UiT Norges arktiske universitet

Fakultet for Humaniora, Samfunnsvitenskap og Lærerutdanning

Ikke-tenkende: En Swedbergiansk analyse av kunstig intelligens i kommunikasjon

Kan kunstig intelligens delta i kommunikasjon, sett i lys av kommunikasjonsbegrepene til Niklas Luhmann og Jürgen Habermas?

Jonas Parcival Kjæret

Masteroppgave i Lektorutdanning 8-13, SOS-3981, Juni 2023

Innhold

1 Innledning 3

2 Metode 8

3 Teori 12

3.1 Jürgen Habermas' kommunikasjonsbegrep 13

3.2 Niklas Luhmanns kommunikasjonsbegrep 17

4 Espositos forklaringsmodell av KI i sosiologi 21

5 Kunstig intelligens: Inspill til ny typologi 27

5.1 Maskinlæring: Sosiologisk typologisk lesning 31

5.2 Tre nivåer av kunstig intelligens 34

6 Analyse av Habermas og Luhmann opp mot min typologi 39

6.1 Habermas og min typologi 39

6.2 Luhmann og min typologi 44

6.3 Avsluttende argumentasjon av analysen 49

7 Konklusjon og veier videre 50

8 Referanseliste 53

Forord

Jeg vil aller først rette en takk til mine foreldre som har vært tålmodige, hjelpsomme og gjort det lett å være meg. Videre vil jeg rette en takk til alle mine gode venner, både i nord og i sør, som har villig lyttet mens jeg har utredet om kunstig intelligens og modeller som for mange har gått hus forbi. Jeg vil rette en spesiell takk til alle de på lesehuset øst, som har ledd og grått i tide og utide. Særlig vil jeg rette en takk til min venn Runar for hans kløkt og inngripende spørsmål; uten ditt blikk og dine gode analytiske egenskaper hadde jeg ikke vært der hvor jeg er i dag. Til slutt vil jeg og rette en takk til min veileder Truls Tunby Kristiansen som har hjulpet meg med å forstå en av de mer utfordrende sosiologer, Niklas Luhmann.

1 Innledning

«It's not a knife : a knife can't decide whether to kill someone or to save life in surgery. The decision is always in human hands. AI is the first tool that potentially can replace us in decision making» – Yuval Noah Harari (Grallet og Pons, 2023)

De siste månedene har det skjedd store fremskritt innen utviklingen av kunstig intelligens. Den 30. november 2022 lanserte OpenAI: ChatGPT (OpenAI, 2023). ChatGPT er en stor språkmodell (large language model). I korte trekk er en stor språkmodell den mest avansert form for kunstig intelligens vi har i skrivende stund. Store språkmodeller bruker avanserte algoritmer og maskinlæring for å forutsi og generere naturlig språk basert på store mengder tekstdata. Disse modellene kan brukes til oppgaver som tekstgenerering, maskinoversettelse og spørsmål og svar. Jeg ble først kjent med ChatGPT i januar i år. Det tok kort tid før jeg forundret meg over hvor utrolig kraftig denne språkmodellen var, og uavhengig av småfeil her og der, ga den informasjon og forståelse hinsides mine villeste fantasier. Det var dog en ting som jeg forundret meg over: Kan ChatGPT tenke? Hvis ChatGPT kan tenke har det store ringvirkninger for mitt kommende yrke som lærer og kommende generasjon sosiologer.

Denne ideen førte meg videre til tanken om at et av de viktigste spørsmålene i forskningen på kunstig intelligens for sosiologien bør være: Er det en forskjell på kunstig intelligens og mennesker? Foreløpig handler samfunnsforskning generelt, og sosiologien spesifikt, om hvordan mennesker i samfunn interagerer og begår handling ilag. Om kunstig intelligensmodeller blir avanserte nok til at de virker som mennesker må de og ved nødvendighet være en del av forskningsfeltet til sosiologien. Når en skal prøve å definere «hva er mennesket?» tas ofte Kant sitt kjente sitat: «Cogito ergo sum» opp som forklaringsmodell: Jeg tenker derfor er jeg. Det er to sentrale ord i denne påstanden: Jeg og tenker. Foreløpig kan store språkmodeller som ChatGPT benytte seg av presens jeg, men kan den tenke? Om den kan tenke er det heller ikke utenkelig å påstå at slike kunstige intelligensmodeller, som ChatGPT, vil være tilsvarende et menneske. Det er slike situasjoner som kan gjøre om på samfunn.

For å se på denne utfordringen var en observasjon jeg gjorde meg at det tilsynelatende foregår kommunikasjon mellom ChatGPT og meg som menneske. Det første og mest åpenbare spørsmålet å stille seg er: Er det *kommunikasjon* mellom mennesker og kunstig intelligens(er)? Deretter kan en stille seg spørsmål som: Hva slags type kommunikasjon er

det? Og hvordan foregår denne kommunikasjonen? Ifølge en klassisk definisjon (se for eksempel Al-Fedaghi (2012) for avklaring av overføringsmodellen til Shannon-Weaver) ser det tydelig ut som kommunikasjon skjer, men menneskelig intuisjon tilsier at det ikke faktisk er samme «kommunikasjonen» slik den forekommer mellom to mennesker. Er dette tilfellet? Hvordan kan vi lage et slikt skille, dersom intuisjonen vår skulle stemme? Et mer avslørende spørsmål med dette i tankene er: Hva er kommunikasjon?

Disse spørsmålene ledet meg I retning av særlig to sosiologer: Niklas Luhmann og Jürgen Habermas. Dette fordi begge sosiologene har skrevet utfyllende om begrepet kommunikasjon, og kommunikasjon er en integrert og uatskillelig del i deres samfunnsforståelse. Etter et litteratursøk, viste det seg at Luhmann har en protigé innen feltet kunstig intelligens: Esposito. Esposito benytter seg av Luhmann sitt rammeverk og beskriver kunstig intelligens koblet til kommunikasjon (Esposito, 2022). Det som derimot ikke forekommer hos Esposito er nyvinningen ChatGPT. Det sees derfor som hensiktsmessig å sette Espositos argumentasjon i lys av den pågående utviklingen innen kunstig intelligens. Underveis i skrivingen av denne oppgaven dukket nykommeren «GPT-4» opp, som i all hovedsak er en forbedring av ChatGPT (OpenAI, 2023). Sosiologiens forskningsområde er mennesker i samfunn, og en vital del er derfor hvordan kommunikasjon oppfattes og foregår. Hvordan kunstig intelligens interagerer med og innbefattes inn i samfunnet er derfor et svært viktig felt hvor sosiologien har flere mangler. Blant disse mangelene, er mangelen på en grunnleggende analyse av variasjonen som eksisterer innen kunstig intelligens feltet. En kan argumentere for at variasjonen av kunstig intelligens er langt større en variasjonen av mennesker. Dette må komme frem når en snakker om kunstig intelligens i sosiologien. Det må dog nevnes at Liu (2021) sitt bidrag, som vil bli nevnt senere, er av uvurderlig verdi. Den begynnelsen på et rammeverk som Liu har laget, er en av grunnen til at det dypdykket jeg gjør i denne studien er mulig.

Det som prøves gjort i denne oppgaven er derfor todelt: 1. å lage et rammeverk for å bedre forstå feltet kunstig intelligens i sosiologien, og 2. prøve ut dette rammeverket ved å koble det opp mot Habermas- og Luhmanns kommunikasjonsbegrep. Jeg har valgt meg disse to sosiologiene da de på hver sin side har fruktbare begreper for å forklare kommunikasjon som foregår mellom individer. Med en slik radikal endring som kunstig intelligens-forskningen bringer med seg, trengs det og nye måter å se på og analysere allerede eksisterende rammeverk. Min problemstilling går dermed som følger: «Kan kunstig intelligens delta i

kommunikasjon, sett i lys av kommunikasjonsbegrepene til Niklas Luhmann og Jürgen Habermas?».

Videre er forskeren inspirert av Swedberg (2014) sin metodologiske tilnærming til sosiologi. Denne metodologien vil bli mer utfyllende presentert i metodekapittelet, og kan forstås som en Swedbergiansk teoretiseringsprosess. I korte trekk går metodologien ut på å gjøre seg en observasjon som deretter navngis, etterfulgt av utvikling av analyseverktøy som kan brukes for å analysere gitt observasjon (Swedberg, 2014). Denne masteroppgaven er, som Swedberg definerer, en «førstudie» (Swedberg, 2014, s. 25-28). Denne førstudien vil innbefatte kommunikasjonsbegrepet til Niklas Luhmann forstått gjennom Baraldi, Corsi og Esposito (2021) og kommunikasjonsbegrepet til Jürgen Habermas. I denne oppgaven vil det gjøres et dypdykk i hva kunstig intelligens er, samt lages en typologi som vil gi et bedre grunnlag for sosiologisk forståelse av hva kunstig intelligensmodeller bygger på. Selv om sosiologien har klare mangler, har den også særegnede styrker. Spesielt er styrken inkludering/kombinering fra andre forskningsfelt noe sosiologien burde trekke oftere på. Som jeg vil argumentere for senere, er nettopp inkludering av KI som studiefelt en styrke, eller kanskje mer en nødvendighet, for å forstå og analysere moderne samfunn.

Sagt på en annen måte, potensialet som finnes i kunstig intelligens er uant og enormt. Sosiologien bør derfor delta aktivt i forskningen av kunstig intelligens, og samtidig inkludere fra andre fagfelt det som lar seg inkludere for å bedre våre sosiologiske forklaringsmodeller. På lik linje er det viktig å ikke innsnevre for mye, og for tidlig. Denne førstudien søker dermed å fremme et rammeverk for videre forskning på feltet kunstig intelligens i sosiologien. For å immøtekomme Swedbergs metodologiske tilnærming har jeg innhentet mine data på følgende måte. Først gjorde jeg et bredt søk i Oria på sosiologi og kunstig intelligens. For å lære mer om begrepet og forskningen rundt kunstig intelligens søkte jeg på maskinlæring, machine learning, artificial intelligenc, AI, KI og kunstig intelligens. Med bakgrunn i mitt begrep ønsket jeg å forstå hvordan oppbyggingen av kunstig intelligens er beskrevet i KI-feltet. Slik fant jeg Ongsulee (2017) og Sarker (2021) sine artikler. For å studere fenomenet har jeg diskutert med ulike typer KI. I litteratursøk etter informasjon om kunstig intelligens koblet til sosiologi havnet jeg i en sidegate. Denne sidegaten førte til en god del interessante funn.

For det første, vanskeligheten med å definere hva kunstig intelligens faktisk er. Det finnes svært mange tilnærminger, og alle med sine styrker og svakheter. Esposito (2022) har sin

tolkning av kunstig intelligens som i all hovedsak kun dreier seg om forskjellige typer algoritmer. Derimot, hvis en går til Sheikh, Prins og Schrijvers (2023) ser en at det finnes gode argumenter for at en ikke burde kategorisere kunstig intelligens kun som algoritmer (s. 15). Uten å gå i dybden på ressonementet, hevder Sheikh, Prins og Schrijvers at kunstig intelligens-feltet har flyttet definisjonen, av kunstig intelligens, ettersom nyere teknologi har kommet til (Sheikh, Prins og Schrijvers, 2023, s. 15-20). Dette innebærer at hver gang et nytt gjennombrudd i feltet har dukket opp, har en forskjøvet hva kunstig intelligens må kunne oppnå for å tilpasses den gitte definisjonen. Som Sheikh, Prins og Schrijvers skriver: «Pamela McCorduck kaller dette for 'KI-effekten': Så snart en datamaskin finner ut hvordan man gjør noe, erklærer folk at det 'bare er en beregning' og ikke faktisk intelligens» (Sheikh, Prins og Schrijvers, 2023, s. 17; min oversettelse).

Med denne innsikten har jeg derfor valgt å følge samme definisjon av kunstig intelligens som Sheikh, Prins og Schrijvers legger frem. Denne definisjonen er å finne på side 19 hos forfatterne, men jeg velger å bruke definisjonen i sin helhet slik den er beskrevet av High-Level Expert Group AI (2018), og går som følger:

> «Artificial Intelligence refers to systems that display intelligent behaviour by analysing their environment and taking action — with some degree of autonomy — to achieve specific goals» (High-Level Expert Group AI, 2018)

Det er denne definisjonen jeg kommer til å benytte meg av, og følgende oversatt til norsk: «Kunstig intelligens referer til systemer some viser intelligent oppførsel ved å analysere dets miljø og med det gjøre handling – til en viss grad autonomt – for å oppnå spesifikke mål». Dermed, når jeg i løpende tekst kun refererer til kunstig intelligens, er det denne definisjonen som ligger til grunn for argumentasjonen. Men, som det i oppgaven vil bli utførlig diskutert, er kunstig intelligens mer omfattende enn som så.

For det andre, det trengs en tydelig begrepsavklaring når en har med kunstig intelligens å gjøre. Jeg vil i senere kapittel forklare om variasjonen og typer av kunstig intelligens, men før det vil det her komme en tydeliggjøring av visse begreper:

1. Modell: En kunstig intelligensmodell er et representasjonssystem som er designet for å imitere et bestemt fenomen eller prosess. Det kan være omfattende (som ChatGPT som består av mange ulike komponenter) eller det kan være simpelt (som filtrering av informasjon som består av færre komponenter).

2. Teknikk (også kalt metode i noen foraer): En kunstig intelligens teknikk omhandler en spesifikk måte at noe gjøres på. Disse kan variere i teknikalitetsgrad, noen er vanskeligere å forstå enn andre, samt at noen kan kombineres for å forbedre en kunstig intelligensmodell. Det finnes i skrivende stund fire slike overordnede teknikker som vil bli utførlig beskrevet i kapittel 5.1.
3. Kunstig intelligens: Kunstig intelligens omfavner to hovedtyper, ekspertsystemer og maskinlæring. Det er maskinlæring som er hovedfokus for denne oppgaven.
4. Maskinlæringsnivåer: Det finnes i skrivende stund tre forskjellige nivåer innen maskinlæringsfeltet 1. overordnet nivå (nevnt som «simpel» maskinlæring senere i oppgaven), 2. nevrale nettverk, 3. dype-nevrale nettverk. Disse vil bli utførlig beskrevet i kapittel 5.2.

For det tredje, det er svært få artikler om emnet kunstig intelligens i sosiologiske termer. Blant artiklene jeg fant av kunstig intellgiens og sosiologi, er det særlig Rolf Lidskog sin «Samhället utmanat?» (2020) og Zheng Liu sin artikkel «Sociological perspectives on artificial intelligence: A typological reading» (2021) som utmerker seg. Liu gir tre typer betraktninger på hvordan å forstå kunstig intelligens i sosiologien, det han kaller «Scientific AI» (s. 4-5), «Technical AI» (s. 6-8), «Cultural AI» (s. 8-9). Samtlige av studiene Liu peker på, får frem viktige poenger om hvordan mennesker og algoritmer interagerer med hverandre, samt hvordan algoritmer kan spille inn og påvirke samfunn. Det som derimot tilsidesettes er variasjonen som finnes i kunstig intelligens, og som jeg har pekt på: Kunstig intelligens forstått kun som algoritmer. Det må allikevel presiseres at Liu sin artikkel er mye av grunnen til at denne masteroppgaven ble til. Dette kommer av Liu sin betraktning av kunstig intelligens i sosiologien som: «… a research field in its own right, involving contributions by scholars from across the full spectrum of sociology to debate AI's relations with and impacts on all major concerns of the discipline» (Liu, 2021, s. 9). Denne masteroppgaven er således et forsøk på en videreføring og presisering av forskningsfeltet (og typen) som Liu beskriver som «teknisk kunstig intelligens» (2021, s. 6-8).

Den andre artikkelen som har vært formativ for min forståelse og analyse av feltet kunstig intelligens i sosiologien er Rolf Lidskog sin artikkel «Samhället utmanat?» (2020). Lidskog argumenterer for det som oftest tas opp når det er snakk om kunstig intelligens og samfunn: en superintelligens som kan ta over for mennesker. Lidskog bygger på Tegmark sin bok «Life 3.0» (2017) og Nick Bostroms bok «Superintelligence: Paths, Dangers, Strategies» (2014).

Som med Lidskog, legger Bostrom og Tegmark ut om forskjellen i hovedtypene smal-kunstig intelligens og generell kunstig intelligens. Denne fremgangsmåten er både produktiv og fruktbar, men har sine begrensninger. Jeg vil følgelig argumentere for at det trengs større innsikt i hva grunnpilarene til kunstig intelligens er, samt utrede om den store forskjellen som eksisterer innad i feltet kunstig intelligens. Videre argumenterer jeg for at det trengs en tydeligere nyansering av det som i realiteten er maskinlæringsteknikker. Med dette som bakgrunn prøver jeg i denne masteroppgaven derfor å utarbeide en typologi som kan brukes for å analysere sosiologiske fenomen, slik som kommunikasjon. Med bakgrunn i en Swedbergiansk teoretiseringsprosess.

2 Metode

Hva er så en Swedbergiansk teoretiseringsprosess? Det er en metodologisk fremgangsmåte gitt av Swedberg i hans bok: «The Art of Social Theory» (Swedberg, 2014). Den metodologiske fremgangsmåten som er valgt for denne masteroppgaven er basert på Swedberg sin analyse og teori om hvordan å teoretisere i sosiologien. Jeg ble først kjent med forskningsmetoden i forelesning med Gunnar C. Aakvaag og Roar Hagen (på den tiden begge professorer i sosiologi ved UiT Norges Arktiske Universitet). Før metoden presenteres vil jeg gi et par argumenter for hvorfor nettopp denne forskningsmetoden er valgt.

For det første, metoden bygger på interesse og engasjement hos forsker. Denne fremgangsmåten ble jeg først kjent med gjennom boken «Sosiologisk fantasi : essays» (Veiden, 1998) for et par år siden. Denne boken skildrer 11 forskjellige essays hvor, datidens, unge lovende norske sosiologer fikk i oppgave å skrive om noe de var opptatt eller fant som spesielt eller interessant. Sosiologisk Fantasi får frem på en usedvanlig god måten hvordan sosiologi kan diskuteres både som spennende og interessant, og samtidig ha full frihet til kreativ tenking og formidling. Det var særlig siste kapittelet til redaktør Pål Veiden jeg ikke klarte å legge fra meg, og en kunne på åpen gate observere en ung sosiolog i shorts nesten gå inn i andre mens han leste en bok. Jeg opplever at den sammen fremstillingen, dog med klarere rammer, gis av Swedberg.

For det andre, min tolkning av Swedberg går ut på at en bør jobbe med førstudien, og ta med alt som en tenker kan være relevant å ha med for hovedstudien. Denne førstudien er som Swedberg skriver: «en kreativ prosess» (2014, s. 26; min oversettelse). Denne kreative prosessen bør ikke forbli i forskerens hode eller i forskerens notater. Derimot bør den skrives

ned og fortelles om, slik at andre har mulighet til å bygge videre (eller komme med motargumenter) på førstudien/hovedstudien. Oppbyggingen av oppgaven bærer preg av nettopp dette. Det er mye som kunne vært gjort annerledes, og eventuelt fjernet, men da hadde en, argumenterer jeg for, mistet en del av prosessen som ledet til resultatet.

For det tredje, planen var å først gjøre en førstudie for deretter en hovedstudie. Jeg hadde tenkt å kort utdype begrepet jeg har diskutert meg frem til: ikke-tenkende kommunikasjon, for så å gjøre en hovedstudie hvor jeg analyserte begrepet opp mot ChatGPT. Teoretiseringsprosessen viste seg å være svært omfattende. Jeg valgte derfor å isteden fokusere fullt og helt på førstudien1. Hva er så Swedbergs teoretiseringsprosess?

Swedberg sitt rammeverket for teoretisering inbefatter forskningsmetoden abduksjon. Abduksjon er, ifølge Swedberg, «... [en] teori om hvordan å komme opp med forklaringer basert på praktiske perspektiver av en forsker» (Swedberg, 2014, s. 101; min oversettelse). Dette innebærer, som Swedberg poengterer, at en både fokuserer på forklaringen og veien til forklaringen: «... the emphasis is not only on the explanation but also on the process of coming up with an explanation or how to get there» (2014, s. 101). Nettopp denne prosessen er hva som er hovedfokuset for denne oppgaven. Videre argumenterer Swedberg for en todeling av forskning innen sosiologi: en førstudie («prestudy») og en hovedstudie («main study») (2014, s. 28).

Det er en slik førstudie som er forsøkt gjort i denne masteroppgaven. Førstudie defineres av Swedberg som: «... formulering av en tentative teori som vil bli brukt i hovedstudien» (Swedberg, 2014, s. 26; min oversettelse). Hensikten med førstudien er å skape en kreativ prosess hvor forskeren prøver seg frem med ulike ideer om hvordan et tema eller emne kan forskes videre på. Ifølge Swedberg består en førstudie av:

> «Observasjon [av fenomen] ...
>
> Utbrodering av teori (navngi observasjonen; utvikle konsepter, analogier, typologier eller lignende for å fange prosessen, mønstre, o.l. ...
>
> Fullstendigjøre den tentative teorien gjennom en forklaring» (Swedberg, 2014, s. 28; min oversettelse).

Swedberg legger dermed opp til at en kan følge en mal for hvordan å forske i sosiologien. Først må en gjøre seg en observasjon (Swedberg selv definerer det som en «sosial observasjon») (kapittel 2, s. 29-51), deretter navngi denne observasjonen, utvikle en typologi eller lage noen konsepter rundt observasjonen (kapittel 3, s. 52-79). Swedberg peker og på evnen til å lage analogier, metaforer eller se mønstre (kapittel 4, s. 80-97) som fordelsaktig for å sette observasjonen i et nytt lys eller forstå observasjonen på en annen måte (Swedberg, 2014, s. 80-81). Først etter at dette er gjennomført bør en utvikle en generell teori om observasjonen (Kapittel 5, s. 98-123).

Hvordan er denne metoden inkorporert i denne masteroppgaven? Det første Swedberg mener en som forsker bør gjøre er å velge seg noe å forske på, og aller helst noe en selv synes er spennende/interessant eller noe som er et problem (Swedberg, 2014, s. 37). Som beskrevet i bakgrunnen for oppgaven, synes jeg den stadige utviklingen av kunstig intelligens er spennende. Særlig nyvinningen ChatGPT (og forbedringen GPT-4) er banebrytende og har potensielt store samfunnsomveltninger (se for eksempel «Italia forbyr ChatGPT» (Nrk, 2023), «How to use ChatGPT as a learning tool» (Abramson, 2023), «NTNU forbyr bruk av ChatGPT på eksamen: Nødvendig å nevne det eksplisitt» (Hystad og Fanghol, 2023), «Kunstig intelligens utfordrer kritisk tenkning og demokrati» (Albrigtsen, 2023), og så videre). Jeg har selv diskutert mye med ChatGPT, så da må det være en eller annen form for kommunikasjon mellom meg og den kunstige intelligensen?

Fenomenet som førstudien bygger på er dermed kommunikasjon mellom mennesker og kunstig intelligens(er). Jamfør Swedbergiansk teoretisering trenger denne observasjonen et navn. Det som best beskriver det jeg observerer er: ‘ikke-tenkende kommunikasjon’. I begrepet ligger det en underliggende ide om at den kunstige intelligensen jeg kommnuniserer med ikke egentlig ‘tenker’. Det å definere hva det vil si å tenke er ingen lett sak. Det som derimot er klart er at mennesker påvirkes av emosjoner når de tenker mens maskiner (deriblant kunstig intelligens) ikke gjør det (Cuzzolin, Morelli, Cîrstea og Sahakian, 2020, s. 1). En kan dermed postulere at en vital del av det «å tenke» er nettopp avhengig av emosjoner. I videreføringen av dette argumentet er derfor «ikke-tenkende» å forstå som mangeler på emosjoner. Emosjoner er en viktig pådriver til menneskers kognitive egenskaper men kognitive egenskaper og ferdigheter avhengig av emosjoner?

Videre kan en argumentere for at jamfør denne måten å forstå begrepet «å tenke», er kunstig intelligens ikke i stand til å tenke. Derav observasjonen «ikke-tenkende kommunikasjon». En

logisk vei videre er å stille seg spørsmålet: kan det kalles kommunikasjon det som foregår mellom mennesker og kunstige intelligenser? Kanskje det kommer an på hva slags type kunstig intelligens det er? F.eks. burde en skille mellom «store språkmodeller» som ChatGPT (OpenAI, 2023) eller Bard (Google, u.å.) kontra læringsalgoritmer innebygd i blant annet Facebook, Twitter og TikTok? Sagt på en annen måte, må en ha dialog for at kommunikasjon skal oppstå med kunstig intelligens? Og er det forskjell på typer dialog? Disse spørsmålene avdekker en usikkerhet/utydelighet i hva som ligger i begrepet «kunstig intelligens». Selv med definisjon gitt over fra High-Level Expert Group AI. Det er derfor hensiktsmessig å trekke inn forskningsfeltet på kunstig intelligens, som beskriver forskjellene innad i- og mellom typer av kunstig intelligens(er). Dette grunnlaget gis av både Ongsulee (2017) og Sarker (2020), og vil bli forklart nærmere i kapittel 5.

Videre bygger en Swedbergiansk teoretiseringsprosess på, som nevnt ovenfor, å «... utvikle konsepter, analogier, typologier eller lignende for å fange prosessen, mønstre, o.l.» (Swedberg, 2014, s. 28; min oversettelse). Det finnes allerede en overordnet typologi av kunstig intelligens for sosiologisk analyse, gitt av Liu (2021). Denne typologien fungerer som et godt grunnlag for å analysere kunstig intelligens' rolle i samfunnet, hvordan utviklingen har skjedd, og hvordan kunstig intelligens som fenomen påvirker indidivider. Der den er mangelfull, er i presisjonen av selve KI begrepet. Jeg vil derfor både benytte meg av typologiene som Liu (2021) presenterer og samtidig utvikle en ny typologi basert på en sosiologisk tilnærming til Ongsulee (2017) og Sarker (2020) sine definisjoner av fire former for kunstig intelligens, innen maskinlæring. Den sosiologiske tilnærmingen til begrepet kommunikasjon gis av henholdsvis Niklas Luhmann (1995) (og gjennom Baraldi, Corsi og Esposito, 2021), Elena Esposito (2022) og Jürgen Habermas (1984; 1987).

Det er sett på som hensiktsmessig å ta med flere enn en sosiolog i denne teoretiseringsprosessen, siden det skaper et bedre grunnlag for hovedstudien. Sagt på en annen måte, fruktbarheten i førstudien kommer frem nettopp fordi det benyttes flere teorier. På den ene siden trengs det dog en insnevring, da for mange teorier ville skapt kaos. Denne insnevring er gjort med bakgrunn i debatten mellom Habermas og Luhmann (se for eksempel Gorm Harste sin bok: «The Habermas-Luhmann Debate» (2021) for en utførlig introduksjon til debatten). Videre har en protigée av Luhmann, Elena Esposito (2022), lagd et svært godt utgangspunkt for hvordan å tolke Luhmann koblet opp mot kunstig intelligens. Forøvrig, slik som med Liu (2021), kan en argumentere for at Esposito sin bok: «Artificial Communication:

How Algorithms Produce Social Intelligence» (2022) er mangelfull for denne førstudien. Som nevnt trekkes Habermas inn nettopp for å virke som en motpol til Luhmann, og med det Esposito.

3 Teori

I denne masteroppgaven utforsker jeg det fascinerende samspillet mellom kunstig intelligens og kommunikasjon i sosiologisk kontekst. Som beskrevet har kunstig intelligens raskt blitt en dominerende kraft i moderne samfunn. Det at kunstig intelligens har innvirkning på ulike sfærer av menneskelig interaksjon og sosiale strukturer er av økende interesse for forskere. For å analysere og forstå denne komplekse dynamikken vil jeg henvende meg til to sentrale teoretikere fra sosiologien: Jürgen Habermas og Niklas Luhmann. Med tanke på omfanget av teoretikernes samfunnsanalyser er det gjort et strategisk utvalg med hovedfokus på begge tenkernes kommunikasjonsbegrep. Dette for å kunne svare på problemstillingen: *«Kan kunstig intelligens delta i kommunikasjon, jamfør kommunikasjonsbegrepet til Niklas Luhmann og/eller Jürgen Habermas?»*. En kunne, gjennom Luhmanns analytiske verktøy, undersøkt hvordan kunstig intelligens påvirker informasjonsflyt, sosial koordinering og endringer i maktstrukturer. Eller, ved å anvende Habermas' begreper som offentlig sfære, diskursiv etikk og kommunikativ handling, undersøkt hvordan kunstig intelligens påvirker og utfordrer disse demokratiske kommunikasjonsidealene.

Dette skal jeg ikke gjøre. Jeg har heller valgt å se på hvordan å inkorporere kunstig intelligens som forskningsfelt i sosiologien. Dette fordi Habermas og Luhmanns teorier er kraftfulle rammeverk for å utforske de komplekse samspillene mellom kunstig intelligens og kommunikasjon. Ved å integrere deres teoretiske perspektiver har jeg prøvd å analysere hvordan kunstig intelligens kan inngå i kommunikasjon, om det i det hele tatt lar seg gjøre. I dette kapittelet vil jeg derfor dykke dypere ned i Habermas og Luhmanns teorier om kommunikasjon. Senere i oppgaven vil jeg, gjennom bruk av min typologi diskutere kunstig intelligens opp mot både først Habermas sitt kommunikasjonsbegrep og deretter Luhmann sitt kommunikasjonsbegrep. Jeg vil argumentere for at dette er begynnelsen på å prøve å avdekke innsikter som kan bidra til en tydeligere forståelse av kunstig intelligens' rolle i samfunnet. Hva legger så Jürgen Habermas i begrepet kommunikasjon?

3.1 Jürgen Habermas' kommunikasjonsbegrep

I bøkene «The Theory of Communicative Action: Volume 1» (Habermas, 1984) og «The Theory of Communicative Action: Volume 2» (Habermas, 1987) legger Habermas ut om sitt begrep om kommunikativ handling. I korte trekk er dette en teori om det moderne samfunn og hvordan individer gjennom kommunikasjon kan foreta handling. Med Habermas sine egne ord:

> «... In modern societies there is such an expansion of the scope of contingency for interaction loosed from normative contexts that the inner logic of communicative action "becomes practically true" in the deinstitutionalized forms of intercourse of the familial private sphere as well as in public sphere stamped by the mass media» (Habermas, 1987, s. 403).

Som vi ser, er ikke Habermas akkurat kjent for å skrive på en enkel måte. En kan her tolke Habermas dithen at det i moderne samfunn er såpass mange muligheter for at mennesker kan interagere på måter som ikke nødvendigvis er styrt av tradisjonelle normer eller sosiale strukturer. Dette innebærer at reglene som vanligvis veileder kommunikasjon og sosial interaksjon utfordres og endres. Videre fører dette til større fokus på individuell utfoldelse og personlig autonomi. En viktig del beskrevet ovenfor – «indre logikk av kommunikativ handling» (min oversettelse) - refererer til de underforliggende prinsippene som styrer kommunikasjon og sosial interaksjon i enhver gitt sammenheng.

Habermas antyder her at disse prinsippene er mindre relevante eller mindre innflytelsesrike i mange områder av moderne liv. Dette inkluderer private familiære interaksjoner, som for eksempel foreldre overfor barn og vice versa. Det inkluderer også offentlige interaksjoner, som gjerne blir formet av massemedier. Habermas snakker ikke selv om sosiale medier (da dette ikke eksisterte) men det er ikke utenkelig at massemedier her og omfatter sosiale medier. Det er summen av dette som fører Habermas til den innsikt om at «kraften i det bedre argument» er en mulig forklaringsmodell på at integrasjon og sosial orden i samfunnet oppstår (Aakvaag, 2008, s. 173-174). Før jeg går inn på kommunikativ handling, er det vært å få frem at i kapittelet «The Uncoupling of System and Lifeworld» hevder Habermas følgende: «Hvert eneste samfunn må imøtekomme det sentrale problemet koordinering av handling: Hvordan får ego alter til å fortsette interaksjonen i den ønskede retningen?» (Habermas, 1987, s. 179; min oversettelse).

Som sosiolog prøver Habermas å gi et svar på hvordan menneskelig interaksjon, altså handling mellom individer, forekommer. Habermas' uttalelse handler derfor om viktigheten av å opprettholde interaksjon og samarbeid mellom individer i et samfunn. «Ego» handler her om invidet som handler, og «alter» er de andre individene som er involvert i samhandlingen. Videre kan en tolke Habermas dithen at denne sammenhengen legger vekt på behovet for å oppnå enighet og koordinasjon mellom ulike handlinger og perspektiver. Kommunikasjon og samhandling spiller derfor en sentral rolle i å oppnå dette. For å få samhandlingen til å fortsette i den ønskede retningen, er det nødvendig å etablere felles forståelse og koordinere handlinger basert på rasjonell argumentasjon og gjensidig forståelse. Dette er et av de interessante aspektetene å analysere kunstig intelligens opp mot. En kan videre tolke Habermas dithen at essensen i begrepet om kommunikativ handling innebærer en felles prosess hvor deltakerne streber etter å oppnå enighet og koordinering gjennom rasjonell dialog og forståelse. Gjennom slik kommunikativ handling kan samfunnet håndtere det sentrale problemet med koordinering av handlinger og oppnå sine mål på en rettferdig og inkluderende måte. Men hvordan oppnås denne koordineringen av kommunikasjon?

Ifølge Habermas oppnås denne koordineringen av handling gjennom to begreper: «(sam)talehandling» (Habermas, 1984, s. 309; Kalleberg i Habermas, 1999, s. 13) og «gyldighetskrav» (Habermas, 1984, s. 307; Aakvaag, 2008, s. 175). Følgende avsnitt vil derfor ta for seg det undertegnede forstår som de viktigste aspektene ved 'talehandling', for deretter å utrede om gyldighetskrav. Fokusområdet for denne oppgaven er ikke kommunikativ handling som overordnet rammeverk, men heller aspektene og nyansene Habermas gir som grunnlag for kommunikasjon. Det sentrale spørsmålet jeg prøver å besvare er om kunstig intelligens kan delta i kommunikasjon, derfor velger jeg å fokusere på talehandling og gyldighetskrav. Talehandlinger («sam»talehandlinger jamfør Kalleberg (Kalleberg i Habermas, 1999, s. 13) deles i tre deler: «regulative, expressive, and constative» (Habermas, 1984, s. 319).

Følgende oversettelser er gjort av undertegnede: «regulative» forstås som ‘regulerende', «expressive» som 'uttrykkende', og «constative» som 'konstanterende'. De tre delene, regulerende, uttrykkende og konstanterende, må forstås som «... pure cases of speech acts» (Habermas, 1984, s. 309). Dette innebærer at de tre delene er idealiserte former, og en setning kan inneholde mer enn bare en av aspektene. En forutsetning som gjøres i oppgaven, er at mennesker kombinerer disse talehandlingene i kommunikasjon med hverandre. Mennesker

kan dermed ved nødvendighet benytte alle tre (men ikke ved nødvendighet må, dette vil være viktig å ha i mente i analysen).

La oss først se på hva som menes med 'konstanterende' talehandlinger. Ifølge Habermas er konstanterende talehandlinger bygget på det han kaller «elementære proposisjoner» som er «assertorisk» (Habermas, 1984, s. 309; min oversettelse). Eksempler på dette kan være: 1. 'der står det et tre', eller 2. 'det er skyer på himmelen'. Setningene er assertorisk fordi det kunne vært i eksempel 1. ikke noe tre, eller i eksempel 2. ingen skyer. Den andre typen talehandling kaller Habermas for 'uttrykkende'. Med dette peker Habermas på at talehandlinger kan være bygget på «elementære erfaringer», «(... i første persons presens)» (Habermas, 1984, s. 309; min oversettelse). Eksempler på dette kan være: 3. 'jeg ser at det står et tre der', eller 4. 'hvor mange skyer ser du på himmelen?'. Til sist har en ‘regulerende’ talehandling. Med 'regulerende' mener Habermas setninger bygget på det han kaller «elementært avgjørende (som i en kommando) eller elementært bevisste/tilsiktede (som i et løfte)» (Habermas, 1984, s. 309; min oversettelse). Eksempler på dette kan være: 5. 'gå til det treet', eller 6. 'jeg lover at du vil se himmelen bare du går bort dit'. Disse tre elementene ved talehandlinger kan dernest kombineres, for eksempel: 7. 'jeg lover at du vil se himmelen bare du kjører en times tid nordover'. I eksempel 7 kombineres regulerende- og konstanterende talehandlinger. Med andre ord innebærer dette at all kommunikasjon har en iboende logikk uavhengig av språket som snakkes. En forutsetning Habermas gjør her, er at det må være snakk om en eller annen form for språk.

Videre påstår Habermas at talehandlinger har tre gyldighetskrav: 1. «...aspektet av riktigheten som taleren påstår for sin handling i forhold til en normativ kontekst (eller, indirekte, for disse normene selv)...», 2. «sannferdigheten som taleren påstår for uttrykket av subjektive opplevelser som han har privilegert tilgang til» 3. «... sannheten som taleren, med sin ytring, påstår for en påstand (eller for de eksistensielle forutsetningene til en nominalisert proposisjon)» (Habermas, 1984, s. 307; min oversettelse). Alle mennesker i et samfunn (om det er seg det norske eller det kongolesiske) må forholde seg til en normativ kontekst. Denne normative konteksten kan forstås på forskjellige måter, for eksmpel med begreper og konsepter. Det første gyldighetskravet handler altså om at den som uttaler og med det påstår noe, må godtas av de som individet snakker til. For eksempel kan en lærer påstå at en elevs besvarelse ikke oppfyller kravene til oppgaven som eleven har prøvd besvart, og dermed ikke er verdig en god karakter. Læreren vil da forsøke å overbevise eleven om at karakteren er

rettferdig og basert på klare normer for vurdering. Nøkkelordet her er rettferdig. Videre vil læreren prøve å skape tillit og aksept hos eleven ved å vise at beslutningen hen har gjort er rettferdig og i tråd med aksepterte standarder.

Det andre gyldighetskravet til Habermas går på at alle mennesker har subjektive erfaringer som ingen andre har tilgang på. Selv eneggede tvillinger som gjør alt likt, ser på det samme og opplever det samme, vil ha subjektive erfaringer med privilegert tilgang siden de til envher tid er seg selv og ikke den andre. Denne subjektive erfaringen blir ofte fremmet i observasjon av noe (eller noen). Nøkkelordet her er sannferdighet. For eksempel kan en lærer påstå at hen har lagt merke til at en elev har hatt vanskeligheter med et bestemt fag eller emne. Dette på tross av at eleven ikke selv har gitt uttryk for det. Læreren vil da prøve å formidle til eleven at hens observasjoner er sannferdige og skal tas på alvor. Ved å gjøre dette vil læreren forsøke å bygge tillit med eleven og skape en åpen kommunikasjonskanal mellom dem.

Det tredje gyldighetskravet til Habermas går på at en uttalelse eller påstand skal være basert på fakta og at det skal være mulig å gi gode grunner for selve uttalelsen. Her trengs det en oppklaring. Fakta er ikke å forstå som i «hard facts», som en ofte taler om i naturfagen, men heller som en delt forståelse av rett og riktig mellom individer/grupper. For eksempel, om tidligere president Donald Trump står foran sine tilhengere deler de et sannhetsbilde som for en utenforstående virker usant (eller rett og slett løgn). Det finnes og uttalelser som går inn under kategorien «nominalized proposition», som en kan forstå som de grunnleggende forutsetningene som ligger til grunn for en uttalelse eller påstand. For eksempel kan påstanden «frihet er en menneskerettighet» ha en eksistensiell forutsetning om at det finnes en slik ting som menneskerettigheter og at disse rettighetene er universelle og ufravikelige. Disse forutsetningene må være oppfylt, hvis ikke vil påstanden være ugyldig.

I dette delkapittelet har jeg beskrevet Habermas teori om 'talehandling' og 'gyldighetskrav', med spesielt fokus på deres anvendelighet i konteksten av kunstig intelligens. I diskusjonen om 'talehandling', fremhevet jeg 'konstanterende', 'uttrykkende' og 'regulerende' aspekter, der hver av disse representerer forskjellige idealiserte former for kommunikasjon som kan kombineres i en gitt setning. Habermas sine tre gyldighetskrav, normativ aksept, subjektive erfaringers sannferdighet og påstanders grunnlag i fakta, ble deretter analysert. Disse kravene gir en nyansert tilnærming til kommunikasjon og legger grunnlag for analysen av kunstig intelligens opp mot Habermas kommunikasjonsbegrep. Dette er det første rammeverket som vil bli diskutert opp mot kunstig intelligens typologien jeg utreder om i kapittel 5. Oppsumert

kan en påstå at Habermas fokuserer på intersubjektiv kommunikasjon mellom individer, samt vektlegger betydningen av felles forståelse, rasjonell diskurs og konsensusbygging. Hans teori om talehandlinger og gyldighetskrav antyder at en vellykket kommunikasjon er basert på gjensidig forståelse, ærlighet og riktig bruk av språket.

3.2 Niklas Luhmanns kommunikasjonsbegrep

Den andre sosiologen jeg har valgt å se på er Niklas Luhmann. Som med Habermas, skriver Luhmann på en avansert og vanskelig måte. Der hvor det har vært lett å fange essensen til Habermas har det vært desto vanskeligere å fange opp hva Luhmann spesifikt tenker om kommunikasjon. Denne grunnen, og at Esposito ellers er en sentral figur i oppgaven, gjør at jeg ser det som mest fruktbart å bygge videre på den analysen Baraldi, Corsi og Esposito legger frem i boken «Unlocking Luhmann: A Keyword Introduction to Systems Theory» (Baraldi, Corsi og Esposito, 2021). Det er styrker og svakheter ved å bygge videre på sekundær analyse av et rammeverk. Den anvendelighet og tydeliggjøring som Baraldi, Corsi og Esposito gjør påstår jeg klargjør denne oppgaven mer enn de ulempene det medbringer.

Før vi går inn på Baraldi, Corsi og Esposito sin analyse av Luhmann (2021), kan en på generell basis forstå Luhmann dithen at kommunikasjon er den grunnleggende prosessen som strukturerer alle sosiale systemer (Luhmann, 1995). Om kunstig intelligens kan delta i kommunikasjonen vil derfor være avgjørende for å forstå og tolke Luhmanns samfunnsteori. Før vi ser på dette må det presiseres at det er det sosiale system som er fokusområde i denne studien, men at i Luhmanns teori eksisterer og «det psykiske system» og flere deler som tilsammen utgjør det fysiske system (Aakvaag, 2008, s. 237).

Tilbake til kunstig intelligens. Som beskrevet tidligere representerer kunstig intelligens et paradigmeskifte i måten informasjon behandles og formidles på. Intuitivt virker det som om noen typer kunstig intelligens-systemer har evnen til å samle inn, prosessere og generere informasjon på måter som tidligere var forbeholdt menneskelig interaksjon. Denne evnen til å engasjere seg i kommunikasjon og håndtere informasjon, kan potensielt påvirke strukturen og funksjonen til sosiale systemer på flere måter. Ved å bygge videre på Esposito (2022) kunne jeg ha anvendt Luhmanns kommunikasjonsteori for å analysere hvordan kunstig intelligens forandrer de etablerte kommunikasjonsmønstrene og systemstrukturene i samfunnet. Jeg kunne og ha undersøkt hvordan kunstig intelligens påvirker informasjonsflyt, dannelsen av kommunikative prosesser og endringer i for eksempel maktrelasjoner. Videre kunne jeg ha

utforsket hvordan kunstig intelligens endrer forholdet mellom mennesker og maskiner, og hvordan det påvirker samspillet og dynamikken i ulike sosiale systemer. Det var dette jeg først hadde tenkt til å gjøre. Men, for at dette skal kunne gjøres må det ligge et fundament som muliggjør en slik tilnærming. Jeg vil i denne førstudien, argumenterer for at grunnlaget for å forstå hva kunstig intelligens som fenomen i sosiologiske termer, ikke er godt nok diskutert. Følgende strategiske utvalg av Luhmanns kommunikasjonsteori vil derfor bli brukt senere i oppgaven, koblet opp mot den samme typologien som blir brukt for å analysere Habermas. Hvordan tolker så Baraldi, Corsi og Esposito Luhmann sitt kommunikasjonsbegrep?

På et overordnet plan er Luhmanns teori om kommunikasjon forankret i ideen om at sosiale systemer er autopoietiske (Baraldi, Corsi og Esposito, 2021, s. 37 og 41), noe som betyr at de opererer på grunnlag av sine egne interne koder og skiller. Kommunikasjon er prosessen der disse kodene og skillene genereres og vedlikeholdes (Baraldi, Corsi og Esposito, 2021, s. 45-48) noe som gjør at systemet fungerer som en sammenhengende helhet. Som Baraldi, Corsi og Esposito peker på, understreker Luhmann viktigheten av kommunikasjon som et middel for differensiering (Baraldi, Corsi og Esposito, 2021, s. 41 og 61-63), der forskjellige elementer innenfor systemet skiller seg fra hverandre gjennom bruk av spesifikke koder og symboler. Ifølge Luhmann er kommunikasjon ikke bare et verktøy for å formidle informasjon, men en måte å skape og vedlikeholde sosial virkelighet på (Baraldi, Corsi og Esposito, 2021, s, 222). Slike sosiale systemer er komplekse nettverk av kommunikasjon som opererer etter sin egen interne logikk (Baraldi, Corsi og Esposito, 2021, s. 37). Grensene til disse systemene er definert av kodene og skillene de bruker, og kommunikasjon er det primære middel for å vedlikeholde disse grensene.

Generelt sett er Luhmanns konsept om kommunikasjon sentralt i hans bredere teori om samfunnet, som ser sosiale systemer som autopoietiske (Baraldi, Corsi og Esposito, 2021, s. 37-40) og autonome enheter som opererer etter sin egen interne logikk (Baraldi, Corsi og Esposito, 2021, s. 40 og 51). Kommunikasjon blir dermed ikke bare en måte å overføre informasjon på, men en måte å skape og vedlikeholde mening innenfor disse systemene, og er avgjørende for å forstå naturen av sosial virkelighet. En kan dermed tolke Luhmann dithen at det foregår komplekse prosesseser som innebærer skapelse og vedlikehold av mening innenfor et system.

Som Baraldi, Corsi og Esposito skriver skriver om Luhmanns forståelse av kommunikasjon:

> «Communication is the basic element and operation of social systems. It consists of the unity of the difference among three selections: utterance (Mitteilung), information, and understanding (Verstehen) of the difference between utterance and information» (Baraldi, Corsi, Esposito, 2021, s. 45).

Det er altså kombinasjonen av disse tre bestandelene: Meddelelse og informasjon, og forståelsen av forskjellen mellom disse to. Den norske litteraturen bruker meddelelse, noe jeg støtter meg på (se for eksempel Aaakvaag, 2008, s. 233). Dette innebærer at kommunikasjon handler om mer enn bare å oppfatte informasjon. Videre argumenterer Baraldi, Corsi og Esposito for at kommunikasjonsbegrepet til Luhmann inneholder to nøkkelkomponenter, informasjon og ansvaret til deltakeren som produserer en meddelelse (Baraldi, Corsi, Esposito, 2021, s. 45). Ansvaret for å uttale informasjonen er et tydelig valg som skiller seg fra informasjonen i seg selv, og denne distinksjonen er nødvendig for at kommunikasjon skal kunne skje.

Et viktig element er at informasjon er et valg fordi å kommunisere om ett emne ekskluderer andre potensielle emner. Kommunikasjon innebærer produksjon og forståelse av informasjon av minst to deltakere, og det kan ikke reduseres til bare persepsjon av sensoriske innputt. Videre legger Baraldi, Corsi og Esposito vekt på at i kommunikasjonsbegrepet til Luhmann blir informasjon ikke bare overført fra en person til en annen, men aktivt produsert og forstått av deltakerne (Baraldi, Corsi og Esposito, 2021, s. 45). Dette er et skift med den klassiske Shannon-Weaver-overføringsmodellen (jamfør fremstillingen til Al-Fedaghi (2012)) som ofte brukes om kommunikasjon. For Luhmann handler kommunikasjon derfor om at en uttalelse av informasjon er et valg fordi den viser talerens intensjoner, motivasjoner, årsaker og kunnskap, samt ansvar for å snakke. Som Baraldi, Corsi og Esposito påpeker:

> «Through understanding, communication can stress who has uttered what. Therefore, understanding makes it possible for further communication to refer to either previous utterances (someone's motives or intentions) or uttered information (what), thus generating communication →processes. Understanding, rather than for thinking, is important for the reproduction of communication, although thinking is related to communication [→Interpenetration and Structural Coupling]» (Baraldi, Corsi og Esposito, 2021, s. 46).

Dette innebærer at uttalelsen av informasjon ikke er nok for at kommunikasjon kan oppstå. Kommunikasjon krever forståelse, som er valget som skiller mellom informasjonen som blir formidlet og årsakene til å formidle den. Forståelse er viktig fordi det tillater kommunikasjon å referere til tidligere uttalelser, motiver og intensjoner, og dermed skape kommunikasjonsprosesser (Baraldi, Corsi og Esposito, 2021, s. 45-46). Baraldi, Corsi og Esposito understreker at forståelse i Luhmanns teori ikke handler om objektiv informasjon eller autentiske grunner for å snakke, men om attribusjonen av valg som informasjon og årsaker til uttalelse. Derfor kan ethvert nivå av forståelse realisere kommunikasjon, selv om informasjonen eller årsakene til uttalelsen er misforstått eller villedende. Reproduksjonen av kommunikasjon avhenger og av forståelse, ikke bare tenkning (Baraldi, Corsi og Esposito, 2021, s. 46). Forøvrig, en presisering Baraldi, Corsi og Esposito gjør, er at det er kun sosiale systemer som kan kommunisere (Baraldi, Corsi og Esposito, 2021, s. 38). Et sosialt system er følgelig et autopoietisk, selvreferensielt system som opprettholder seg selv, og som er differensiert fra sin omgivelse. Det er et system som skaper mening, samt dets handlinger og endelige elementer er kommunikasjon (Baraldi, Corsi og Esposito, 2021, s. 221).

Et annet viktig poeng Baraldi, Corsi og Esposito tar opp er at «Utterance, information and understanding can be separated for analytical purposes, but they are a unity in communication, which cannot be decomposed» (2021, s. 46). Det å kunne skille mellom meddelelse, informasjon og forståelse er viktig å ha med, særlig i et studie av kunstig intelligens (dette vil bli utdypet mer senere i oppgaven). Videre er det viktig å poengtere nettopp at kommunikasjon er en såkalt «event». Dette innebærer at de tre bestanddelene er «realisert samtidig» (Baraldi, Corsi og Esposito, 2021, s. 46; min oversettelse). Det at bestanddelene realiseres samtidig gjør ikke at de ikke kan diskuteres hver for seg, men derimot at når en diskuterer de må de alltid sees i lys av nettopp det at de forekommer samtidig. Henholdsvis påpeker Baraldi, Corsi og Esposito at hver kommunikasjonshendelse er ny og unik siden den forsvinner umiddelbart. Imidlertid går kommunikasjonsprosesser utover spesifikke hendelser og krever en forbindelse mellom hver kommunikasjon gjennom forståelse. Dette skaper et rekursivt nettverk av kommunikasjoner som er det som definerer enheten til et sosialt system (Baraldi, Corsi og Esposito, 2021, s. 46).

For å komme tilbake til et siste viktig poeng i Luhmanns systemteori er «mening» (Baraldi, Corsi, Esposito, 2021, s. 137-139; min oversettelse). Som Baraldi, Corsi og Esposito skriver: «The system operations of communication and thought are realized in the medium of

meaning» (2021, s. 138). Dette innebærer at kommunikasjon og tenkning er sentrale systemoperasjoner i sosiale systemer. Disse operasjonene foregår ikke direkte gjennom materielle eller fysiske elementer, men heller gjennom et abstrakt medium som kalles «mening». Meningen fungerer som det mediumet gjennom hvilket kommunikasjon og tenkning oppstår og utvikler seg. Henholdsvis, kan mening betraktes som et symbolsk system som tillater utveksling og forståelse av informasjon mellom mennesker. Det er gjennom mening at kommunikasjon finner sted, hvor deltakere i et sosialt system tolker og gir mening til symboler, tegn, språk og andre kommunikative uttrykk. På samme måte er tenkningens operasjoner, som persepsjon, resonnering og konseptuell dannelse, også forankret i meningens medium. Tenkning involverer bearbeidelse av mening og konstruksjon av mentale modeller, og dette skjer innenfor systemets autopoietiske rammeverk. Dette er det andre rammeverket som vil bli diskutert opp mot kunstig intelligens typologien jeg utreder om i kapittel 5.

I disse to delkapitlene har vi sett på kommunikasjonsbegrepet til først Jürgen Habermas og så Niklas Luhmann. Jeg har valgt å ta for meg disse to tenkerne siden de på hver sin side har svært spennende og innholdsrike begreper om kommunikasjon, og at de har hver sin særegne måte å diskutere nettopp fenomenet kommunikasjon. Jeg vil i påfølgende kapittel se på hva Elena Esposito skriver i sin bok «Aritificial Communication: How Algorithms Produce Social Intelligence» (2022). Jeg har valgt å delegere et eget kapittel til Esposito, nettopp fordi jeg legger opp til en kombinasjon av å presentere teorien til Esposito og samtidig diskutere det jeg mener er mangelfult i sosiologien med tanke på studiet av kunstig intelligens.

4 Espositos forklaringsmodell av KI i sosiologi

Det vil i kommende avsnitt argumenteres for hvorfor sosiologien er mangelfull med tanke på studiet av kunstig intelligens. Jeg vil fokusere mest på Esposito (2022) sin fremstilling. Det må dog understrekes at Esposito sitt bidrag er uvurderlig, dette av to grunner. For det første, fordi Esposito legger frem en analyse av kunstig intelligens som er nyansert. Det er altfor lett å trekke frem de utfordrende sidene ved kunstig intelligens (ofte basert i sci-fi dystopier). Samfunnsanalyser er verdiladet, om en vil det eller ikke. Argumentasjon i seg selv er en måte å overtale en leser til at de aspektene som legges frem er de en burde fokusere på. Det er dermed ikke en dårlig analyse Esposito legger frem. Det som derimot er mangelfult er utgangspunktet, hvor kunstig intelligens ene og alene forstås som algoritmer. Jeg vil komme

tilbake til dette. En kan allikevel komme med svært verdifull innsikt, selv om utgangspunktet er mangelfult/feilaktig. Dette fører oss til den andre grunnen, det faktum at Esposito får frem flere gode poenger. Blant disse er Captcha (s. 108), retten til å bli glemt (s. 65-77), bias i data (s. 109), tvetydighet som mangelfull hos kunstig intelligens (s. 109-110), med fler.

Disse overnevnte poengene er såpass gode, og kan være relevante som motargument for den analysen jeg selv legger opp senere i oppgaven. Jeg vil derfor ta for meg hver av de overnevnte poengene. Før jeg gjør det, må det en begrepsavklaring til. Esposito sidestiller kunstig intelligens med algoritmer. Som vil komme frem i analysen, er jeg uenig i denne sidestillingen. Henholdsvis, når jeg bruker begrepet algoritme sikter jeg til Esposito sin tolkning, mens når jeg benytter begrepet kunstig intelligens sikter jeg til den definisjonen gitt av High-Level Expert Group AI som nevnt i innledningen (2018).

Det første poenget til Esposito jeg vil ta opp, er det jeg har valgt å kalle «den omvendte Turing-testen», det vil si: Captcha. Ifølge IT-selskapet Google står CAPTCHA for: «Completely Automated Public Turing test to tell Computers and Humans Apart» (Google, u.å.). Turing-testen har vært et viktig bidrag inn på forskningsfeltet på maskiner og datateknologi og går som følger: se for deg at du sitter foran en datamaskin med et kommunikasjonsprogram. På den andre siden av kommunikasjonsprogrammet er det enten en datamaskin som styrer, eller et menneske som styrer. Turing påsto at når mennesker ikke lenger forstår forskjellen på om det er et menneske eller en datamaskin, er det en «intelligent datamaskin» (Copeland, 2000, s. 519-522; min oversettelse). I skrivende stund, og som Esposito påpeker, er det ikke lenger maskiner som prøver å overtale mennesker om at de er mennesker. Derimot er det maskiner som prøver å overtale andre maskiner om at de er mennesker. Med denne problemstillingen oppsto det som Esposito (2022, s. 108) peker på: «Captcha». Captcha kan dermed sies å være en algoritme som er laget for å skille mennesker fra maskiner, gjennom en rekke øvelser som maskiner (herunder og diverse kunstig intelligenser) ennå ikke får til.

Det andre gode poenget til Esposito går på bias i data (2022, s. 66). De brillene vi som individer bruker har mye å si for hvordan vi forstår verden. Algoritmer er intet mindre. Den innputten som gis påvirker hvilke slutninger algoritmen gjør. Det er derfor en innebygd bias i envher algoritme basert på hvilken data som den får tilgang på. Videre påpeker Esposito at en ikke kan med sikkerhet fastslå at algoritmeutviklerne faktisk er opphavet til de implementerte prosedyrene. Problemet, slik Esposito ser det, ligger ikke så mye i at maskinene gjenspeiler

skaperne sine skjevheter, men heller det motsatte: Deres skjevhet stammer hovedsakelig fra at deres funksjonalitet ikke samspiller med skaperne sine verdier (Esposito, 2022, s. 109)

Esposito peker på her det at innputtdataen ikke alltid inneholder ønsket informasjon som en vil at algoritmen skal behandle, samt at prosedyrene for implementering bygger på bias ukjent fra skaperne. Et kjent eksempel er Microsofts Tay.ai (Vincent, 2016). Tay var et forsøk fra Microsoft på å lage en konversasjonsrobot på Twitter. Som Vincent skriver «Det tok mindre enn 24 timer for at Twitterfolket korrumpterte en uskylidg KI chatbot» (Vincent, 2016; min oversettelse). I løpet av disse 24 timene hadde altså twitterskaren fått Tay til å bli rasistisk, kvinnehatende og ny-nazistisk. En kan tolke Esposito (2022) dithen at selv med gode intensjoner og sikkerhetsrutiner er kommunikasjon såpass komplekst at det som skjedde med Tay godt kan skje igjen:

> «Machines participate in a communication that is neither neutral nor egalitarian, and they learn to work correspondingly, in ways that can be biased very differently from the preferences of their designers» (Esposito, 2022, s. 109).

To viktige påpekninger som Esposito får frem her er at kommunikasjon nok hverken er nøytral eller egalitær. Uavhengig av modellen som ligger til grunn for å forstå kommunikasjon mellom mennesker, er kommunikasjon i seg selv sjeldent (om aldri) nøytral. Dette fordi kommunikasjon er en såpass kompleks prosess, som involverer både den som uttrykker en melding og den som mottar meldingen. Videre kan det sies at meldingen kan påvirkes av en rekke faktorer som kulturelle forskjeller, personlige oppfatninger, maktforhold, med fler. Argumentet til Esposito bygger på at siden algoritmen interagerer med så og si hele verden, vil dette kun forsterke problematikken rundt nøytralitet. I tillegg er det stor sannsynlighet for, slik som med Tay.ai, at algoritmen ikke lenger følger preferansene til de som skapte algoritmen.

Det tredje poenget til Esposito handler om mangelen/utfordringen til algoritmer om å forstå tvetydighet:

> «For algorithms, however, ambiguity is notoriously a challenge. Machines not only struggle with understanding the ambiguity of human communication, they struggle harder to generate ambiguous communication—that is, to use in competent ways the ambiguity required by legal arguments» (Esposito, 2022, s. 110).

Dette innebærer at maskiner har vanskeligheter med å forstå og generere tvetydig kommunikasjon, i motsetning til mennesker. Spesielt bruk av tvetydighet som kreves av juridiske argumenter, utgjør en utfordring for maskiner ifølge Esposito. Debatten mellom forklaring og tolkning i loven gjenspeiler denne vanskeligheten. For å motagere denne utfordringen peker Esposito på feltet forklarbar kunstig intelligens (se Turri (2022) for en utfyldende diskusjon av forklarbar KI). Dette nettopp for å gjøre beslutningene som algoritmen gjør mer gjennomsiktige. I korte trekk innebærer dette å gi detaljerte prosedyretrinn som illustrerer hvordan maskinen kom frem til sin beslutning. Essensen er å gjøre beslutningsprosessen mer klar og gjennomsiktig, selv om maskinen sliter med å forstå eller generere tvetydighet på samme måte som mennesker kan.

Dette fører oss til det mest sentrale begrepet til Esposito, «artificial communication» (2022). Som jeg har vært innom tidligere, argumenterer Esposito for er en grunnleggende endring av hvordan å forstå algoritmer. Esposito (2022, s. 2-3) argumenterer for at moderne maskinlæringsalgoritmer ikke nødvendigvis lærer å forstå informasjon på samme måte som mennesker gjør, men i stedet blir mer effektive ved å fokusere på å utnytte de store mengdene data som genereres av brukere på nettet. Videre antyder Esposito at disse algoritmene ikke gjenskaper menneskelig intelligens, men heller reproduksjon av menneskelige kommunikasjonsevner. Bruken av store datamengder tillater disse algoritmene å lage prognoser og generere resultater basert på mønstre som de finner i dataene, uten nødvendigvis å forstå betydningen bak mønstrene (2022, s. 2-3). I praksis kan en forstå Esposito dithen at disse algoritmene ikke er ekte intelligente på samme måte som mennesker, men heller bruker statistiske metoder for å lære av store datamengder.

Denne forenklingen av kunstig intelligens som kun algoritmer er jeg uenig i. Et eksempel på hvorfor dette ikke stemmer er å finne når Stockfish (den ledende sjakk-computeren) spilte mot KI-modellen AlphaZero. Som Silver et al. skriver:

> «AlphaZero searches just 60,000 positions per second in chess ..., compared with 60 million for Stockfish and ... . AlphaZero may compensate for the lower number of evaluations by using its deep neural network to focus much more selectively on the most promising variations ... —arguably a more humanlike approach to searching, as originally proposed by Shannon» (Silver et al., 2018).

Det som her kommer frem, og som Silver et al. (2018) peker på, er at AlphaZero som gjøre færre utregninger per sekund er allikevel en klart bedre sjakkspiller enn Stockfish. Dette bygger de på argumentet til Shannon, som allerede i 1950 postulerte følgende:

> «The number of possible positions, of the general order of 64! / 32!(8!)^2 (2!)^6 , or roughly 1043, naturally makes such a design unfeasible. It is clear then that the problem is not that of designing a machine to play perfect chess (which is quite impractical) nor one which merely plays legal chess (which is trivial). We would like to play a skilful game, perhaps comparable to that of a good human player» (Shannon, 1950, s. 4).

Shannon gjør her en grov overslagsberegning for å se på antall lovlige muligheter i et sjakkspill. Siden dette er en god pekepinn på hvorfor ren algoritmisk utregning ikke er praktisk vil jeg her i korte trekk legge ut om beregningen. I matematikk forstås «!» som fakultet av et tall. For eksempel er 4! = 4*3*2*1. For det første, 64! representerer alle mulige måter å arrangere 64 unike elementer (som i dette tilfellet er de 64 feltene på et sjakkbrett). For det andre, 32! er antall måter å arrangere 32 unike elementer (som kan representere de 32 brikkene i et komplett sjakksett). For det tredje, (8!)^2 representerer antall måter å arrangere 8 unike elementer (som kan representere bønder av hver farge) kvadrert, siden det er 8 bønder for hver av de to fargene i sjakk. For det fjerde, (2!)^6 representerer 2! antall måter å arrangere to elementer på, som kan representere de to fargene (hvit og svart) for hver type brikke (tårn, løpere, hester, dronninger). Henholdvis er det 6 typer brikker som kommer i par (2 tårn, 2 løpere, 2 hester, og 2 dronninger for hver farge), derfor hever vi 2! til sjette for å representere alle mulige arrangerte kombinasjoner av disse brikkene.

Denne noe matematisk tekniske fremstillinge viser at uttrykket til Shannon beregner antall mulige unike sjakkstillinger, gitt en viss antagelse om at alle posisjoner er like sannsynlige og ignorere regler som begrenser lovlige trekk (for eksempel sjakkregler om plassering av kongen, osv.). Som nevnt bør det bemerkes at dette er en grov overslagsberegning, og det faktiske antallet lovlige sjakkstillinger er noe lavere på grunn av de nevnte regler og begrensninger. Et mer overkommelig tall å forholde seg til, og som kalles «Shannons tall» jamfør det overnevnte ressonementet, er 10^120 (som beskriver antall mulige trekk i et gjennomsnittlig sjakkparti) (Chess Journal, u.å.). Som det kommer frem hos Chess Journal er dette et konservativt estimat. Det estimatet får frem, er at det er praktisk uegnet å kun bygge

KI-modeller basert på store data, de modellen som bygger på andre prinsipper (som AlphaZero gjør) er bedre egnet en de rent algoritmisk baserte.

Omformuleringen av kunstig intelligens til «artificial communication» brukes for å argumenterer for flere punkter. I kapittel 2 utbroderer Esposito (2022) om hvordan en kan forstå algoritmer og algoritmisk «tenking» (eller mangel på nettopp tenking). Dette bygger på ideen om at algoritmer «tenker» gjennom bruken av lister (Esposito, 2022, s. 19). Disse listene er det som sørger for nettopp styrken til algoritmer. Esposito argumenterer for at det å gi opp på ideen om reproduksjon av menneskelig intelligens, er det som sørger for at utviklingen innen kunstig intelligens har gått såpass fort. Med Espositos egne ord: «Algorithms do not reason the way we do in order to do what we do with abstract reasoning» (Esposito, 2022, s. 27). Som en kan se, er et av hovedpoengene til Esposito at det i forskningen på kunstig intelligens har en gått bort ifra tankegangen om å prøve å kopiere menneskelig intelligens. Med andre ord, peker Esposito på at styrken og effektiviteten til algoritmer avhenger av deres evne til å utføre beregninger uten behov for abstraksjon.

Videre antyder Esposito (2022) at de nylige fremskrittene innen selv-lærende algoritmer er muliggjort ved å gi opp ideen eller ambisjonen om å gjenskape menneskelig intelligens gjennom programmeringsteknikker:

> «Modern machine-learning algorithms are so efficient not because they have learned to imitate human intelligence and to understand information, but rather because they have abandoned the attempt and the ambition to do so and are oriented toward a different model. Machine-learning algorithms that use big data, I claim, are artificially reproducing not intelligence but communication skills, and they do so by parasitically exploiting the participation of users on the web.» (Esposito, 2022, s. 2-3)

Riktig påpekt, resonerer ikke algoritmer på samme måte som mennesker gjør og er heller ikke avhengige av abstrakt resonnering for å utføre oppgaver (foreløpig). Esposito sammenligner dette med gamle kulturer uten alfabet, der lister ofte ble brukt som en måte å formidle informasjon på uten behov for abstraksjon (Esposito, 2022, s. 19-26). Essensen i argument går på at algoritmer er i stand til å fungere effektivt nettopp fordi de ikke er avhengige av samme former for resonnering som mennesker. Til tross for dette er algoritmenene allikevel i stand til å oppnå imponerende resultater gjennom ikke-menneskelige metoder for å prosessere og

analysere data (Esposito, 2022, s. 26-29). Derav skiftet fra kunstig intelligens, til «artificial communication».

Oppsumert argumenterer jeg her for at det er klare mangler i sosiologien på studiet og fenomenet kunstig intelligens, med spesiell oppmerksomhet mot Espositos (2022) analyse. Jeg verdsetter Espositos bidrag for dets nyanserte syn på kuntig intelligens, men jeg kritiserer det for å definere kunstig intelligens for enkelt, kun som algoritmer. Særlig tre sentrale punkter i Espositos analyse er av viktighet. For det første, captcha som en «omvendt Turing-test», der maskiner ikke klarer å overbevise andre maskiner om at de er mennesker. For det andre, bias i data, som oppstår fra innputtdata og innebygde prosedyrer uavhengig av skaperne. Dette følger av Espositos argument om at den kunstig intelligens modellens skjevhet kommer fra det faktum at dens funksjonalitet ikke samstemmer med skapernes verdier. For de tredje, utfordringen med tvetydighet for algoritmer, hvor algoritmer i skrivende stund sliter med å forstå og generere tvetydig kommunikasjon.

Hovedpoenget til Esposito er dermed at moderne maskinlæringsalgoritmer ikke etterligner menneskelig intelligens, men heller fokuserer på å utnytte store mengder data for å produsere resultater basert på mønstergjenkjenning. Denne tilnærmingen kaller Esposito for «artificial communication» og grunnlegges i hennes syn om at KI-utviklingen har skutt fart nettopp fordi en har gitt opp å prøve å kopiere menneskelig intelligens. Stemmer denne antagelsen? Det er argumenter både for og imot, men for å komme dit må vi først innom kunstig intelligens og typologien min, som jeg nevnte såvidt tidligere i kapittelet.

5 Kunstig intelligens: Inspill til ny typologi

Den innsikt Esposito (2022) her har gitt er definitivt uvurderlig, men har og en klar mangel: Kunstig intelligens er ikke bare én type algoritme. Derimot er kunstig intelligens et omfattende felt som trengs å utredes om for at en ikke skal gå feil i analysen (og konklusjonen) om KI som fenomen. Som en konsekvens av denne innsikt legger jeg her opp til følgende tredeling: 1. legge ut om og presisere hva kunstig intelligens er, 2. diskutere diverse kunstig intelligenser opp mot sosiologisk teori innen kommunikasjon, 3. bygge opp om følgende argumentasjon: for å forstå samtiden trenger sosiologien å sette seg inn i potensielt samfunnsomveltende teknologi og tenking.

Før en kan stille seg spørsmålet: Foregår det kommunikasjon mellom kunstig(e) intelligens(er) og mennesker? Trengs det en presisering til: Hva er egentlig grunnpilarene i forskningen på maskinlæring? Hvis en ser på forskningsfeltet innen kunstig intelligens er det særlig fire typer maskinlæringsteknikker som legges frem (Sarker, 2021; Ongsulee, 2017). I kommende avsnitt vil disse typene bli utførlig diskutert. I tråd med Swedbergiansk teoretisering vil jeg derfor i dette kapittelet redegjøre for en måte å forstå de forskjellige kunstige intelligensene i sosiologisk kontekst. For å forstå *om* kunstig intelligens kan inngå i kommunikasjon, må det først klargjøres *hva* kunstig intelligens faktisk er. For å forstå hva kunstig intelligens er, må vi innom kunstig intelligens slik det forstås i sosiologien i dag.

I innledningen presenterte jeg flere sosiologiske tilnærminger til kunstig intelligens, deriblant Zheng Liu (2021) sin artikkel «Sociological perspectives on artificial intelligence: A typological reading». I artikkelen argumenterer Liu for en to-deling av begrepet kunstig intelligens: «... strong (or general) and weak (narrow) AI» (Liu, 2021, s. 3). I korte trekke skiller disse to seg fra hverandre i det at generell kunstig intelligens kan gjennomføre hvilken som helst menneskelige egenskap av intellektuel eller kognitiv oppgave/handling. «Svak/smal» kunstig intelligens kan derimot kun gjennomføre oppgaver/handling på en «menneskelignende måte» (Liu, 2021, s. 3; min oversettelse)). Siden førstnevnte ikke er oppnådd vil svak/smal kunstig intelligens kun omtales som kunstig intelligens (jamfør definisjonen til High-Level Expert Group AI) i resten av denne masteroppgaven. Henholdsvis, denne masteroppgaven er et dypdykk inn i «techinical AI» (teknisk KI) studiet som Liu legger frem (Liu, 2021, s. 6-8). I typologien teknisk kunstig intelligens beskriver Liu hvordan sosiologer har diskutert algoritme perspektivet i sosiologien (Liu, 2021, s. 7). Denne typologien vil jeg utdype om, samt prøve å forbedre.

Videre peker Liu på at det er særlig to typer «teknikker» som har revolusjonert forskningen og utviklingen av kunstig intelligen: «maskinlæring (ML) og dyp-læring (DL)» (Liu, 2021, s. 3; min oversettelse). Ja, det stemmer at både maskinlæring og dyp læring har vært med på å revolusjonere feltet kunstig intelligens. Derimot, som vil bli utførlig utredet, er dette en mangelfull forklaring på forskjellen mellom maskinlæring og dyplæring. Det jeg her argumenterer for, er større forståelse og innsikt i hva maskinlæring faktisk er. Jeg trekker derfor inn to artikler fra maskinlæringsfeltet: Ongsulee (2017) og Sarker (2021).

Ifølge både Ongsulee (2017) og Sarker (2021), kan en dele feltet innen maskinlæring i fire deler: «supervised learning», «unsupervised learning», «Semi-supervised learning», og

«reinforcement learning» (Ongsulee, 2017, s. 2; Sarker, 2021, s.3-4). Tidemann beskriver de i Store Norske Leksikon som: «veiledet læring» (supervised learning), «ikke-veiledet læring» (unsupervised learning), «semi-veiledet læring» (semi-supervised learning) og «forsterket læring» (reinforcement learning) (Tidemann, 2022). Det er denne oversettelsen som benyttes gjennom oppgaven. Videre er det i forskningen på kunstig intelligens, i all hovedsak, snakk om disse fire delene (som inngår i maskinlæring) og en annen type kalt «ekspertsystemer» (se for eksempel Tan (2017) for en utredning av slike systemer). Jeg har valgt å kun forholde meg til maskinlæring, dette fordi ekspertsystemer (slik som Tan (2017) beskriver dem) ikke er læringsbasert på samme måte som maskinlæring er. Det er og vanlig i dagligtale å bruke kunstig intelligens og maskinlæring omhverandre, mens ekspertsystemer er noe mindre kjent. En viktig innsikt om begrepet maskinlæringsstrategier, er at maskinlæringsstrategier ikke er det samme som kunstig intelligens-modeller. Det vil si, en modell kan benyttes til å løse et (eller flere) problem eller fenomen for oss mennesker og kan inneholde flere maskinlæringsstrategier. Maskinlæringsstrategier som begrep kan derfor forstås som rene typer, mens kunstig intelligens-modeller er den faktiske maskinen (og/eller kunstig intelligensen).

Delkapittel 5.1. vil derfor ta for seg Ongsulee (2017) og Sarkers (2021) beskrivelse av maskinlæringsteknikker. Med bakgrunn i ideen om at kunstig intelligens er 'ikke-tenkende' har jeg til hver av teknikkene utarbeidet et begrep som omfavner hvordan å forstå dette i sosiologiske termer. For at det skal være mulig å forstå maskinlæringsteknikkene har jeg valgt å først legge frem en kort redegjørelse. Begrepene vil bli utførlig diskutert i kommende delkapittel. Typene i min typologi heter følgende: 1. *veiledet ikke-tenkende*, 2. *ikke-veiledet ikke-tenkende*, 3. *semi-veiledet ikke-tenkende* og 4. *selv-lært ikke-tenkende*.

Hvis vi tillater oss en liten digresjon her, hvordan kan begrepene forstås om det er snakk om mennesker i samfunn? La oss først se på «*veiledet ikke-tenkende*»: Dette kan referere til individer eller grupper som følger instruksjoner eller veiledning uten å kritisk vurdere, reflektere over eller utøve uavhengig tenkning. Veiledningen kan komme fra autoriteter, tradisjoner, massemedia, sosiale normer og så videre. I denne kategorien kan handlinger og valg være sterkt påvirket av ytre krefter, som sosialt press eller autoritet. Hva så med «*ikke-veiledet ikke-tenkende*»? Dette kan referere til individer eller grupper som ikke tenker kritisk eller reflekterende, men som heller ikke har noen klar veiledning eller autoritet de følger. I

stedet kan deres handlinger være mer spontane, impulsstyrte, eller basert på innfall. Det er mindre struktur, og valg kan være sterkt påvirket av nåværende omstendigheter eller følelser.

Videre har vi «*semi-veiledet ikke-tenkende*», som kan referere til individer eller grupper som delvis følger veiledning eller autoriteter, men som også utviser noen spontane eller impulsstyrte handlinger. Det kan være en blanding av struktur og ustruktur, avhengig av situasjonen. I denne kategorien kan handlinger være påvirket av både ytre krefter og indre tilfeldighet. Og sist, «*selv-lært ikke-tenkende*», som kan referere til individer eller grupper som lærer og anpasser seg på egen hånd, uten veiledning fra autoriteter eller eksisterende strukturer, men som fremdeles ikke utøver kritisk eller reflekterende tenkning. Handlinger i denne kategorien kan være basert på prøving og feiling, med endringer som oppstår over tid basert på personlige erfaringer snarere enn ytre veiledning.

Det denne argumentasjonsrekken får frem, er at om vi antropomorfiserer en kunstig intelligens, altså menneskeliggjør den, vil den allikevel ha klare mangler. De tre første typene kan en argumentere for at fortsatt vil være under menneskelig kontroll, den siste derimot er her hvor det oppstår komplikasjoner. Et eksempel på en slik komplikasjon er å finne i Stanford Encyclopedia of Philosophy hvor Vncent C. Müller tar opp konseptet «singularitet» (Müller, 2020), som forøvrig er omdiskutert. Argumentet, som omtales som «Good-Chalmers argument», om singularitet går som følger:

- «Premiss 1: Det kommer til å oppstå kunstig intelligens (KI) (skapt av menneskelig intelligense (MI) som medfører at KI = MI)
- Premiss 2: Hvis det eksisterer KI, vil det og eksistere KI+ (skapt av KI)
- Premiss 3: Hvis KI + eksisterer, vil det og eksistere KI++ (skapt av KI+)
- Konklusjon: KI++ vil eksistere (og singularitet vil skje)» (Müller, 2020; min oversettelse).

Singularitet innebærer i denne sammenheng at menneskelig intelligens vil sakke akterut ettersom kunstig intelligens vil skape sterkere og sterkere maskiner/kunstig intelligenser som igjen skaper nye kunstige intelligenser, osv. Det som ofte forbindes med singularitet er at det til slutt vil være en herskende enhet som styrer alt og alle. Som Müller får frem, er dette en reell debatt som foregår. På lik linje med at en hverken kan bekrefte eller avkrefte om kunstig intelligens vil klare å skape en KI+, kan en heller ikke avfeie muligheten for det. For å forstå

hvorfor jeg har valgt begrepsendringen vil det i kommende delkapittel legges ut om Ongsulee (2017) og Sarkers (2021) sine definisjoner av maskinlæringsteknikker.

5.1 Maskinlæring: Sosiologisk typologisk lesning

For det første, handler «veiledet læring» ifølge Ongsulee (2017, s. 2-3) om at algoritmen får et sett med inndata sammen med tilsvarende korrekte utdata, og lærer ved å sammenligne sin faktiske utdata med de korrekte utdataene for å finne feil. Modellen modifiserer seg selv for deretter å tilpasse seg. Sagt på en annen måte, det er fire stadier: Treningsdata, læring fra treningsdata, prediksjon, evaluering. Først får kunstig intelligens-modellen treningsdata som brukes til å trene modellen. Treningsdataene er merket, det vil si at de allerede har de riktige svarene eller resultatene (kalt «etiketter» eller «merkelapper»). Deretter lærer modellen fra treningsdataene ved å forsøke å finne mønstre eller sammenhenger mellom inndataene og de tilsvarende utdataene. Videre, når modellen er trent, kan den brukes til å gjøre prediksjoner basert på nye, ukjente inndata. Til slutt blir modellen evaluert på et separat sett med data, kjent som testdata. Dette fordi ved å sammenligne modellens prediksjoner med de faktiske etikettene på testdata, kan man få en følelse av hvor nøyaktig modellen er sannsynligvis i fremtidige, ukjente situasjoner.

Et eksempel på en veiledet læringsalgoritme kan være å forutsi prisen på et hus basert på dets egenskaper, som størrelse, antall rom, beliggenhet, og så videre. En kan trene en veiledet-læringsalgoritme ved å bruke et datasett med hus og deres tilsvarende priser for å lære hvordan man kan lage nøyaktige forutsigelser. Et annet eksempel, og som er med å påvirker oss i det daglige, er de innebygde algoritmene til Facebook eller Instagram, henholdsvis personaliserte anbefalinger. Algoritmen lager personaliserte anbefalinger som er basert på brukerens tidligere atferd og interesser av innhold. For eksempel, når en bruker går gjennom sin Facebook- eller Instagram-strøm, vil algoritmen analysere deres tidligere interaksjoner og preferanser for å velge hvilke innlegg, videoer eller annonser som passer best for brukeren. Algoritmen vil også justere anbefalingene basert på brukerens tilbakemelding og interaksjoner med innholdet, og dermed lære og forbedre sine anbefalinger over tid. En kan på sett og vis forstå dette som at mennesket er med på å påvirke utfallet. Dette er dermed den første typen kunstig intelligens i min typologi: Veiledet ikke-tenkende.

Den andre typen bygger på Ongsulee sin beskrivelse av ikke-veiledet maskinlæring: «The system is not told the «right answer.» ... The goal is to explore the data and find some

structure within» (Ongsulee, 2017, s. 4). Dette innebærer at algoritmen ikke har noen forkunnskaper om dataen, og at den kun skal klassifiserer basert på gitte parametere. Ikke-veiledet maskinlæring benyttes ofte når en skal lete etter sammenheng eller finne informasjon i et datasett som ennå ikke er systematisert. Med systematisert menes det her at informasjonen er der, men at det ennå ikke er satt i system. For eksempel kan en kunstig intelligens bygget på ikke-veiledet maskinlæring være effektiv til å oppdage hvem som kjøper visse produkter. Dette kalles for «clustering» (Ongsulee, 2017, s. 4; Sarker, 2021, s. 3). «Clustering» innebærer å gruppere en stor mengde data i mindre grupper, eller «cluster», basert på likheter i egenskapene til dataen, uten at en har noen forhåndskunnskap eller etiketter på hva «clusteren» faktisk representerer.

Ta for eksempel en nettbutikk som ønsker å utforske mønstre i kjøpsatferd blant kunder. Denne nettbutikken kan bruke «clustering» for å gruppere kunder basert på deres kjøpsmønstre og egenskaper. Algoritmen vil analysere egenskapene til hver kunde, som kjøpshistorikk, kjøpsbeløp, alder, geografi, kjønn, og så videre. Den vil deretter gruppere kundene i såkalte «clusters» basert på likheter i disse egenskapene. Resultatet kan gi verdifull innsikt i kjøpsatferd og forbrukertrender blant kundene, som videre kan brukes til å tilpasse markedsføringsstrategier og forbedre kundeopplevelsen. I sosiale medier kan en uovervåket læringsalgoritme brukes for å analysere brukerdata for å identifisere grupper av brukere med lignende interesser og preferanser. Algoritmen kan gruppere brukere basert på deres atferd, som hvilke sider de liker, hvilke poster de engasjerer seg mest med, eller hvilke hashtags de ofte bruker. Hovedforskjellen mellom veiledet- og ikke-veiledet maskinlæring, er at ikke-veiledet maskinlæring selv leter etter årsakssamenhenger, og som Sarker påpeker: «... analyzes unlabeled datasets without the need for human interferences» (Sarker, 2021, s. 4). Utfallet blir dermed bestemt av algoritmen, ikke mennesket. Den andre typen kunstig intelligens i min typologi er dermed: Ikke-veiledet ikke-tenkende.

Den tredje typen maskinlæring bygger på en kombinasjon av de to foregående teknikkene og kalles «semi-veiledet læring» (Ongsulee, 2017, s. 3; Sarker, 2021, s. 4). Som Sarker beskriver: «The ultimate goal of a semi-supervised learning model is to provide a better outcome for prediction than that produced using the labeled data alone from the model» (Sarker, 2021, s. 4). Denne typen maskinlæringsteknikk ble tidligere benyttet for blant annet å utvikle ansiktsgjenkjenningsprogram (Ongsulee, 2017, s. 3). Et eksempel på semi-veiledet læring i sosiale medier i dag, finner vi hos Facebook i algoritmen deres: «Personer du kanskje

kjenner» (Facebook, u.å.). Denne algoritmen bruker en kombinasjon av de foregående teknikkene. Ved å gi egen innputt, gjennom å akseptere eller avvise forespørselen om å legge til andre mennesker, påvirker brukeren kunstig intelligensmodellen. Dermed, ved å kombinere både veiledet- og ikke-veiledet læring, kan semi-veiledet læring bidra til å forbedre kvaliteten og effektiviteten til algoritmen i å gi en personaliserte forespørsler om individer, i sosiale medier. Henholdsvis blir den tredje typen kunstig intelligens i min typologi: Semi-veiledet ikke-tenkende.

Den fjerde og siste maskinlæringstypen er «reinforcement learning» (forsterket læring) (Ongsulee, 2017, s. 3; Sarker, 2021, s. 4). Et eksempel på en kunstig intelligens-modell basert på forsterket læring er Google sin AlphaGo som spesialiserer seg på brettspillet Go (DeepMind, u.å.). AlphaGo ble internasjonalt kjent i 2016 da den overraskende slo verdensmester Lee Sedol i en femkampsserie, noe som markerte første gang en kunstig intelligensmodell overgikk menneskelig ytelse på høyt nivå i dette spillet (Moyer, 2016). Forsterket læring bygger på følgende premisser: Det trengs en «agent» (det vil si noen eller noe som er det som skal lære), et miljø, en form for handling, og belønningssystem, som sumasumarum ender i læring. I eksempelet med AlphaGo er det følgende premisser: Agenten (den som utfører) er her AlphaGo, miljøet er selve Go-brettet samt motstanderen, handlingen er det som tillates i spillet Go, belønningen er poengsummen som agenten får basert på kvaliteten på trekket som den har gjort, læringen forekommer gjennom å prøve og feile gjentatte ganger. AlphaGo vil dermed teste ut og gjøre trekk om og om igjen til den finner de beste trekkene. Med andre ord, forsterket læring bygger på dikotomien positiv- versus negativ stimuli («reward versus penalty») (Sarker, 2021, s. 4). Dette er det første konkrete eksempelet på maskinlæring som kopierer en del av menneskelig væremåte. Den fjerde og siste typen i min typologi er dermed: Selv-lært ikke-tenkende.

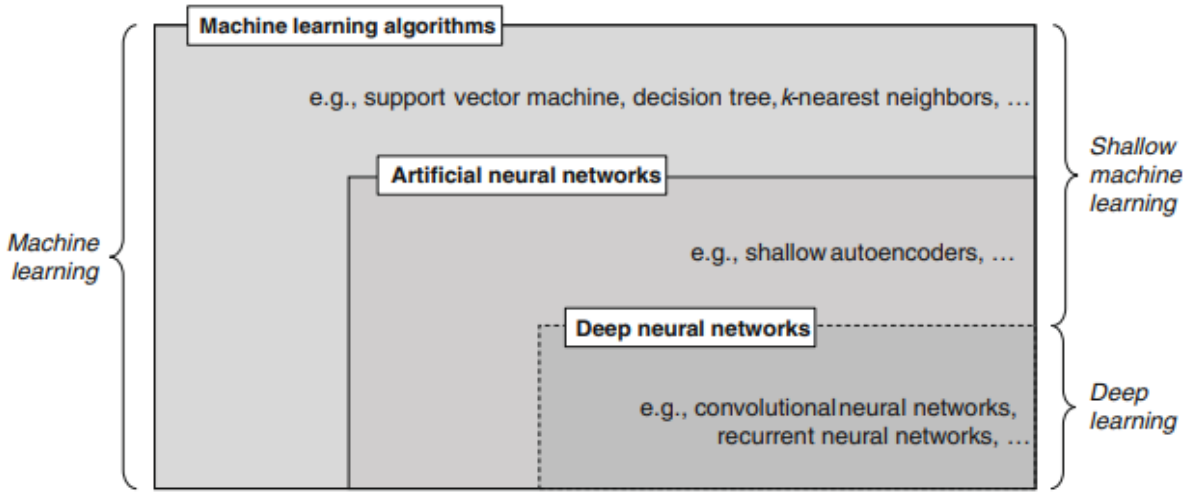
Oppsumert er typene i min typologi følgende, veiledet ikke-tenkende, ikke-veiledet ikke-tenkende, semi-veiledet ikke-tenkende og selv-lært ikke-tenkende. Videre er typologien her ment å være såpass generell at den kan brukes for flere aspekter av forskningen på kunstig intelligens. Dette fordi, som vil bli vist i kommende delkapittel, finnes det flere nivåer innen kunstig intelligensforskningen. Disse nivåene er helt nødvendig å redegjøre for, siden de har innvirkning på typologien jeg her har laget. For å tydeliggjøre hvordan typene skal brukes kommer her et eksempel: En kunstig intelligensmodell slik som «spam filter» (IBM, u.å.b.) vil inngå i typen «veiledet ikke-tenkende kunstig intelligensmodell». Jeg har understreket

«kunstig intelligensmodell» siden det er slike fenomener sosiologien vil ha med å gjøre, ikke hvordan den underforliggende mekanikken i modellen fungerer/virker. Jeg argumenterer for at denne typologien vil gjøre det lettere å forske på fenomener i sosiologien knyttet til kunstig intelligens. Med andre ord, det jeg her har prøvd å få frem, er at ved å se til kunstig intelligensfeltet er det fruktbart å forstå variasjonen som eksisterer der. Dette fordi det har stor innvirkning inn på hvordan vi i sosiologien analyserer kunstig intelligens som fenomen.

5.2 Tre nivåer av kunstig intelligens

Før jeg gjør et dypdykk inn i OpenAI sin kunstig intelligens ChatGPT (OpenAI, 2023), trengs den en oppklaring i hva som ligger i begrepene «artificial neural network» (kunstig-nevrale-nettverk) og «Deep learning» (dyp-læring) (Sarker, 2021, s. 13-15; Ongsulee, 2017, s. 3-4). Maskinlæring, kunstig nevrale nettverk og dyp-læring, blir ofte snakket om og diskutert som lignende begreper. Dette er svært misvisende. Maskinlæring er paraplybetegnelsen som omfatter både kunstig-nevrale-nettverk, dyp-læring, og andre typer maskinlæring. Kunstig-nevrale-nettverk er igjen en paraplybetegnelse for diverse spesifikke maskinlæringsteknikker. Mens dyp-læring defineres av Ongsulee som: «The study of artificial neural networks and related machine learning algorithms that contain more than one hidden layer» (2017, s. 3). Det vil si at dyp-læring er å finne når en benytter seg av visse typer nevrale nettverk (men ikke med nødvendighet alle). Det mest vesentligst å merke seg, er at hovedforskjellen mellom dyp læring og generell maskinlæring ligger i at dyp læring er mer avansert og etterligner menneskelig kognisjon i større grad enn det tradisjonell maskinlæring gjør. Dyp-læring kan dermed defineres som en type arkitektur innen kunstig nevrale nettverk, som igjen er å finne i maskinlæring. I litteraturen på kunstig intelligens brukes begrepene «dyp-læring» og «dype-nevrale-nettverk» omhverandre. For at det skal være enklere å følge med i argumentasjonsrekken har jeg valgt å kun benytte begrepet dyp-læring.

Janiesch, Zschech og Heinrich (2021) sitt venn-diagram er svært nyttig for å se kobling mellom maskinlæring, kunstig nevrale nettverk og dyp-læring:



Figur 1. «Venn-diagram av maskinlæringskonsepter og klassifikasjoner» hentet fra (Janiesch, Zschech, Heinrich, 2021, s. 687; min oversettelse).

Som en ser av venn-diagrammet, er maskinlæringsalgoritmer det overordnede rammeverket. Dette kan, i forskjellige sammenhenger, inneholde kunstige nevrale nettverk, som er såkalt «shallow» (dvs. inneholder kun en «hidden layer», se Pasupa og Sunhem (2016) for en utfyllende studie av forskjellen mellom «shallow» kunstig nevrale nettverk og dyp læring). Eksempelet Janiesch, Zschech og Henirich gir er «shallow autoencoders» (2021, s. 687). Sagt på en enklere måte, innebærer dette at de fire overordnede maskinlæringsteknikkene kan forbedres gjennom bruk av først kunstig nevral nettverk, og så dyp læring. Jeg har laget en tabell for å tydeliggjøre dette poenget. I første kolonne gis et eksempel av først «simpel» maskinlærings-algoritme, deretter kunstig nevral nettverk, og til sist dyp-læring.

Tabell 1. En oversikt over maskinlæringsnivåene i kunstig intelligens. Denne fremstillingen er basert på Ongsulee (2017), Sarker (2021), Murtagh (1991), Wang og Zhou (2007) og Kim, Choi og Kim (2019).

Overordnede typer i maskinlæring	**Veiledet maskinlæring**	**Ikke-veiledet maskinlæring**	**Semi-veiledet maskinlæring**	**Forsterket læring**
«simpel» maskinlærings-teknikker	Decision Tree	K-means Clustering	Constrained-based clustering	Q-Learning
Kunstig nevrale nettverk-teknikker	Multi-Layer Perceptron	Self-Organizing Maps	Co-Training	Neural Q-Learning
Dyp-lærings-teknikker	Convolutional Neural Networks	Deep Belief Networks	Virtual Adverserial Training	Deep Q Network

I tabellen min er det bare gitt *ett* eksempel på maskinlæringsteknikk. Dette fordi det eksisterer svært mange, og hensikten her er bare å lage en forståelig oversikt og å bygge videre på venn-diagrammet til Janiesch, Zschech og Heinrich. Et av argumentene mine i analysen bygger på nettopp skillet mellom typer kunstig intelligens. Det er derfor vitalt å få frem dette skillet. For å utdype tabellen vil jeg her ta for meg eksempelet spamfilter (IBM, u.å.b). Den kunstig intelligensmodellen «spamfilter» vil her bli diskutert først med arkitektur av «simpel» maskinlæringsteknikk, deretter arkitektur av kunstig nevrale nettverks-teknikk og så arkitektur av dyp-læringsteknikk. Spam defineres som «uønsket e-post», også kalt «søppelpost», og omfatter ofte reklame e.l. (Det Norske Akademis Ordbok, 2023). Oftest er spam å forstå som uønsket oppmerksomhet som en gjerne skal være foruten.

Jobben til spamfiltre er å si ja eller nei til inkommende e-poster. De første spamfiltrene benyttet antageligvis «ren» veiledet maskinlæring (se for eksempel Abdulrahman og Salim (2022) for en komparativ analyse av bruken av Decision Trees i spamfilter). Ettersom teknologien utviklet seg og kunstig nevrale nettverk kom til, forbedret en spamfiltrene. Sharmin, Troia, Potika og Stamp (2022) foreslår enda en forbedring for spamfiltre ved å benytte dyp-læringsstrategien «Convolutional Neural Networks». Som en her ser, kan allerede eksisterende maskinlæringsteknologi videreutvikles med først bruk av kunstig

nevrale nettverk, for så igjen forbedres med dyp-læring. Tabellen må derfor ikke leses som ekskluderende typer for kunstig intelligens. Med utgangspunkt i tabellen og spamfilter eksempelet har jeg laget en supplerende modell som tydeliggjør det siste poenget:

Figur 2: lagd av undertegnede basert på tabell 1. (Venstre til høyre) Fra mindre avansert spamfilter til mer avansert spamfilter

«Several stage neural networks» (Alkhat og Khatib, 2016) kan innebære bruk av dyp-læring, og er derfor tatt med som eksempel for den nyeste iterasjonen av effektiv spamfilter.

Etter en slik lang utredning om variasjon av kunstig intelligens, er et reelt spørsmål å stille: Hva har dette med sosiologi å gjøre? Dette spørsmålet har to svar: 1. Det eksisterer ikke en slik overordnet analyse/utredning av variasjonen innen kunstig intelligens i sosiologiske termer, som fører med seg; 2. at uten en slik overordnet analyse/fremlegg lar det seg heller ikke gjøre å analysere kunstig intelligens opp mot sosiologisk teori (uten å gjøre antagelser basert på mangelfult grunnlag). Som jeg her har vist er det altfor stor variasjon innen kunstig intelligens til at en bare kan anta at alle kunstig intelligensmodeller fungerer og operer likt. Et påfølgende spørsmål å stille seg er: Hvorfor er det relevant for sosiologien å se på byggestenene av kunstig intelligenser, samt sette seg inn i dette feltet? Jeg argumenterer, gjennom bruk av Habermas og Luhmanns kommunikasjonsteorier, for at å faktisk forstå om kunstig intelligens kan delta i kommunikasjon, kreves det klargjøring av hva kunstig intelligens som fenomen er. Mitt argument bunner i at, uavhengig av analyseverktøy, kan en ende opp med å gjøre feilslutninger om en ikke tar inn over seg diversiteten som eksisterer i kunstig intelligensmodeller.

Sagt på en annen måte, dyp-læring er mer avansert enn generelle maskinlæringsalgoritmer og kunstig-nevrale-nettverk. Et eksempel er å finne hos Michaud, Liu, Girit og Tegmark i «The Quantization Model of Neural Scaling» (2023), hvor emergensteori blir brakt inn for å forklare hvordan store språkmodeller har evner utover det mindre språkmodeller har. Store språkmodeller bygger på dyp-læring, og i artikkelen til Michaud, Liu, Girit og Tegmark

postuleres det at slike kunstig intelligens systemer får en eller flere emergente egenskaper. Dette innebærer egenskaper som tidligere iterasjoner av slike systemer ikke hadde, og som kun kommer av at systemet er blitt bedre og mer avansert. Emergensteori har vært anvendt før i sosiologien for å prøve å forklare at noe mer oppstår når en gruppe individer kommer sammen (se for eksempel Sawyers (2001) artikkel). I korte trekk handler emergensteori om at helheten av noe utgjør mer enn bare delene den er bygget opp av (O'Connor, 2020). På samme måte virker det intuitivt som om OpenAI sin kunstig intelligens: ChatGPT (OpenAI, 2023), innehar noen slike emergente egenskaper.

Denne intuitive oppfattelsen støttes av Wei et al. (2022) i deres artikkel: «Emergent Abilites of Large Language Models». I korte trekk fant Wei et al. at fenomenet med emergente evner oppstår i store språkmodeller, og som ikke er til stede i mindre modeller. Disse evnene, som ikke kan forutsies av skalalover (Wei et al., s. 1-2), viser en dramatisk forbedring i ytelse etter at en viss skaleringsgrense er nådd. Skalering av språkmodeller omfatter i denne artikkelen økning i beregningsmengde, antall modellparametre og treningsdatasettets størrelse. Det diskuteres og at fremveksten av nye evner avhenger av en rekke faktorer, inkludert datakvalitet og optimal modelltreningspraksis. Denne noe tekniske fremstillingen viser at selv 'gamle' ChatGPT var i stand til å få emergente egenskaper. Wei et al. diskuterer og, i kapittel 5.5, at det er et sosiologisk skift ved bruk av NLP («natural language processing») mot allmenformål («general-purpose»), noe som er muliggjort ved økende skala av språkmodeller. Med dette peker Wei et al. på at generelle modeller, som GPT-3 og PaLM, utkonkurrere tidligere oppgavespesifikke modeller på ulike benchmarks (Wei et al, 2022, s. 10). Evnen til allmennformålmodeller til å utføre usette («unseens») oppgaver med få eksempler har åpnet for nye applikasjoner utenfor NLP-forskning, inkludert bruk i robotikk, brukerinteraksjon, multi-modal resonnering og kommersielle tjenester som OpenAI's Chat-GPT (Wei et al, 2022, s. 10).

Det Wei et al. (2022) får frem her, er at allmennformålmodeller er kommet for å bli. Disse kunstig intelligensmodellen, som pekt på tidligere, har skapt furore og usikkerhet i flere sektorer. For eksempel må lærere i skolen og universitet tenke nytt og annerledes rundt hvordan å kvalitetssikre eksamen dersom en har tilgang på slike store språkmodeller. Jeg velger her å benytte store språkmodeller siden det er disse som er relevante i skrivende stund. I en artikkel fra Bubeck et al. (2023): «Sparks of Artificial General Intelligence: Early experiments with GPT-4», viser forfatterne hvordan den nye iterasjonen av GPT, GPT-4, er

såpass avansert at den kan sees på som «... en tidlig (men fortsatt ufullstendig) versjon av en generell kunstig intelligens (GKI) system» (Bubeck et al., 2023, s. 1; min oversettelse). Dette på grunn av omfanget og dybden av GPT-4s kapabiliteter. I korte trekk handler artikkelen om å demonstrerer hvordan GPT-4 kan løse varierte og vanskelige oppgaver i mange felt, uten behov for spesiell «prompting» (det vil si måten å legge frem oppgaven), og ofte med ytelse nær menneskenivå. Forfatterne diskuterer og begrensningene ved GPT-4 og utfordringene fremover for å utvikle dypere og mer omfattende versjoner av GKI. De avslutter med noen refleksjoner om samfunnsmessige påvirkninger av den nylige teknologiske fremgangen og fremtidige forskningsretninger. Hvordan passer så alt dette her inn hos Habermas og Luhmann?

6 Analyse av Habermas og Luhmann opp mot min typologi

I denne analysen vil jeg først undersøke kommunikasjonsbegrepet til Jürgen Habermas opp mot typologien min, deretter gjøres det samme med Niklas Luhmann. Jeg har delt KI-systemer inn i følgende typer: «Veiledet ikke-tenkende KI-modell», «ikke-veiledet ikke-tenkende KI-modell», «semi-veiledet ikke tenkende KI-modell» og «selv-lært ikke-tenkende KI-modell». Begge sosiologene tilbyr unike perspektiver på kommunikasjonsprosessen som kan gi verdifulle innsikter i vår forståelse av hvordan kunstig intelligens fungerer og interagerer i samfunnet. Habermas sin teori om kommunikativ handling, bygger på grunnpilarene talehandling og gyldighetskrav. Det er disse to grunnpilarene jeg vil se opp mot typologien jeg har utredet om. For at dette skal la seg gjøre har jeg valgt ut en type kunstig intelligensmodell som vil diskuteres. Dette gjøres og for Luhmann men da med fokus på hans tredeling av meddelelese, informasjon og forståelsen av forskjellen mellom disse to. Disse teoretiske rammeverkene presenterer to distinkte måter å forstå kommunikasjon i kontekst av kunstig intelligens.

6.1 Habermas og min typologi

Det vil i dette delkapittel bli diskutert Habermas sine talehandlinger samt gyldighetskrav. Som beskrevet tidligere, er talehandlinger rendyrkede former. En forutsetning Habermas gjør er at alle mennesker kan med nødvendighet, men ikke må ved nødvendighet, benytte seg av de tre rendyrkede formene for talehandling. En forutsetning jeg gjør basert på dette, er at kunstig intelligens må og kunne (men ikke må) ved nødvendighet, benytte alle tre formene for rendyrkede talehandlinger. Dette innebærer at der en kunstig intelligens ikke kan benytte en

eller flere av Habermas sine talehandlinger er det heller ikke nødvendig å gå videre for å se på gylidghetskravene som beskrives. Dette fordi KI-modellen dermed ikke kan inngå i helheten Habermas legger ut om. For å tydeliggjøre har jeg laget en tabell av min typologi:

Tabell 2. Eksempler på KI-modeller innenfor min typologi

	Veiledet ikke-tenkende KI-modell	Ikke-veiledet ikke-tenkende KI-modell	Semi-veiledet ikke-tenkende KI-modell	Selv-lært ikke-tenkende KI-modell
Eksempel	Spamfilter	Kundepersonlighet	Personer du kanskje kjenner	GPT-4

La oss ta for oss den første typen kunstig intelligensmodell: Veiledet ikke-tenkende. For at det skal være lettere å følge med i argumentasjonsrekken tar jeg utgangspunkt i en eksisterende kunstig intelligens basert på veiledet ikke-tenkende: «Spam filter» (IBM, u.å.b). Et spamfilter får instrukser om hvordan det skal operere, og gjennomfører så disse instruksene. På den ene siden kan en argumentere for at spamfiltrene deltar i kommunikasjon gjennom *en* av de tre talehandlingene til Habermas: «Regulerende». Dette fordi det er individer som forteller spamfilteret hva som forkastes og hva som godkjennes. Derimot er det vanskelig å peke på eksempler hvor spamfilteret enten gjør nytte av/deltar i «uttrykkende» eller «regulerende». Således er det heller ikke nødvendig å se på om spamfilteret kan benytte seg av gyldighetskravene, siden veiledet ikke-tenkende kun deltar i kommunikasjon gjennom en av tre nødvendige talehandlingene. Jamfør Habermas' teori kan en argumentere for at modeller basert på veiledet ikke-tenkende ikke deltar i tilstrekkelig grad til at en kan kalle det for kommunikasjon.

Den andre typen maskinlæring som er beskrevet er ikke-veiledet ikke-tenkende. Et eksempel på en kunstig intelligensmodell basert på ikke-veiledet ikke-tenkende, gitt av IBM, er «kundepersonlighet» (IBM, u.å.c; min oversettelse). En «kundepersonlighet» er en ansamling data av et individ eller gruppe individer. Denne mengden med data benyttes for at selskaper bedre skal gi for eksempel tilbud eller annonsering av produkter til grupper. Det er her snakk om prediksjonsmodeller. For en utfyllende beskrivelse av prediksjonsmodeller, se Shoshana Zuboffs bok «The Age of Surveillance Capitalism: The fight for the future at the new frontier

of power» (2019). En slik kunstig intelligensmodell kan få frem, jamfør Habermas' teori, «konstanterende» setninger. For eksempel: Individ A har gjort x, y og z, og vil derfor med høy sikkerhet være interessert i å se/kjøpe Q. Slike assertoriske setninger bygger på premisset om at det ikke er hundre prosent sikkert, noe denne kunstig intelligensmodellen bygger på. Ikke-veiledet ikke-tenkende KI-modeller kan derimot ikke delta i eller benytte seg av uttrykkende- eller regulerende talehandlinger. Dette fordi uttrykkende talehandling fordrer en eller annen form for persepsjon, noe ikke-veiledet ikke-tenkende KI-modeller mangler. Ikke-veiledet ikke-veiledet typen mangler og mulighet for å delegere eller gi kommando, og kan med det ikke benytte regulerende talehandling. En kan dermed argumentere for at ikke-veiledet ikke-tenkende, ikke kan delta i kommunikasjonsbegrepet til Habermas.

Den tredje typen av kunstig intelligensmodeller, er som beskrevet: Semi-veiledet ikke-tenkende. Semi-veiledet ikke-tenkende bygger som nevnt på en kombinasjon av de to foregående teknikkene. Med dette i mente, kan en her argumentere for at jamfør Habermas talehandlinger, kan semi-veiledet ikke-tenkende både benytte seg av «regulerende-» og «konstanterende» talehandlinger. Dette nettopp fordi semi-veiledet ikke-tenkende KI-modeller er en kombinasjon av de to foregående formene for KI-systemer. Videre kan semi-veiledet ikke-tenkende KI-modller ikke benytte seg av «uttrykkende» talehandlinger, av samme grunn som hverken veiledet ikke-tenkende- og ikke-veiledet ikke-tenkende KI-systemer kan. Sagt på en annen måte kan en forstå Habermas dithen at det krever en eller annen form for subjektiv oppfatning for å kunne uttale setninger/kommunisere om slike forhold. Denne subjektive oppfatningen mangler veiledet ikke-tenkende, ikke-veiledet ikke-tenkende og semi-veiledet ikke-tenkende.

Som beskrevet tidligere er et eksempel på en kunstig intelligensmodell i typen semi-veiledet ikke-tenkende, Facebook sin algoritme: «Personer du kanskje kjenner» (Facebook, u.å.). Uten å gå for mye i detalj, innebærer dette at Facebooks algoritme vet visse parametere:

- «At man har felles venner
- nettverkene dine, for eksempel byen du bor i nå, skole eller jobb
- å være i samme Facebook gruppe
- å bli tagget i samme bilde eller innlegg
- kontakter du har lastet opp» (Facebook, u.å.).

Modellen baserer seg med det på visse data som er «markert» (labeled), og visse data som er «umarkert» (unlabeled) (se Sarker (2021), eller Ongsulee (2017), for utfyllende om hva «labeled» versus «unlabeled» data innebærer). I korte trekk handler det her om at modellen både får beskjed om hva den skal gjøre av mennesker, samt at visse deler må den «finne ut av selv». Eksempelet her viser hvordan semi-veiledet ikke-tenkende modeller benytter både strategier fra veiledet- og fra ikke-veiledet ikke-tenkende kunstig intelligensmodeller.

Den fjerde typen av kunstig intelligensmodeller er selv-lært ikke-tenkende. Jeg vil her ta for meg kunstig intelligensmodellen GPT-4 (OpenAI, 2023) som eksempel for en selv-lært ikke-tenkende. La oss nå vurdere hvordan GPT-4 kan håndtere hver av disse typene. For det første, regulerende talehandlinger er elementært avgjørende eller elementært bevisste/tilsiktede. Eksempler kan inkludere anmodninger, forslag, påbud eller instruksjoner. GPT-4 er designet til å reagere på en hel rekke forskjellige slike (både avgjørende og bevisste/tilsiktede), inkludert de som ville kreve en respons til en regulerende talehandling. For eksempel, hvis en bruker gir en instruksjon eller forespørsel, vil GPT-4 generere en respons som er relevant for den givne instruksjonen. Dette er ikke for å si at GPT-4 har noen forståelse av sosiale normer eller konvensjoner, men snarere at den er i stand til å simulere en respons til en regulerende talehandling basert på hvordan den er trent. For det andre, uttrykkende talehandlinger er de som uttrykker talerens subjektive tilstand. Dette kan inkludere følelser, ønsker, tro osv. Her kan en arugmentere for at GPT-4 møter en betydelig begrensning, fordi som en kunstig intelligensmodell har den ingen subjektive tilstander å uttrykke. Selv om den kan generere tekst som simulerer en uttrykkende talehandling, vil denne teksten ikke være et uttrykk for noen faktisk subjektiv tilstand. En kan allikevel argumentere for at gjennom simulering virker det som om GPT-4 innehar en slik subjektiv tilstand, til tross for at den faktisk ikke har det.

For det tredje, konstanterende talehandlinger er de som har til hensikt å formidle eller bekrefte en assertorisk påstand. GPT-4 er i stand til å generere slike talehandlinger til en hvis grad. Den kan generere tekster som inneholder informasjon som den ble trent på, og som dermed kan betraktes som en form for konstanterende talehandling. Men, som nevnt tidligere, har den ingen evne til å bekrefte eller verifisere sannheten av den informasjonen som den genererer. Der er her jeg argumenterer for viktigheten i å forstå begrepet ikke-tenkende kommer inn. Siden GPT-4 kan simulere forskjellige typer talehandlinger i noen grad, er det viktig å merke seg at den faktisk ikke kan utføre disse handlingene på samme måte som en menneskelig taler ville gjort. Derfor er GPT-4 en ikke-tenkende kunstig intelligensmodell som simulerer menneskelig interaksjon og væremåte. Hvordan passer så GPT-4 inn i Habermas sine

gyldighetskrav? Habermas sine gyldighetskravene har relevans når vi vurderer GPT-4, og forsåvidt andre KI-baserte språkmodeller, i sammenheng med å simulere eller utføre talehandlinger. La oss se hvordan GPT-4 kan relateres til hver av disse gyldighetskravene.

For det første har vi «aspektet av riktighet». Dette aspektet referer til «riktigheten som taleren påstår for sin handling i forhold til en normativ kontekst». GPT-4 er designet og trent til å generere tekst som er relevant og hensiktsmessig gitt innputten og kontekst. Dette betyr at GPT-4 prøver å generere svar som er «riktige» i den forstand at svarene er konsistente med den gitte konteksten, og normer for språkbruk. Imidlertid er det verdt å påpeke at mens GPT-4 kan generere normativt korrekte svar, kan den ikke forstå eller påstå normativitet i samme grad som et menneske kan, siden den ikke har noen bevissthet eller moralsk forståelse.

For det andre, «sannferdigheten som taleren påstår for uttrykket av subjektive opplevelser som han har privilegert tilgang til». Her møter vi en begrensning hos GPT-4. Som en kunstig intelligensmodell har GPT-4 ingen subjektiv opplevelse og dermed heller ingen privilegert tilgang til noen form for subjektiv opplevelse. En kan med det argumentere for at den ikke kan møte Habermas' gyldighetskrav til sannferdighet, uten en modifisering. Som i argumentet over, virker det intuitivt som GPT-4 klarer dette, gjennom burk av begrepet ikke-tenkende argumentere jeg derfor for at den klarer å simulere en slik opplevelse i interaksjon med mennesker.

For det tredje, «sannheten som taleren, med sin ytring, påstår for en påstand». GPT-4 kan generere informasjon som er sann, i den forstand at det er i tråd med den kunnskapen den ble trent på. Dette er på lik linje med mennesker, om mennesker blir feilinformert kan de fortsatt tro at deres ytring er sann. På lik linje fungerer GPT-4, om den feilinformeres tror den fortsatt at den informasjonen den er gitt er sann. Den kan heller ikke bekrefte sannheten av nye opplysninger som den ikke har blitt trent på, og den har heller ingen evne til å verifisere sannheten av dens egne uttalelser. Så, selv om GPT-4 kan på sett og vis tilfredsstille Habermas gyldighetskrav i noen grad, er det viktig å understreke at dette bare er innenfor rammen av dens design og begrensninger. Henholdsvis, mangler GPT-4 subjektive opplevelser og evnen til å bekrefte eller undersøke sannheten av sin egen kunnskap, noe som er sentrale aspekter av menneskelig kommunikasjon og talehandlinger.

I lys av en dypere utforskning av Habermas' talehandlinger, gyldighetskrav og deres relevans for kunstig intelligens, blir betydningen av begrepet mitt «ikke-tenkende» enda mer tydelig.

Dette fordi det peker på at ikke alle KI-modeller kan sidestilles når en snakker om Habermas sitt kommunikasjonsbegrep. Min analyse av ulike KI-modeller: «Spamfilter», «kundepersonlighet», «personer du kanskje kjenner» og «GPT-4», viser dermed at disse modellene har begrensninger når det kommer til talehandlinger. En sentral forutsetning jeg gjør er at for å kunne snakke om gyldighetskrav må og KI-modellen kunne benytte seg av de tre typene for talehandling. Dette er det kun GPT-4 som kvalifiserers for, og en kan med det fastslå at KI-modeller som inngår i typene: «Veiledet ikke-tenkende», «ikke-veiledet ikke-tenkende» og «semi-veiledet ikke-tenkende» ikke kan inngå i kommunikasjonsbegrepet til Habermas. Derimot er det noe mer usikkert med GPT-4, som her representerer «selv-lært ikke-tenkende» KI-systemer.

På den ene siden er GPT-4 en ikke-tenkende kunstig intelligensmodell som simulerer menneskelig interaksjon og væremåte, uten evnen til å være bevisst sine handlinger. På den andre siden kan GPT-4 generere svar som er normativt korrekte og konsistente med den gitte konteksten, dog uten subjektiv opplevelse. Oppsumert kan en dermed si at GPT-4 ikke klarer å møte Habermas krav til sannferdighet i sin reneste form. Videre kan GPT-4 heller ikke generere informasjon som er sann og basert på den kunnskapen den ble trent på, kan den heller ikke bekrefte sannheten av nye opplysninger. Derfor, selv om GPT-4 i noen grad kan tilfredsstille Habermas sine tre gyldighetskrav, er det viktig å understreke at dette bare er innenfor rammen av dens design og begrensninger som en ikke-tenkende kunstig intelligensmodell. Denne analysen har forhåpentligvis bidratt til en dypere forståelse av forholdet mellom kunstig intelligensmodeller og menneskelig kommunikasjon. Hva så med Luhmann sitt kommunikasjonsbegrep?

6.2 Luhmann og min typologi

Luhmann sitt begrep om kommunikasjon er ganske så annerledes enn Habermas sitt. Der hvor Habermas legger klare føringer for hva kommunikasjon mellom individer er, har Luhmann en mer overordnet tilnærming. Det trengs derfor en annen vinkling inn på hvordan å analysere kommunikasjonsbegrepet til Luhmann opp mot typologien min. Luhmanns tre bestandeler av kommunikasjon, meddelelse, informasjon og forståelsen av forskjellen mellom disse to, er et godt startspunkt. Hvordan passer så «spamfilter» (som her representerer en veiledet ikke-tenkende kunstig intelligensmodell) inn i Luhmanns begrep om kommunikasjon?

For å svare på dette spørsmålet har jeg valgt å dele opp meddelelse, informasjon og forståelsen i tre bestanddeler og analysere dem hver for seg. Som Baraldi, Corsi og Esposito (2021) påpeker realiseres disse samtidig. Jeg vil derfor først se på en av typene i min typologi opp mot hver bestanddel, for deretter å se på helheten som en «event». Følgende kunstig intelligensmodeller vil bli diskutert: Spamfilter (veiledet ikke-tenkende), kundepersonlighet (ikke-veiledet ikke-tenkende), «personer du kanskje kjenner» (semi-veiledet ikke-tenkende), og GPT-4 (selv-lært ikke-tenkende). Hvordan passer så de forskjellige ikke-tenkende kunstig intelligensmodellene inn i Luhmann sitt kommunikasjonsbegrep?

La oss først undersøke hvordan å forstå «spamfilter» inn i dette rammeverket. I Luhmanns teori refererer meddelelse til selve handlingen med å overføre en melding. Meddelelsen er objektiv og kan bli oppfattet av alle som er til stede. I sammenheng med et spamfilter, kan vi se meddelelsen som selve e-posten eller meldingen som sendes. Videre refererer informasjon til det som blir sagt i meddelelsen - det underforliggende budskapet. Denne delen av kommunikasjon er subjektiv og kan tolkes forskjellig av ulike mottakere. Et spamfilter prøver å oppdage denne informasjonen ved å analysere innholdet i meddelelsen og bestemme om det er spam eller ikke-spam. Deretter er prosessen med å tolke og gi mening til informasjonen som er mottatt. Mennesker gjør dette gjennom sin egen personlige erfaring og kunnskap. Et spamfilter kan ikke egentlig «forstå» informasjonen på samme måte som et menneske, men den kan bruke algoritmer og maskinlæring for å «forstå» om en melding sannsynligvis er spam basert på tidligere erfaringer.

På dette vis kan en se på spamfilter som en type «kommunikator» som tolker meddelelser (e-poster, meldinger, e.l.), analyserer informasjonen i dem, og «forstår» (gjennom algoritmer og maskinlæring) om disse meldingene sannsynligvis er spam eller ikke-spam. Ved å bruke dette rammeverket argumenterer jeg her for at spamfilteret er en aktiv deltaker i kommunikasjonsprosessen, selv om det ikke er et menneske. I videreføringen kan det argumenteres for at spamfilteret er en type sosialt system som bidrar til å strukturere og filtrere kommunikasjonen som foregår i det digitale rommet. Jamfør en slik tolkning er Elena Esposito sitt begrep «artificial communication» svært fruktbart. Hva så med ikke-veiledet ikke-tenkende, i dette tilfellet kundepersonlighet?

En kundepersonlighet, er som nevnt, en representasjon av den ideelle kunde basert på markedsforskning og faktiske data om eksisterende kunder. Det er et viktig verktøy for bedrifter for å forstå og kommunisere effektivt med deres målkunder. I denne kunstig

intelligensmodellen er meddelelsen den faktiske kommunikasjonen som sendes til kunden. Jamfør kundepersonlighet, kan dette være en markedsføringsmelding som er spesielt utformet for å appellere til en bestemt kundepersonlighet. Videre, kan meddelelsen være designet for å matche personaens interesser, behov, demografi og så videre. Henholdsvis brukes dette for å øke sjansene for at meldingen vil bli mottatt positivt. I meldingen er det informasjon, og er det som er innholdet eller betydningen av meddelelsen. I henhold til kunstig intelligensmodellen kundepersonlighet, kan dette være den spesifikke informasjonen som er relevant for en gitt kundepersona. For eksempel, hvis kundepersonaen er en ung, miljøbevisst person, kan informasjonen fokusere på hvordan produktet eller tjenesten er miljøvennlig. Dette bunner i forståelse, som er mottakerens tolkning av informasjonen. I forhold til den ikke-veiledet ikke-tenkende KI-modellen: Kundepersonlighet, kan dette innebære hvordan personaen oppfatter og reagerer på informasjonen basert på deres individuelle behov, interesser og verdier. Som med forrige KI-modell, gir denne fremstillingen god grunn til å bygge videre på Eposito sitt begrep «artificial communication» som et frukbart begrep. Hva så med den tredje kunstig intelligensmodellen «personer du kanskje kjenner»?

Som nevnt, er «personer du kanskje kjenner» en funksjon på Facebook som foreslår potensielle venner basert på ulike faktorer. Dette kan være faktorer som: Felles venner, skole, arbeidsplass, og mer. Den semi-veiledede ikke-tenkende kunstig intelligensmodellen bak denne funksjonen analyserer disse forskjellige faktorene for å foreslå mennesker som en sannsynligvis vil kjenne eller være interessert i å legge til som «venn». Jamfør Luhmanns tredeling kan «personer du kanskje kjenner» KI-modellen plasseres som følger. Meddelelsen tilsvarer de faktiske forslagene som Facebook gir deg om «personer du kanskje kjenner». KI-modellen lager denne meddelelsen basert på informasjonen den har om deg og de foreslåtte vennene, i henhold til de overnevnte faktorene. Videre kan informasjon forstås som den faktiske informasjonen som brukes av algoritmen til å lage forslagene. Dette inkluderer blant annet informasjon om ens eksisterende venner, ens aktiviteter på Facebook, og annen relevant informasjon. Samt, at denne informasjonen er subjektiv, da den er basert på ens unike profil og aktivitet.

Disse to delene henger sammen i forståelsen hvor KI-modellen tolker informasjonen den har for å lage meningsfulle forslag. Den bruker komplekse beregninger for å analysere informasjonen og bestemme hvilke forslag som mest sannsynlig vil være relevante for brukeren. I denne sammenhengen fungerer KI-modellen som et system som håndterer og tolker informasjon, noe som for så vidt er en sentral del av Luhmanns

kommunikasjonsbegrep. Det må merke seg at jeg har tolket KI-modellen dithen at det er et eget system som på lik linje med psykiske og fysiske systemer benytter seg av kommunikasjonssystemet. Dette passer og godt overens med Esposito (2022) sin tolkning, og er ennå et argument for hvorfor begrepet «artificial communication» er fruktbart.

Jeg har her lagt frem tre gode grunner til at Esposito (2022) gjør en glimrende analyse fundert i Luhmann om at noen kunstig intelligenssystemer kan forstås isteden som «artificial communication». Dette kan virke som motstridende til tidligere bemerkninger om kunstig intelligens som fenomen. Frem til dette punktet er Esposito (2022) sin analyse svært fruktbar, men som vil komme frem, er den mangelfull nettopp med fremveksten av store språkmodeller. Esposito sin analyse bygger på at KI-systemer kun er algoritmer, noe de tre foregående KI-modellen i all hovedsak er.

For å analysere den siste typen i min typologi vil jeg derfor i kommende avsnitt legge ut om GPT-4 som jeg argumenterer for er en selv-lært ikke-tenkende kunstig intelligensmodell. Som beskrevet av OpenAI (2023) er GPT-4 designet for å generere tekst basert på brukerens input, noe som innebærer en form for meddelelse. Videre produserer GPT-4 en meddelelse basert på de kodene og mønstrene den har lært gjennom sin trening, noe som er i tråd med Luhmanns ide om at sosiale systemer opererer på grunnlag av sine egne interne koder og skiller. I henhold til ideen om informasjon, kan en argumenere for at GPT-4 kan håndtere informasjon, siden den er bygd på grunnlag av en stor mengde data som den har blitt trent på. Således er informasjonen GPT-4 gir, ikke bare en kopi av tidligere innlært informasjon, men genereres fortløpende av GPT-4 basert på brukerinputen og de interne kodene og reglene som den har lært.

Det kanskje mest utfordrende punktet å argumentere for, jamfør GPT-4, er forståelse. Siden GPT-4 ikke har egen forståelse eller bevissthet i menneskelig forstand, er det desto vanskeligere å argumentere for at GPT-4 har en egen forståelse av forskjellen mellom meddelelse og informasjon. Det er her begrepet mitt om ikke-tenkende spiller inn for å beskrive forskjellen mellom mennesker og kunstig intelligenssystemer. Som nevnt, siden mennesker har følelser kan de og tenke, dette står i motsetning til KI-systemer. Derimot, gjennom nettopp å kopiere menneskelig væremåte basert på menneskelig forståelse skaper ikke-tenkende KI-systemer en illusjon om at de kan tenke. Denne illusjonen blir bare bedre og bedre og må derfor adresseres. I skrivende stund er det forskning på å forstå/lære opp, store språkmodeller i refleksjon, som og er en menneskelig egenskap (Shinn et al. 2023).

En interessant vinkling med dette aspektet er at det kan argumenteres for at nettopp gjennom å kopiere menneskelige væremåter dukker emergente egenskaper opp i store språkmodeller. Om dette stemmer får tiden vise, men det er alikevel interessant nettopp for forstå KI-systemer i sosiologisk kontekst. Forøvrig, viktigheten i en slik argumentasjon, ligger i å hverken se på KI-systemer som menneskelige (siden de kun prøver å kopiere vår væremåte) og heller ikke neglisjere nettopp det *at* de kopierer vår væremåte. En mulig vei videre, vil være å legge til hos Luhmann et system som inkluderer denne sistnevnte men ekskluderer de foregående. Et mulig navn for et slikt system kan være tuftet på typologien min: «Selv-lært ikke-tenkende system». Dette vil supplere Luhmanns gjeldende struktur: Det psykiske, det fysiske og kommunikasjonssystemet. Dette er tenkt på samme måte som Esposito (2022) supplerer med hennes begrep «artificial communication».

Videre, har GPT-4 forståelse slik mennesker har? Her igjen dukker en utfordring opp, hva menes med «å forstå» noe? Hvis vi ser forståelse i en bred forstand, som evnen til å behandle og reagere på innputt på en måte som er i tråd med de interne kodene og reglene en har lært, kan det argumenteres at GPT-4 har en form for «forståelse». Dette på tross av at uten egen autonomi og evne til selvbevissthet er ikke GPT-4 et fullverdig sosialt system, istedenfor er GPT-4 et ikke-tenkende KI-system som skaper illusjon av å fungere på lik linje med sosiale systemer. Dette begrenser forøvrig ikke dens evne til å delta i kommunikasjon, noe den blir bedre og bedre til ettersom teknologien utvikles. GPT-4 anses derfor som en deltaker i kommunikasjonsprosesser, som en enhet som genererer og reagerer på kommunikasjon på grunnlag av sine egne interne koder og skiller.

Hva så med produksjon og vedlikehold av mening? Siden GPT-4 genererer tekst basert på statistiske mønstre snarere enn noen form for subjektiv forståelse eller tolkning, har den vel ingen egen opplevelse av mening? Mening forstått i smal forstand som en som ytrer sin ide om hva ytringen skal handle om, er ikke det Luhmann legger opp til. Derimot vil jeg argumentere for at Luhmann legger opp til å forstå mening i vid forstand: Som noe som blir skapt i kommunikasjonsprosessen selv. Jeg argumenterer derfor for at GPT-4 bidrar til produksjonen av denne sistnevnte type mening fordi, GPT-4 genererer tekst som kan tolkes og gis mening av brukerne. På denne måten kan det sies at GPT-4 er involvert i produksjonen og vedlikehold av mening i kommunikasjonsprosesser. I denne tolkningen krever det dermed kun tilstedeværelse av et sosialt system for produksjon og vedlikehold av mening i kommunikasjonsprosesser. Alle andre, om det være seg mange eller få, kan være av selv-lært ikke-tenkende KI-systemer (eller andre systemer) så frem at et sosialt system deltar.

I dette delkapittelet har jeg argumentert for hvordan kunstig intelligensmodeller kan inngå i Luhmann sitt kommunikasjonsbegrep. Jeg har benyttet meg av typologien min for nettopp å analysere KI-systemers innpass, hvorav de tre første: «Veiledet ikke-tenkende», «ikke-veiledet ikke-tenkende» og «semi-veiledet ikke-tenkende» har god innpass i Esposito sin teori om «articial communication». Jeg har videre argumentert for hvorfor den siste typen: «Selv-lært ikke-tenkende» er basert på menneskelig-væremåte og at dens funksjon er basert på illusjon av det foregående. Det som her kommer frem, er at jeg ikke forkaster Esposito sin ide men heller legger til en nødvendig nyansering av KI-modeller. På denne måten kan en si at Esposito sitt begrep om «artificial communication» fungerer bra om en kun snakker om algoritmer tuftet på typene: «Veiledet ikke-tenkende», «ikke-veiledet ikke-tenkende» og «semi-veiledet ikke-tenkende». Esposito sitt begrep er derimot ikke anvenlig når en ser på avanserte KI-systemer som er tuftet på «selv-lært ikke-tenkende». En todeling av kunstig intelligensfeltet, i Luhmannsk kontekst, bør få med seg denne innsikten. Jeg foreslår, jamfør Esposito sitt begrep, følgende todeling: «Kunstig kommunikasjon» og «kunstig illusjon». Førstnevnte kategori omfatter det Esposito kaller for «artificial communication» samt mine invendinger, mens sistnevnte omfatter «selv-lært ikke-tenkende KI-systemer».

6.3 Avsluttende argumentasjon av analysen

Med bakgrunn i Luhmann og Habermas sine kommunikasjonsbegreper har jeg analysert ulike typer kunstig intelligensmodeller, som «spamfilter», «kundepersonlighet», «personer du kanskje kjenner» og «GPT-4». Hvilken av disse to tilnærmingen til kommunikasjon er mest fruktbart for å forstå kunstig intelligens som fenomen? Begge tilnærmingene har positive og negative sider ved seg. Som det kanskje kommer frem, har det dog vært mest spennende å se på KI-systemer opp mot Luhmann. Dette av to grunner: 1. fordi det allerede ligger et solid stykke arbeid inn i hvordan å tolke Luhmann opp mot kunstig intelligens som fenomen (hos Esposito), og 2. fordi det er særlig hos Luhmann at mitt bidrag har vært størst. Dette utelukker ikke at Habermas kan vise seg å være mer nyttig ettersom KI-modellene bare blir bedre og bedre. En kan videre spørre seg: Når KI-modeller er såpass avanserte at de til og med klarer å skape illusjon av Habermasiansk kommunikasjon, nærmer vi oss da generell kunstig intelligens?

På den ene siden, innebærer Habermas sin teori talehandlinger og gyldighetskrav, og viser at selv om KI-modeller kan simulere menneskelig kommunikasjon til en viss grad, klarer de ikke å oppfylle de samme kravene som stilles til mennesker. Veiledet og ikke-veiledet KI kan

utføre bestemte typer talehandlinger, men mangler subjektiv oppfatning og kan derfor ikke utføre alle talehandlinger. Det er kun selv-lært ikke-tenkende KI-modeller, som GPT-4, som kan simulere forskjellige typer talehandlinger og generere normativt korrekte svar. Dog kan den ikke bekrefte eller verifisere sannheten av sin egen informasjon. Dette fremhever viktigheten av begrepet «ikke-tenkende» i forståelsen av KIs rolle i kommunikasjon.

På den andre siden, i Luhmanns kommunikasjonsbegrep, som tar en mer overordnet tilnærming, ser en på kommunikasjonens bestanddeler: Meddelelse, informasjon og forståelse av de foregående. I denne konteksten fungerer veiledet KI som spamfilter ved å analysere meldinger og «forstå» om de er spam. Ikke-veiledet KI, som «kundepersonlighet», forstår og kommuniserer med målkunder ved å matche deres interesser og behov. Semi-veiledet KI, som «personer du kanskje kjenner», foreslår nye venner basert på brukerens eksisterende nettverk og aktiviteter. Selv om GPT-4 ikke har menneskelig forståelse, kan den håndtere og generere informasjon basert på brukerinput og de interne reglene den har lært. Avslutningsvis, til tross for forskjellene mellom menneskelig kommunikasjon og KI-modeller, kan KI-modeller skape en illusjon av tenking gjennom å kopiere menneskelig væremåte, noe som kan være en av forklaringene på det at emergente egenskaper oppstår. En slik tilnærming kan være nyttig for å supplere for eksempel Luhmanns teori med nye systemer, slik som det «selv-lærte ikke-tenkende systemet».

7 Konklusjon og veier videre

I denne førstudien har jeg forsøkt å belyse begrepet «ikke-tenkende» i konteksten av kunstig intelligens. Jeg utviklet en ny typologi basert på ulike tilnærminger til kunstig intelligens, i et forsøk på å forstå KI som systemer som ikke tenker. Målet var ikke å adoptere en positiv eller negativ holdning til KI, men heller å oppnå en større forståelse for kompleksiteten i feltet. Gjennom en litteraturstudie og analyse av forskjellige synspunkter, inkludert Swedberg, Habermas, Luhmann, Ongsulee, Sarker, Esposito, Liu og Lidskog, har jeg forsøkt å dykke dypere inn i spørsmål om kunstig intelligens rolle i kommunikasjon og KI-modellers evne til å «tenke» på en lignende måte som mennesker. Min analyse viser at det er en betydelig variasjon mellom forskjellige modeller av kunstig intelligenssystemer: Veiledet ikke-tenkende, ikke-veiledet ikke-tenkende, semi-veiledet ikke-tenkende, og selv-lært ikke-tenkende KI-modeller. De tre første er noenlunde like, men sistnevnte skiller seg ut. Dette fordi selv-lært ikke-tenkende KI-modeller utviser emergente egenskaper, etterligner

menneskelig væremåte, og skaper en illusjon av menneskelighet. Dette står i kontrast til de andre modellene som ikke skaper en lignende illusjon. Videre, kommer dette av at i møte med Habermas og Luhmanns kommunikasjonsteorier, viser det seg at de tre første KI-modellene ikke møter kravene for deltakelse i kommunikasjon, mens selv-lært ikke-tenkende KI-systemer kan betraktes fra et annet perspektiv. Dette fordi KI-systemer, slik som GPT-4, etterligner menneskelig væremåte på en overbevisende måte, noe som skaper diskusjon om deres potensial for å delta i kommunikasjon på en meningsfylt måte. Til slutt må jeg understreke at selv om selv-lært ikke-tenkende KI-systemer kan skape en illusjon av menneskelighet, betyr det ikke at vi skal eller bør antropomorfisere KI-modeller. Vi må tvert imot forstå at slike modeller er fundamentalt forskjellige fra oss mennesker. Dette krever derfor nye tilnærminger og analyser for å virkelig forstå deres kapabiliteter og begrensninger.

Denne fremsitllingen er jeg ikke alene om. Det er tydelig at stemmer innenfor feltet av kunstig intelligens fremhever behovet for å unngå nettopp en slik antropomorfisering. En av de fremste forkjemperne for dette synet er OpenAI sin administrerende direktør: Sam Altman. I dialog mellom Fridman og Altman kommer dette tydelig frem (Fridman, 2023). I podkasten argumenterer Altman sterkt for at GPT-4 skal forstås som et verktøy og ikke som noe menneskelig. Kanskje begrepet mitt ikke-tenkende kan passe inn her for å forklare hvorfor det virker som KI-modellen er menneskelig? Dette kan være god grobunn for videre diskusjon av fenomenet KI.

En annet spennende fenomen som jeg oppdaget for sent, mens som kan brukes til å forklare begrepet mitt om ikke-tenkende KI-systemer er å finne i Kosinski sin artikkel: «Theory of Mind May Have Spontaneously Emerged in Large Language Models» (2023). Denne artikkelen får frem hvordan emergente egenskaper i store språkmodeller har ført til at disse har det som kalles «ToM» (eller «theory of mind»). Som Kosinski skriver: «... humans do not merely respond to observable cues, but also automatically and effortlessly track others' unobservable mental states: their knowledge, intentions, beliefs, and desires» (2023, s. 2). Dette innebærer altså at gjennom emergente egenskaper som store språkmodeller har tilegnet seg, har de og blitt kapable til å forstå mentale oppfatninger tilsvarende nivået til en 7 åring. GPT-4 klarer faktisk å overgå dette nivået (Kosinski, 2023, s. 10).

9. mars 2023 presenterte Tristan Harris og Aza Raskin fra «Center for Humane Technology» et konsept de kaller for «GLLMM (**G**enerative **L**arge **L**anguage **M**ulti-Modal **M**odel)», også kalt «Gollem-class» kunstig intelligenssystemer (Harris og Raskin, 2023). Disse KI-

systemene argumenterer Harris og Raskin for at kan benytte seg av de tre komponentene bilde, tekst og lyd og gjøre om fra den ene til den andre uavhengig hvorhen en starter. For eksempel, om jeg har et lydklipp men jeg skulle gjerne ønske å få det til tekst: Gollem-class KI-systemer vil kunne gjøre dette. Det samme gjelder for om jeg har en tekst og ønsker å se det visuelt, KI-systemene vil kunne lage en svært god representasjon. En mulig hovedsstudie basert på den førstudien jeg her har gjennomført kunne vært å se på nettopp «Gollem-class» kunstig intelligenssystemer og hvordan de benytter illusjonen av å tenke (altså ikke-tenkende) til å utfordrer gjeldende samfunnsstrukturer. Dette kunne vært bunnet i andre sosiologer slik som Bourdieu eller Foucault, med bakgrunn i forståelsen av slike KI-systemer som nettopp ikke-tenkende.

Et godt sted for å studere KI-modellers innvirkning på samfunn, er å finne hos «Center for AI Safety» (forkortet til CAIS) (2023a). CAIS har lagd en liste over åtte forskjellige problemområder hvor ikke-tenkende kunstig intelligensmodeller kan vise seg å føre til store ødeleggelser. I tillegg kom det ut for få dager siden en «Uttalelse om Risiko knyttet til Kunstig Intelligens» (CAIS, 2023b). Denne uttalelsen har flere av de ledende KI-forskerne signert, både ledende figurer fra Google DeepMind, OpenAI, Anthropic, og flere forskere fra diverse universiteter. Jeg ønsker å avslutte denne oppgaven med et sitat fra Harris og Raskin: «*Nukes don't make stronger nukes, but AI makes stronger AI*» (Harris og Raskin, 2023).

8 Referanseliste

Abdulrahman, S. H., & Salim, M. (2022). Using Decision Tree Algorithms in Detecting Spam Emails Written in Malay: A Comparison Study. ITM Web of Conferences, 42, 1001. https://doi.org/10.1051/itmconf/20224201001

Abramson, A. (2023). ChatGPT as a learning tool. American Psychological Association. Hentet 10. april, fra https://www.apa.org/monitor/2023/06/chatgpt-learning-tool

Albrigtsen, Y. (2023). Kunstig intelligens utfordrer kritisk tenkning og demokrati. Utdanningsnytt. Hentet 10. mars, fra https://www.utdanningsnytt.no/chatgpt-kunstig-intelligens-teknologi/kunstig-intelligens-utfordrer-kritisk-tenkning-og-demokrati/351704

Alkhat, I., og Khatib, Al. B. (2016). Filtering SPAM Using Several Stages Neural Networks. IRECOS. https://doi.org/10.15866/irecos.v11i2.8269

Al-Fedaghi, S. (2012). A Conceptual Foundation for the Shannon-Weaver Model of Communication. International Journal of Soft Computing, 7(1), 12–19. https://doi.org/10.3923/ijscomp.2012.12.19

Baraldi, C., Corsi, G., Esposito, E., & Universität Bielefeld. (2021). *Unlocking Luhmann : a keyword introduction to systems theory* (1st ed.). transcript Verlag.

Bostrom, N. (2014). *Superintelligence : paths, dangers, strategies* (First edition.). Oxford University Press.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712. https://arxiv.org/abs/2303.12712

CAIS. (2023a). AI Risk. Hentet 30. Mai, 2023, fra https://www.safe.ai/ai-risk

CAIS. (2023b). Statement on AI Risk. Hentet 30. Mai, 2023, fra https://www.safe.ai/statement-on-ai-risk

Chess Journal. (u.å.). Shannon Number: What is the Shannon Number in Chess? Hentet 19. mai, 2023, frahttps://www.chessjournal.com/shannon-number

Copeland, B. J. (2000, November 1). The Turing Test* - Minds and Machines. SpringerLink. https://doi.org/10.1023/A:1011285919106

Cuzzolin, F., Morelli, A., Cîrstea, B., & Sahakian, B. J. (2020). Knowing me, knowing you: theory of mind in AI. *Psychological Medicine, 50*(7), 1057–1061. https://doi.org/10.1017/S0033291720000835

DeepMind. (u.å.). AlphaGo - The story so far. Hentet 26. mai, 2023, fra https://www.deepmind.com/research/highlighted-research/alphago

Det Norske Akademis Ordbok. (2023). Spam. Hentet 18. mai, 2023, fra https://naob.no/ordbok/spam

Esposito, E. (2022). Artificial communication : how algorithms produce social intelligence. The MIT Press.

Facebook. (u.å.). Personer du kanskje kjenner. I Facebook Hjelpesentral. Hentet 15. mai, fra https://www.facebook.com/help/163810437015615/

Fridman, L. (Vert). (2023, 25. mars). Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the Future of AI (Nr. 367) [Video]. Lex Fridman Podcast. YouTube. https://www.youtube.com/watch?v=L_Guz73e6fw

Google. (u.å.). What is CAPTCHA?. Hentet 22. april, 2023, fra https://support.google.com/a/answer/1217728

Grallet, G., & Pons, H. (2023, 11. mai). Yuval Harari Sapiens versus Yann Le Cun: Meta on artificial intelligence. LePoint. https://www.lepoint.fr/sciences-nature/yuval-harari-sapiens-versus-yann-le-cun-meta-on-artificial-intelligence-11-05-2023-2519782_1924.php

Habermas, J. (1984). The theory of communicative action : 1 : Reason and the rationalization of society (Vol. 1, pp. xlii, 465). Heinemann.

Habermas, J. (1987). The theory of communicative action : 2 : Lifeworld and system : a critique of functionalist reason (Vol. 2, p. 457). Beacon Press.

Habermas, J. (1999) Kraften i de bedre argumenter. Oversatt av Eriksen, A. Basis. Oslo: Ad notam Gyldendal.

Harris, T., & Raskin, A. (2023, 5. april). The A.I. Dilemma - March 9, 2023 [Video]. YouTube. https://www.youtube.com/watch?v=xoVJKj8lcNQ

Harste, G. (2021). The Habermas-Luhmann Debate. Columbia University Press. http://www.jstor.org/stable/10.7312/hars15914

High-Level Expert Group AI. (2018). COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS i Coordinated Plan on Artificial Intelligence. Hentet 15. mai, fra https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52018DC0795&rid=3#:~:

Hystad, J., & Fanghol, T. A. (2023). NTNU forbyr bruk av ChatGPT på eksamen: Nødvendig å nevne det eksplisitt. Khrono. Hentet 15. mai 2023, fra https://khrono.no/ntnu-forbyr-bruk-av-chatgpt-pa-eksamen-nodvendig-a-nevne-det-eksplisitt/779389

IBM. (u.å.a). Artificial Intelligence. Hentet 10. mai 2023, fra https://www.ibm.com/topics/artificial-intelligence

IBM. (u.å.b.). Supervised learning. Hentet 11. mai 2023, fra https://www.ibm.com/topics/supervised-learning

IBM. (u.å.c.). Unsupervised learning. Hentet 11. mai 2023, fra https://www.ibm.com/topics/unsupervised-learning

Janiesch, C., Zschech, P., og Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, 31(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

Kim, D., Choi, Y., & Kim, Y. (2019). Understanding and Improving Virtual Adversarial Training. https://doi.org/10.48550/arxiv.1909.06737

Kosinski, M. (2023). Theory of Mind May Have Spontaneously Emerged in Large Language Models. https://doi.org/10.48550/arxiv.2302.02083

Liu, Z. (2021). Sociological perspectives on artificial intelligence: A typological reading. Sociology Compass, 15(3), n/a–n/a. https://doi.org/10.1111/soc4.12851

Luhmann, N. (1995). Social systems (pp. LII, 627). Stanford University Press.

Michaud, E. J., Liu, Z., Girit, U., & Tegmark, M. (2023). The Quantization Model of Neural Scaling. https://doi.org/10.48550/arxiv.2303.13506

Moyer, C. (2016, 14. mars.). How Google's AlphaGo Beat a Go World Champion: Inside a man-versus-machine showdown. The Atlantic. https://www.theatlantic.com/technology/archive/2016/03/the-invisible-opponent/475611/x

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. Neurocomputing (Amsterdam), 2(5), 183–197. https://doi.org/10.1016/0925-2312(91)90023-5

Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/ethics-ai/

NRK. (2023). Italia forbyr ChatGPT. Hentet 2. april, fra https://www.nrk.no/nyheter/italia-forbyr-chatgpt-1.16360400

O'Connor, T. (2020). Emergent Properties. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Winter 2020 Edition). Hentet 10. mai, fra https://plato.stanford.edu/entries/properties-emergent/

Ongsulee, P. (2017). Artificial intelligence, machine learning and deep learning. 2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE), 1–6. https://doi.org/10.1109/ICTKE.2017.8259629

OpenAI. (2023). ChatGPT: OpenAI's Language Model [Blogginnlegg]. Hentet 16. mai, fra https://openai.com/blog/chatgpt

Pasupa, K. og Sunhem, W. (2016). A comparison between shallow and deep architecture classifiers on small dataset. 1-6. https://doi.org/10.1109/ICITEED.2016.7863293

Lidskog, R. (2020). Samhället utmanat? Sociologisk forskning, 57(2). https://doi.org/10.37062/sf.57.19591

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3), 160–160. https://doi.org/10.1007/s42979-021-00592-x

Sawyer, R. Keit. (2001). Emergence in Sociology: Contemporary Philosophy of Mind and Some Implications for Sociological Theory1. The American Journal of Sociology, 107(3), 551–585. https://doi.org/10.1086/338780

Shannon, C. E. (1950). Programming a Computer for Playing Chess. Philosophical Magazine, 41(314), ss. 1-18.

Sharmin, T., Di Troia, F., Potika, K., & Stamp, M. (2020). Convolutional neural networks for image spam detection. Information Security Journal., 29(3), 103–117. https://doi.org/10.1080/19393555.2020.1722867

Sheikh, H., Prins, C., og Schrijvers, E. (2023). Mission AI. Springer International Publishing AG. https://doi.org/10.1007/978-3-031-21448-6

Shinn, N., Cassano, F., Labash, B., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. https://doi.org/10.48550/arXiv.2303.11366

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science (American Association for the Advancement of Science)*, *362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

Swedberg, R. (2014). *The art of social theory* (Course Book.). Princeton University Press.

Tan, H. (2017). A brief history and technical review of the expert system research. *IOP Conference Series. Materials Science and Engineering, 242*(1), 12111. https://doi.org/10.1088/1757-899X/242/1/012111

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Knopf.

Tidemann, A. (2022). Maskinlæring. Store norske leksikon. https://snl.no/maskinlæring

Turri, V. (2022). *What is Explainable AI?* Software Engineering Institute. Carnegie Mellon University. Hentet 17. april, fra https://insights.sei.cmu.edu/blog/what-is-explainable-ai/

Veiden, P. (1998). *Sosiologisk fantasi : essays* (p. 232). Ad notam Gyldendal.Aakvaag, G. C. (2008) Moderne sosiologisk teori. Oslo: Abstrakt forl.

Vincent, J. (2016, 24. mars). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. The Verge. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Wang, W., & Zhou, Z.-H. (2007). Analyzing Co-training Style Algorithms. Machine Learning: ECML 2007, 454–465. https://doi.org/10.1007/978-3-540-74958-5_42

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. arXiv preprint arXiv:2206.07682. https://doi.org/10.48550/arXiv.2206.0768

Zuboff, S. (2019). The age of surveillance capitalism : the fight for the future at the new frontier of power. Profile Books PublicAffairs.