og «update and adapt cautiously».

Modul B

AI-infunderte systemer - karakteristikker

Amershi et al. beskriver i artikkelen «Guidelines for Human-AI Interaction» (2019) at AI-infunderte systemer («AI-infused systems") som systemer som utnytter AI-funksjonalitet som er direkte eksponert for sluttbrukeren (s. 1). En slik definisjon kan gjøre det enklere å forstå hva et slikt system er, men hva karakteriser slike systemer? Med utgangspunkt i de tre artiklene «Guidelines for Human-AI Interaction» (Amershi et al., 2019), «Will You Accept an Imperfect AI? Exploring Designs» (Kocielnik et al.,2019) og «Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design» (Yang et al., 2020) har jeg kommet frem til fire hovedkarakteristikker for AI-baserte systemer. De er uforutsigbare, unøyaktige, de endrer seg over tid basert på brukers handlinger og de er ugjennomsiktige.

Unøyaktig

Amershi et al. er spesielt opptatt av utfordringer knyttet til AI-infunderte systemer og peker på hvordan dette leder til at *feil er vanlige* i AI-infunderte systemer (2019, s. 2). Også Kocielnik et al. er opptatte av begrensningene som AI-baserte systemer innebærer, og hvordan disse kan kommuniseres slik at sluttbrukerens forventninger kan justeres til et riktig nivå. På lik linje med Amershi et al., peker også de på hvordan slike systemer ofte gjør feil, eller «almost always operate at less than perfect accuracy». Dette forklares med at systemenes algoritmer er probabilistiske (2019, s. 2), som - satt på spissen - vil si at de driver med sannsynlighetsbasert «gjetting». Yang et al. skriver videre at selv enkle AI-applikasjoner kan gjøre «inference errors that are difficult to anticipate» (Yang, 2020, s. 1), og peker dermed på den neste karakteristikken som definerer AI-baserte systemer.

Uforutsigbar

Amershi et al. trekker frem at mange AI-komponenter er *inkonsekvente* og *uforutsigbare*, noe som kan forvirre brukeren og minske tilliten til systemet. Den iboende inkonsekvensen er, ifølge forfatterne "due to poorly understood, probabilistic behaviors based on nuances of tasks and settings, and because they change via learning over time" (2019, s. 1-2), noe også Yang et al. synes å være enige i når de skriver at AI-funksjoner er adaptive og at deres output kan «autonomously diverge at a massive scale» (2020, s. 10). Begge disse sitatene sender oss igjen videre til neste karakteristikk.

Endres over tid basert på brukers handlinger

Som Amershi et al. poengterte, vil AI-infunderte systemer endre seg gjennom læring over tid (2019, s. 2). Og også Kocielnik et al. nevner kort hvordan brukerens innvirkning på systemet er en egenskap som er både avgjørende og unik for AI-systemer (2019, s.3). Yang et al. fokuserer på AI som et utfordrende materiale for UX-designere å jobbe med, og peker på «AI's evolving capabilities and adaptive behaviors» som en viktig årsak til dette (2020, s. 8).

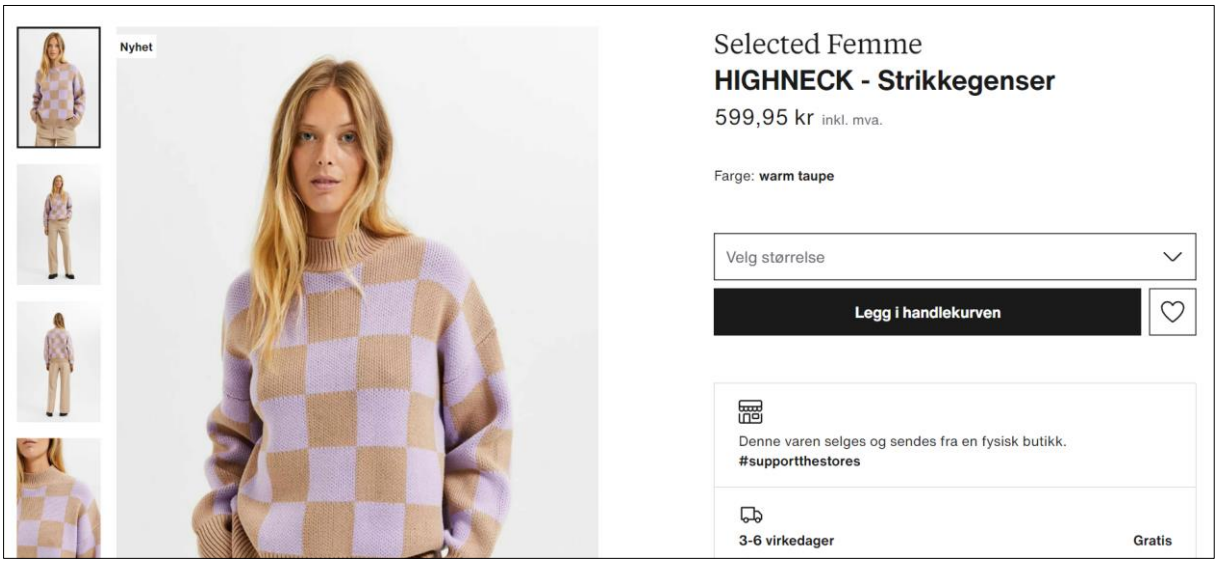
Ugjennomsiktig

Videre peker Amershi et al. på hvordan noen AI-teknologier gjør forsøk på personalisering gjennom å la systemet ta valg på vegne av brukeren, uten at det er eksplisitt synlig for brukeren hvordan dette skjer. En slik *«behind the scenes»-tilnærming* kan komme i konflikt med brukerens mål og forventninger dersom disse valgene ikke stemmer overens med brukers egentlige preferanser (2019, s. 1). På samme måte trekker Kocielnik et al. frem *transparens* som en utfordring med AI-baserte systemer som ikke har vært adressert nok i forskningen (2019. s. 3). I sin beskrivelse av AI som designmateriale skriver Yang et al. at mange designere finner det vanskelig å forstå hva AI kan og ikke kan gjøre (2020, s. 2) og kaller dette «capability uncertainty» (2020, s. 6). Selv om forfatterne også peker på designernes manglende erfaring med AI som materiale og teknologiens rivende utvikling, kan det kanskje også tenkes at dette handler litt om hvordan denne teknologien tar mange «valg» som ikke er synlige for de fleste av oss?

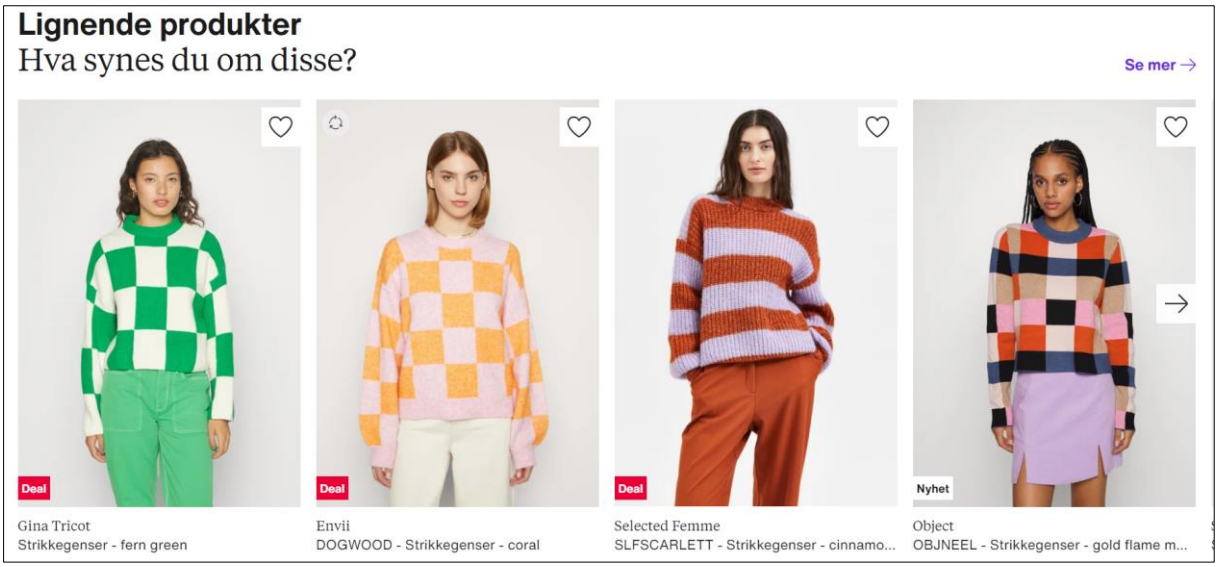
Utover dette, passer det også å nevne at disse karakteristikkene ikke vil kunne definere alle AI-baserte systemer, for som Amershi et al. skriver så vil kommersielt AI-design i stor grad variere, både med tanke på funksjonalitet og interaksjonstype (2019, s. 1-2).

Eksempel

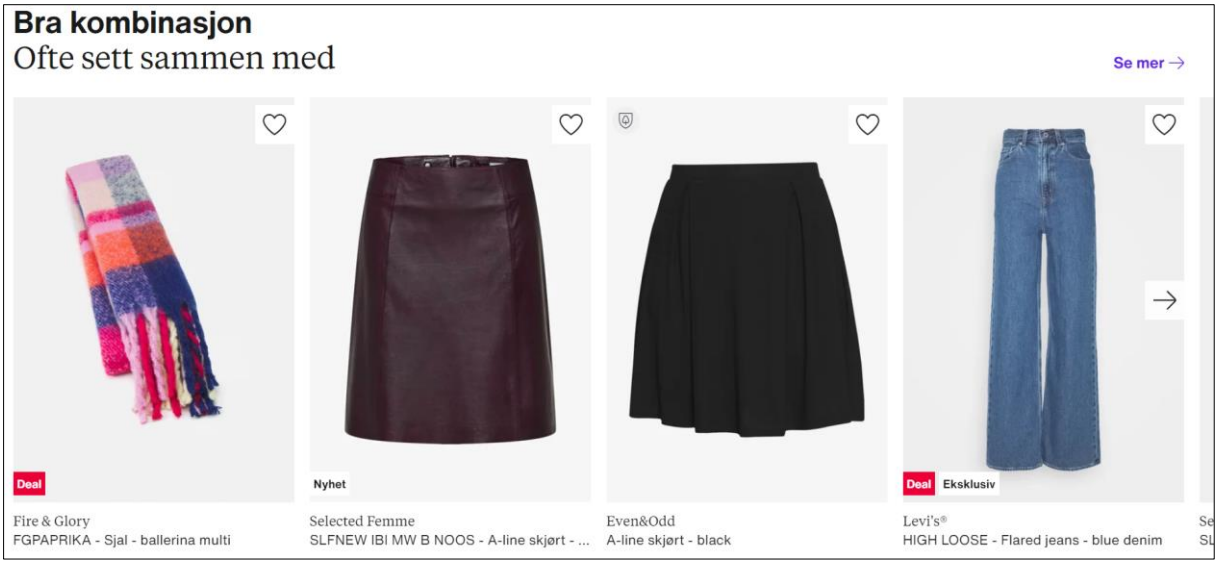
Jeg har tidligere nevnt utformingsideene til programmet Power Point, og ønsker i denne oppgaven å se på en liknende funksjonalitet, nærmere bestemt nettbutikken Zalandos «Liknende produkter» som jeg går ut ifra er både algoritme og AI-basert (bildegjenkjenning). Funksjonen går ut på at når du går inn på et produkt (som i figur 4), så vil du under få forslag til andre produkter som likner på dette (som i figur 5). Slik jeg oppfatter dem vil de produktene som er likest mulig i utseende, produsent, pris og materiale komme opp først.



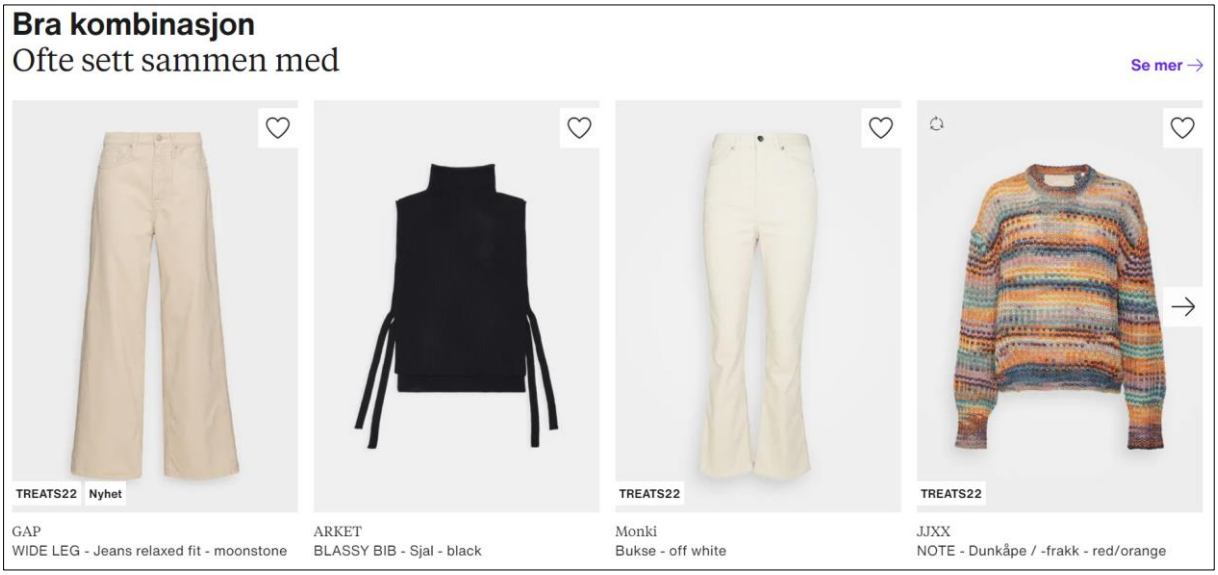
Figur 4 – produktvisning fra zalando.no



Figur 5 – produktforslag til produktet i figur 4



Figur 6 - produktforslag til produktet i figur 4



Figur 7 - produktforslag til produktet i figur 4

Må skrives litt mer ut:

Unøyaktig: erfaringsmessig gir forslagene veldig ofte mening, samtidig vil feil kunne skje. Særlig tenker jeg at målet til denne funksjonaliteten er å komme med forslag som brukeren kan like, så vil det variere hvilke parametre ulike brukere vektlegger. Zalando bøter på dette ved å gi flere forslag, og dermed øker sannsynligheten for at en bruker vil kunne finne noe interessant. Dette kan sies å være en måte å imøtekomme den tiende guidelinen til Amershi et al. (2019); «Scope services when in doubt»

Uforutsigbar: ønsker man å finne tilbake til et produkt som fant ved å trykke seg fra forslag til forslag, vil ikke det å gå tilbake til det produktet som ga deg forslaget første gangen nødvendigvis gi deg det samme forslaget igjen. Se figur 6 og 7.

Endrer seg over tid: relatert til det over, men jeg regner også med at algoritmene som lager forlagene også endrer seg, basert på brukeroppførsel.

Ugjennomsiktig: Jeg får ikke informasjon om hvorfor jeg får de forslagene jeg får. Som nevnt over kan jeg gjette på at jeg får forlag som har samme funksjon, samme materiale/kvalitet, er i samme prisklasse, har liknende mønster eller farge osv, men jeg vet det ikke sikkert…

Human-AI interaction design

Både Amershi et al. (2019) og Kocielnik et al. (2019) diskuterer interaksjonsdesign for AI-baserte systemer, og begge er interessert i brukerperspektivet i denne sammenhengen.

18 retningslinjer

Førstnevnte ser på disse systemenes brukervennlighet med utgangspunkt i 18 retningslinjer for design som de har destillert ut fra mer enn 150 anbefalinger for AI-baserte systemer. Videre har de evaluert disse retningslinjene gjennom tre runder med evaluering; en intern heuristisk gjennomgang, en brukerstudie der 49 brukere evaluerte en AI-drevet del av et produkt de kjente godt med utgangspunkt i retningslinjene, samt en ekspertvurdering av funnene fra brukerstudien. Amershi et al. (2019) mener dermed at disse retningslinjene vil bidra til mer menneskesentriske AI-systemer, dersom de blir fulgt.

Justere forventninger

Kocielnik et al. (2019) er litt mer spesifikke i sin tilnærming. De undersøker først og fremst hvordan man kan justere brukers forventninger til et system som ikke alltid er perfekt. De skiller mellom to typer feil; falsk positiv og falsk negativ, og peker videre på hvordan mange AI-baserte systemer er optimaliserte for å unngå falske positive (s. 3). Kocielnik et al. tar utgangspunkt i en AI-drevet planleggingsassistent for å både undersøke om brukere foretrekker et system som prioriterer å unngå falske positive treff eller et som prioriterer å unngå falske negative treff. De undersøker også om designelementer som skal justere brukers forventninger til systemets presisjonsnivå har noen innvirkning på brukeres opplevelse av systemet. Gjennom et eksperiment med 150 brukere, fant de at selv om de to variantene av planleggingsassistenten hadde samme presisjonsnivå, så oppfattet brukerne at systemet som prioriterer å unngå falske negative treff det som mer nøyaktig enn det som unngår falske positive. Brukerne aksepterte også dette systemet i høyere grad. Videre fant de at designelementer ment for å justere brukers forventninger kunne utlikne forskjellen mellom de to variantene.

Tilbake til Zalando

Med utgangspunkt i to utvalgte retningslinjer fra Amershi et al. (2019) sitt rammeverk, vil jeg igjen ta en titt på Zalandos produktanbefalinger.

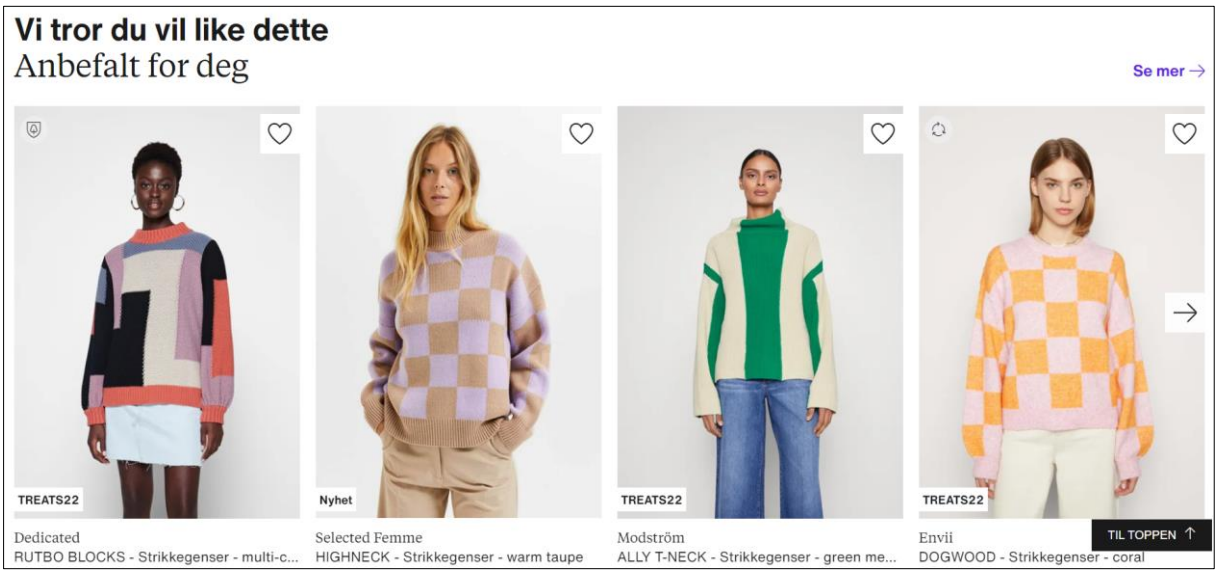
Time services based on context

Amershi et al. (2019) skriver at det er viktig å velge tidspunkt for handlinger eller avbrytelser med omhu. Brukers gjeldende oppgave og kontekst burde altså tas med i beregningen. Dette tenker jeg at Zalando gjør på en ganske god måte. Disse forslagene kommer på tidspunkt i handleprosessen da brukeren antagelig er ganske mottagelig for dem; når hen er inne på en produktside og ser på et gitt plagg. Samtidig er de plassert litt ned på siden, slik at bruker enkelt kan oppsøke dem ved å scrolle nedover, men uten at de forstyrrer informasjonen om det produktet som bruker faktisk har valgt å se nærmere på. Det er også flere kategorier av forslag; «lignende produkter» (se figur 4), «mer fra [produsent]» og «bra kombinasjon» (se figur 5). Noe som kan nå brukere med ulike «handle-mindsets».

Learn from user behavior

Amershi et al. (2019) skriver også at et AI-basert system burde personalisere brukeropplevelsen ved å lære fra brukernes handlinger over tid. Som tidligere nevnte, vet jeg lite om hvordan Zalando sine produktforslag utvikles over tid og hvorvidt de eventuelt baserer denne utviklingen på brukeres klikk og handlinger. Jeg vet imidlertid at disse forslagene spesialtilpasses til hver enkelt bruker innad i et enkelt besøk på siden. Dette blir spesielt tydelig om man trykker seg inn på Zalandos forside etter å

ha klikket litt rundt på produkter og får opp forslagskategoriene «vi tror du vil like dette» (se figur 8) og «dine beste deals». Antagelig vil disse forslagene også kunne overføres fra besøk til besøk dersom man lager en bruker og logger inn også.



Figur 8 - produktforslag på forside av Zalando.com

Stokastiske papegøyer og utfordringene som følger med

Bender et al. (2021) skriver i artikkelen "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" om hvordan stadig større språkmodeller har blitt tatt i bruk for å skape datasystemer som kan «sy sammen sekvenser av språklige skjemaer de har observert i sine enorme treningsdata, i henhold til probabilistisk informasjon om hvordan de kombineres» (s. 617). Forfatterne peker imidlertid på noen problematiske aspekter ved at bruken av slike systemer blir stadig mer utbredt.

Først og fremst peker de på *miljøkostnadene* og det økonomiske kost/nytte-regnskapet. De enorme mengdene data som trengs for å trene slike store språkmodeller koster både mye penger og karbonutslipp. Det settes også et ekstra tankekors ved det faktum at de delene av jordas befolkning som rammes hardest av konsekvensene av klimautslipp, ikke nødvendigvis får høste godene som slike systemer kan bringe, blant annet fordi språkmodellene ikke er tilgjengelige på deres språk (s. 612-613).

Videre trekker Bender et al. frem hvordan *treningsdataen til store språkmodeller også kan by på utfordringer* fordi den er så enorm og dermed og så uoversiktlig. De argumenterer for at dette kan resultere i modeller med «stereotype og nedsettende assosiasjoner når det gjelder kjønn, rase, etnisitet og funksjonshemming» (s. 613). De argumenterer for at det skjer en utsiling av hvem som blir representert i denne dataen som gjerne blir hentet fra ulike internettfora. Denne silingen gir hegemoniske synspunkter fortrinn og skjer i alle trinn av dataaggregering og -innsamling. På den måten kan ubalanse i makt og bias som skader marginaliserte grupper forsterkes (s. 614).

Forfatterne av denne artikkelen mener også at de store språkmodellene utgjør en risiko for forskningsfeltet som forsker på dem. Dette fordi de med sin tilsynelatende overbevisende funksjonalitet kan lede forskningsinnsatsen inn på blindspor eller avlede forskningen fra å nå det langsiktige målet om å lage systemer som forstår språk (s. 615-616).

Til slutt peker Bender et al. på hvordan disse systemene kan gjøre skade når de settes i bruk i «den virkelige verden» på grunn av alle de ovennevnte utfordringene, samt muligheten for mulig misbruk og potensialet for at vi mennesker tillegger denne kommunikasjon mening, når den egentlig bare er sannsynlighetsdrevet herming (og dermed ikke har mening) (s. 616-617).

Chatbots / conversational user interfaces

Chatbots are one type of AI-infused systems. Based on the lectures, and the mandatory articles, discuss key challenges in the design of chatbots / conversational user interfaces.

Utfordringer

Trekke inn de fire punktene fra over (må skrives ut):

Unøyaktig: mange veilednings-chatboter sliter nok med dette.

Uforutsigbar:

Endrer seg over tid: særlig relevant for mer avanserte samtale-chatboter

Ugjennomsiktig: viktig med tanke på de fleste chatboter.

Kan også trekke inn punktene fra Bender et al.:

(Miljøutfordringer: gjelder bare avanserte samtale-chatboter som bruker store språkmodeller?)

Treningsdata som forsterker maktubalanse og uheldige bias

Folks forventninger om mening/menneskelighet – veldig relevant. Kan også knyttes opp mot Kocielnik og våre forventninger til presisjonsnivå.

Mulige løsninger?

Revisit Guidelines G1 and G2 in Amershi et al. (2019). Discuss how adherence to these could possibly resolve some of the challenges in current chatbots / conversational user interfaces.

Make clear what the system can do:

vil særlig bøte på punktene unøyaktig og ugjennomsiktig

Make clear how well the system can do what it can **do:**

kan justere forventninger mtp unøyaktighet og uforutsigbarhet. Kanskje også forberede bruker på endringer?

Optionally, you may read Følstad & Brandtzaeg (2017), Luger & Sellen (2016), and Hall (2018) from the optional literature to complement your basis for answering.

Referanser

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for Human-AI Interaction (s. 1–13). ACM.

Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? (s. 610–623).

Bratteteig, T. & Verne, G. (2018). Does AI make PD obsolete? Exploring challenges from Artificial Intelligence to Participatory Design.

Det norske akademis ordbok (2022). Automatisk. Hentet den 21.10.2022 fra https://naob.no/ordbok/automatisk

Design Principles (u.å). Don Normans Principles of Design. Hentet 21.10.2022 fra https://principles.design/examples/don-norman-s-principles-of-design

Discover Magazine (2022, 21. Mars). AI Is Revolutionizing Prosthetic Arm Control. https://www.discovermagazine.com/technology/ai-is-revolutionizing-prosthetic-arm-control

European Commission (2019, 8. april). A definition of Artificial Intelligence: main capabilities and scientific disciplines. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines?fbclid=IwAR3oPKZL2dYvaN89dkNwbelVZLWyXOB7qtefz309Gc0IWmbLxac1ONJJHqk

Grudin, J. (2009). AI and HCI: Two fields divided by a common focus. The AI magazine, 30(4), 48–57.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and machines (Dordrecht), 30(1), 99–120.

Holmquist, L. (2017). Intelligence on tap: artificial intelligence as a new design material. Interactions (New York, N.Y.), 24(4), 28–33.

Kilter (u.å). *Kilter AX-1: The next generation of ultra high presicion weeding*. Hentet 21.10.2022 fra https://www.kiltersystems.com/ax1

Kilter (2021, 15. november). Kilter AX-1 @ Hasle gård [video]. YouTube. https://www.youtube.com/watch?v=842pkkoUlwM

Kocielnik, R., Amershi, S., & Bennett, P. (2019). Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems (s. 1–14). ACM.

McCarthy, J., Minsky, M., Rochester, N. & Shannon, C. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.

Microsoft (u.å.) *Seeing AI*. Hentet 21.10.2022 fra https://www.microsoft.com/en-us/ai/seeing-ai

Reese, H. (2022, 23. februar). *What Happens When Police Use AI to Predict and Prevent Crime?* JSTOR Daily. https://daily.jstor.org/what-happens-when-police-use-ai-to-predict-and-prevent-crime/

Regjeringen (2021, mai). *Veileder i universell utforming*. https://www.regjeringen.no/no/dokumenter/veileder-i-universell-utforming/id2850026/?ch=3

Spielberg, S. (regissør) (2002). Minority report. Twentieth Century Fox.

Store Norske Leksikon (2022, 12. mai). Robot. https://snl.no/robot

Thrun, S. (2004). Toward a Framework for Human-Robot Interaction. Human-computer interaction, 19(1-2), 9–24.

Waymo (u.å.). Hentet 21.10.2022 fra https://waymo.com/

W3C (2018, 5. juni). *Web Content Accessibility Guidelines (WCAG) 2.1*. 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 (s. 1–13). ACM.

Zalando (u.å.). Hentet 08.11.2022 fra https://www.zalando.no/selected-femme-strikkegenser-warm-taupe-se521i0t0-o11.html

Vedlegg

Vedlegg1

Hva jeg har gjort med tilbakemeldingen jeg fikk

Jeg var så heldig at jeg fikk to tilbakemeldinger grunnet litt forvirring rundt listene over hvem som skulle gi tilbakemelding til hvem. Ettersom jeg ikke rakk å gjøre meg helt ferdig med oppgaven før innlevering og hadde en svært tvilsom kildeføring, gikk mange av tilbakemeldingene på dette. Jeg har derfor gjort ferdig det jeg ikke rakk til første innlevering og ryddet opp i kildehenvisninger og referanselista. Jeg fikk også tilbakemelding på at et av eksemplene mine om hvordan AI kan diskriminere kanskje var litt på siden av temaet, og har derfor valgt å ta bort dette.