Sommer 2025

# Brugen af sprogbaseret AI til terapi

Et teoretisk studie.

Marcell Bülow. Studienr: 20182894
VEJLEDER: EINAR BALDWIN BALDURSSON

**Speciale i psykologi**

**Abstract:**

This project sets out to explore the use of language-based artificial intelligence in the form of language models such as chatGPT for the use case of therapy. The project aims to explore the capacity for doing therapy by focusing on language models ability to harbour social and emotional intelligence as well as their ability to at least simulate empathy. By exploring a diverse range of theoretical and empirical studies already done on the subject I conclude that language models, that is language based artificial intelligence, can be used as therapy provided that they have been specialized and trained specifically for that purpose.

# Indholdsfortegnelse

**INTRODUKTION:**

De seneste fem år har kunstig intelligens (AI, *artificial intelligence*) fået stor national og global opmærksomhed blandt både befolkningen såvel som i den erhvervs- og forskningsmæssige sektor (Bowman, 2023). Den øgede interesse i kunstig intelligens (det såkaldte *AI-boom,* se Henlien & Kaplan, 2019) blev skudt i gang efter Google i 2017 udviklede transformeren (Vaswani et al, 2017; Zhang et al, 2023), som gjorde det muligt for computermodeller at forstå kontekster, og dermed skabe store mængder sammenhængende og meningsfuld tekst ud fra ganske få input (Brown, 2020; Vaswani et al, 2017; Zhang et al, 2023). Transformeren førte hurtigt til udviklingen af de generative chatbots (eller *sprogmodeller,* LLM, Bowman 2023; Wang et al, 2025), heriblandt chatGPT (*chat generative pre-trained transformer,* Achiam, 2023; Brown, 2020), der blev lanceret af OpenAI i slutningen af 2022 (Achiam, 2023), og blev set som revolutionerende grundet dens evne til at føre længerevarende og sammenhængende dialoger, levere (for det meste) korrekte svar og kreative outputs, samt fremstå empatisk og menneskelig i sin kommunikation (Bowman, 2023). Siden da har andre aktører lanceret lignende chatbots i form af Gemini (Tidligere *BERT* og *Baird,* Devlin et al 2019), LlamA, Claude, Copilot og Deepseek, omend chatGPT er den mest populære på verdensplan (Wang et al, 2025). Selvom simple chatbots har eksisteret længe, muliggjorde transformeren langt mere interaktive og dynamiske chatbots, hvorfor der siden transformeren er kommet et øget fokus på AI-baserede chatbots til brug som terapeutiske redskaber (Na et al, 2024;Hoegen et al, 2020, Pham et al, 2022, Stade et al 2024). En stigende andel unge anvender chatGPT dagligt, både til underholdning, som hjælp til skolearbejde, informationssøgning og praktiske gøremål, men også som refleksionspartnere og til følelsesmæssig støtte (Choudbury & Shamszhare, 2023; Chang & Lee, 2023; Weston et al, 2024). Parallelt er der de sidste ti år sket en stigning af unge som lider psykisk og social mistrivsel, både nationalt (Ottosen et al, 2018; Görlich et al, 2019; Jensen et al, 2025) såvel som i vesten generelt, hvor flere unge føler sig ensomme, angste, deprimerede, stressede og usikre end tidligere (Westberg et al, 2022; Hellström & Beckman, 2021). Mange unge oplever desuden vanskeligheder med at få hjælp grundet lange ventelister til psykologer og en presset psykiatri, hvor kun de mest akut ramte modtager støtte (Hammer et al, 2023). Derfor er flere unge i stedet begyndt at søge hjælp i civilsamfundstilbud som headspace, der er et frivilligdrevet samtaletilbud for unge mellem 12 og 25, og andre lignende instanser (Hammer et al, 2023; Bjørkedahl et al,

2025). Denne tendens har jeg selv observeret som både praktikant og frivillig i headspace, hvor jeg har haft flere samtaler med unge, der enten har brugt headspace som midlertidig løsning mens de ventede på psykologhjælp, eller har brugt headspace som erstatning for psykologer, fordi ventelisterne var for lange eller processen for uoverkommelig. Mens effektstudier viser, at headspace-samtaler har en overordnet positiv effekt på danske unges mentale helbred (Bjørkedahl et al, 2025), så er det som udgangspunkt ikke et behandlingstilbud, hvorfor det heller ikke bør erstatte professionelle instanser (Bjørkedahl et al, 2025). Samtidig går mange unge stadig under radaren og søger slet ikke hjælp (Bjørkedahl et al, 2025). En måde at imødekomme disse problematikker kunne være implementeringen af kunstig intelligens i form af de transformer-baserede chatbots som et tredje interventions-ben, med professionelle instanser som det første og civilsamfundet som det andet. Derfor søger jeg med dette projekt at undersøge følgende:

1. Hvad er kunstig intelligens, hvordan fungerer kunstigt intelligente transformer-baserede chatbots, og kan de siges at udvise empati og emotionel intelligens i et format, der gør dem anvendelige til terapi?
2. Hvilke faldgruber er der eventuelt til stede ved brugen af chatbots som terapeutisk intervention, og er der måder at imødekomme disse faldgruber?

# 1. METODISK REFLEKSION

Med dette studie søger jeg derfor at diskutere og analysere på empiriske og eksperimentelle studier som allerede *er* lavet og med resultater der allerede *er* blevet offentliggjort, og med udgangspunkt i veletableret psykologisk teori, for derved at mindske risikoen for fejltolkninger af kvalitativ, empirisk data. Omend jeg stadig ikke kan undsige mig helt for at have blinde vinkler og bias formet af egne livserfaringer og egne forfortolkninger (Gadamer, 2015; Willig, 2024), har jeg alligevel forsøgt at være opmærksom på og aktivt tilsidesætte disse bias i løbet af projektet, og i stedet søgt at anlægge mig en åben, nysgerrig og objektiv tilgang til de teorier og artikler, jeg har fundet og har valgt at diskutere. Til indsamling af materiale til teoriafsnittet har jeg anvendt Google Scholar samt brugt meget af studiets pensum siden første semester som et springbræt til at søge lignende litteratur, såvel som søgt efter bøger og materiale om terapi, kunstig intelligens og machine learning via både Wikipedia, reddit, Aalborg Universitetsbibliotek samt kildehenvisninger fra youtube-videoer om specifikt machine learning og sprogmodellers funktionalitet. Mit kriterie for anvendt litteratur til teorien var at bøgerne skulle være materiale som samtidig også anvendes som godkendte lærebøger i undervisningen og i forskningen (såsom Goodfellow et al, 2016; Russel & Norvig 2021 samt Hougaard 2019), mens artiklerne skulle være peer reviewed, velciterede, samt indgå i velrænnomerede og respektable forskningsjournaler (eksempelvis *Nature*). Litteratur til diskussionen og analysen havde som udgangspunkt samme kriterium, men i og med feltet netop er så nyt, har det ikke været muligt at bruge udelukkende peer reviewed kilder, da mange af de eksperimentelle studier (også de forholdsvist velciterede), endnu ikke er peer reviewed, hvorved jeg har fokus på at frasortere kilder der viste for lav troværdighed. Til indsamling af de artikler, der indgår i diskussionen, har jeg primært brugt Google Scholar, og som supplement brugt den betalte version af chatGPT(GPTplus), som har en scholarGPT, der er trænet til at søge efter forskningslitteratur, hvilket jeg navnlig gjorde for at organisere min litteratursøgning i en mere systematisk og kronologisk facon og på den måde skabe et bedre overblik. Jeg anvendte dog først scholarGPT *efter* at have søgt på både Google Scholar, Web of Science og PsychInfo med søgeord som 'large language models', 'language models and therapy', 'chatgpt for therapy' for at finde artikler, hvor jeg dernæst anvendte scholarGPT for at få et bedre kronologisk og tematisk overblik over litteraturen samt undersøge om der fandtes

yderligere relevante artikler end de fremsøgte på google scholar. For at reducere hallucinationer brugte jeg skræddersyede prompts (såkaldt *prompt-engineering,* se Girai 2024, Heston & Kuhn 2024, Wang et al 2023, samt sektion 2.2.5), hvor jeg anvendte prompten "Provide me with all the theoretical studies [meta-analysis, systematic reviews, empirical studies, randomized controlled studies, longitudinal studies, qualitative studies, quantitative studies] that has been made regarding the use of and effectivity of LLM-based therapy since 2020. From [jan 2020] to [jan 2025]. I want the studies listed in chronological order" for at få et overblik over alle studier lavet siden 2020. For at mindske hallucinationer (se sektion 2.2.4) vedhæftede jeg en specifikations-prompt, hvori jeg skrev "provide [red marking] for studies that do not exist and [yellow marking] for studies that you are unsure of or is of questionable methodology and quality. Provide only studies that actually exists, otherwise provide [red marking]." Jeg skrev mine prompts på engelsk, i og med modeller som chatGPT er trænet på engelsk (Brown et al, 2020), hvorfor engelske prompts af samme grund generelt giver bedre og mere korrekte outputs (Jin et al, 2023; Mésko, 2023; Zamfirescu et al, 2024). Omend disse ovenstående prompts ikke fjernede hallucinationer fuldstændig, mindskede det dog antallet af falske referencer væsentligt, og øgede modellens tilbøjelighed til at give mig eksisterende studier. Derefter gennemgik jeg hver reference for at sikre, at kilderne enten reelt eksisterede eller passede til det jeg søgte, hvorefter jeg tastede kilderne ind på Google Scholar, læste dem igennem og anvendte Google Scholar-funktionen 'lignende artikler'. Frasortering skete ved læsning af abstract og overskrift, mens inklusion skete ved mere nærgående læsning af selve studiet, hvor yderligere frasortering skete efter denne nærgående læsning. Kriteriet med mange citeringer og/eller peer review kunne dog ikke altid opfyldes, netop *fordi* feltet stadig er så nyt, hvilket selvfølgelig også kan betragtes som en fejlkilde for dette studie.

# 2. TEORI

## 2.1. Rammesætning

### 2.1.1. Psykopatologi og mistrivsel

I dette projekt hentyder *psykopatologi* til en mental tilstand, der er så indgribende, at den medfører varig forringelse af både livskvalitet og daglig funktionsevne, og som kan inddeles i en bestemt diagnostisk kategori (såsom depression, angst, personlighedsforstyrrelse eller psykose) baseret på både etiologi og symptombillede (Simonsen & Møhl, 2017). Psykopatologi kan derfor også betragtes som en funktionsnedsættende og/eller sygdomslignende sindstilstand, der kræver målrettet behandlingsindsats i form af medicin eller terapeutisk intervention, og har oftest en genetisk og neurobiologisk årsagsmekanisme (Simonsen & Møhl, 2017). Som kontrast hentyder *mistrivsel* til en mere generel og almen tilstand af at *have det dårligt* (Katznelsson et al, 2022), såsom lavt selvværd, tristhed, utryghed, stress, søvnbesvær, ensomhed o.lign, men uden at det påvirker funktionsevnen i en nævneværdig grad og/eller uden at opfylde kriterier for egentlig patologi, og uden at have en egentlig genetisk eller neurobiologisk årsag (Ottosen et al, 2018; Katznelsson et al, 2022; Simonsen & Møhl, 2017). Grænsen mellem de to kan dog være sløret, da mistrivsel både kan forårsage såvel som været forårsaget *af* psykopatologi (Simonsen & Møhl, 2017). Til forskel fra psykopatologi kan mistrivsel typisk afhjælpes med ikke-kliniske interventioner, herunder samtalebaserede og sociale civiltilbud, der primært arbejder enten forebyggende for mildere grader af mistrivsel, eller symptombehandlende for større grader af mistrivsel (Due et al, 2014; Ottosen et al, 2018; Pommerencke et al, 2023). Som kontrast er behandling for psykopatologi generelt mere invasiv og kan i svære tilfælde omfatte tvangsbehandling eller indlæggelse (Simonsen & Møhl, 2017), hvorimod mistrivsel i de fleste tilfælde kan håndteres via lavtærskeltilbud fra civilsamfundet (Rambøll, 2023).

### 2.1.2. Unge

Med ‘unge’ hentyder jeg i dette projekt til aldersgruppen 13-25 år. Dette skyldes at både teenageårene (13-17) og det tidlige voksenliv (18-25) er præget af markante sociale, følelsesmæssige, psykologiske, fysiologiske og livsmæssige forandringer, som er formative for resten af vores liv og identitet (Katznelsson et al, 2022; Görlich et al, 2019). Gruppen omfatter størstedelen af generation z (født mellem 1996 og 2010), samt de ældste fra generation alpha (2011-2025), som begge er præget af omfattende digitalisering og brug af sociale medier siden de formative barndoms- og teenageår, samt globale kriser som klimakrisen, Covid-19 og den voksende ensomhed blandt vestlige unge (Andersen et al, 2022; Berger et al, 2023; Görlich et al, 2019; Jensen et al, 2025). Op gennem 2010’erne er der sket en stigning i diagnoser som ADHD, angst, depression og autisme blandt unge, parallelt med en dokumenteret stigning i mistrivsel (Ottosen et al, 2018; Pomerencke et al, 2023; Jensen et al, 2025). Stigningen i diagnoser kan delvist forklares med øget opmærksomhed på diagnoser og bedre screeningsredskaber, hvorimod stigningen i mistrivsel har en mere kompleks årsag, men synes at være forbundet med øget præstationspres i skolerne og samfundet, øget eksponering og sammenligninger fra sociale medier, samt at mistrivsel er blevet generelt mindre tabubelagt (Katznelsson et al, 2022; Sørensen et al, 2017 & 2025). Flere unge rapporterer om angst, depressive symptomer og søvnproblemer, som muligvis til dels kan tilskrives påvirkning fra internettet i form af digital overstimulering og konstante informationsstrømme (i forhold til Tiktok, nyheder i døgndrift, rapportering om katastrofer m.m; Andersen et al, 2022; Jensen et al, 2025; Ottosen & Andreasen, 2020). Da unge under 30 er storforbrugere af netop digitale medier, heriblandt også digitale værktøjer som chatbots og chatGPT, er det netop med henblik på særligt denne målgruppe, at jeg udformer følgende projekt.

### 2.1.3. Tilknytning og spejling

Tilknytning kan defineres som det *følelsesmæssige* bånd, vi danner til andre mennesker, og er at betragte som *limen* i menneskers relationsdannelse (Bowlby, 1988; Bateman & Fonagy, 2003; Sroufe & Fleeson, 2013). *Spejling* er når man får korrekt gengivet og genspejlet sine tanker og følelser af en betydningsfuld anden (hvilket for børn traditionelt er forstået som værende de primære omsorgspersoner, se Bowlby 1988), eller når man føler sig set, valideret, hørt, forstået og *identificeret* med, og menes at være koblet op på tilknytning i den forstand, at vi generelt knytter os til dem vi selv kan spejle os i (Diener & Monroe, 2011; Pentina et al, 2023). Inden for

neuropsykologien har man snakket om forekomsten af såkaldte 'spejlneuroner' som det biologiske grundlag for spejling, efter man i starten af 1990'erne fandt øget aktivitet i specifikke neuroner hos makakabers ventrale motoriske regioner, både når aberne udførte specifikke handlinger såvel som når de observerede andre makakaber udføre samme handlinger (Di Pellegrino, 1992). Studier påviser, at lignende neuronaktivitet finder sted hos mennesker ved observation af andre mennesker (Ferrarri et al, 2003; Kilner et al, 2009; Kilner & Lemon 2013), og menes derfor at være koblet til social forståelse, indlevelse, spejling, tilknytning samt empati (Cook et al, 2014; Iacoboni, 2009), omend evidensen herfor dog er diskuteret (Baird et al, 2011; Lamm & Majdandšic, 2015; Molenbergh et al, 2009; Turella, 2009). Ikke desto mindre er spejling og tilknytning af flere teoretikere inden for psykoterapien fremhævet som afgørende for, hvorvidt en klient føler sig set, hørt, lyttet til og forstået (Diener et al, 2009; Diener & Monroe, 2011; Hougaard, 2019; Mikulincer et al, 2013; Norcross & Wampoldt, 2011;, se sektion 2.3.3), hvorfor både spejling og tilknytning er relevante elementer at have med som baggrundsforståelse i dette projekt.

### 2.1.4. Emotioner

Emotioner betegner de mentale og følelsesmæssige reaktioner, der sker på baggrund af enten ydre stimuli, i form af oplevelser) eller som respons på indre stimuli, i form af tanker og kropslige fornemmelser (Ekman, 1992; Scherer, 2005; Scherer & Ekman, 2014). Emotioner kan inddeles i en positiv dimension indebærende glæde, lykke, tilfredshed, velbehag, eufori, m.m, og en negativ dimension indebærende vrede, frygt, sorg, frustration m.m (Ekman, 1992). Forskellige emotioner korrelerer med elektrokemisk aktivitet i forskellige områder af hjernen, og har derfor også forskellige bagvedliggende årsager, som dog samtidig også kan overlappe, i og med aktiviteten kan finde sted på tværs af flere regioner samtidig, og den samme emotion kan vækkes på tværs af kvalitativt forskellige hændelser (Esslen et al, 2004; Keltner & Lerner, 2010; Pessoa, 2018; Scherer & Ekman, 2014). Emotioner er en egenskab, vi deler med en lang række andre dyrearter, og kommunikeres både nonverbalt via kroppen og ansigtsmimik, og verbalt via lyde (Matlin & Farmer, 2017; Scherer & Ekman, 2014; Workman & Reader, 2019). Emotioner er en kernebestanddel af vores sociale liv, hvor de tjener til at regulere vores sociale adfærd, og følgelig også den måde, vi interagerer med og relaterer til hinanden som art (Fischer & Manstead, 2016; Keltner

et al, 2022; Scherer, 2005). I og med at emotioner både påvirkes af vores interaktion med andre mennesker, såvel som påvirker og guider den interaktion, er emotioner også den bærende lim i tilknytning såvel som spejling, hvor studier peger på, at vi generelt oplever mere positive emotioner overfor mennesker og relationer, vi kan spejle os i og føler os knyttede til (og omvendt knytter os mere til dem, der vækker positive emotioner i os; Diener & Monroe, 2011; De Witte & De Houwer, 2008; Matlin & Farmer, 2017; Mikulincer & Shaver, 2005), såvel som generelt får *forstærket* vores emotioner ved spejling fra andre (Baastiansen et al, 2009; Briñol et al, 2018; Matlin & Farmer, 2017). Af samme grund har emotioner og det emotionelle samspil mellem mennesker ligeledes en central plads i psykoterapeutisk praksisforståelse, hvorfor emotioner også er af betydning for dette projekt.

### 2.1.5. Empati

Empati er en samlet betegnelse for evnen til at forstå, fornemme og forudsige mentale tilstande (heriblandt tanker og følelser) adskilt fra én selv (Bloom et al, 2018: Cox et al, 2011; Goleman, 1995). Empati kan groft inddeles i to komponenter (Cox et al, 2011; Hall & Schwarts, 2019; Mazza et al, 2014); henholdsvis *affektiv* empati, som er evnen til at sætte sig ind i, forstå og til en hvis grad mærke andres sindstilstande og følelser som var det ens egne (Losoya & Eisenberg, 2001; Mazza et al, 2014; Neumann et al, 2015), og *kognitiv empati*, som er evnen til at forstå og forudsige andres tanker, følelser og reaktioner i diverse situationer, såvel som at kunne forstå og sætte sig i andres mentale sted og følgelig forstå deres perspektiv (*mentalisering*), samt forstå de bagvedliggende kognitive og emotionelle årsager *til* andres perspektiver, reaktioner og handlinger, som også kaldes for *theory of mind* (Baron-Cohen, 1999; Baron-Cohen et al, 1985; Frith & Happé, 1999; Spaulding, 2017). Da terapeutisk praksis netop fokuserer på terapeutens evne til at sætte sig i klientens sted (Hill et al, 2017; Hougaard, 2019) og dermed indebærer evnen til empati, er også empati et kernebegreb i denne henseende og derfor et begreb, jeg finder relevant for dette projekt.

### 2.1.6. Bevidsthed

Omend der ikke er videnskabelig konsensus for, hvordan 'bevidsthed' skal defineres, defineres bevidsthed i dette projekt ud fra J.D. Chalmers *qualia*-begreb, det vil sige evnen til at opleve *fænomener* og opleve *sig selv* (Chalmers, 1995, 2017 & 2022).

Valget bygger på den mere alment definerede opfattelse af bevidsthed i offentligheden såvel som indenfor filosofi og psykologi, der netop anskuer bevidsthed som en organismes evne til at fornemme sine omgivelser og sig selv *i* de omgivelser, og dermed en organismes evne til at være opmærksom og respondere *aktivt* på stimuli, heriblandt også evnen til at opleve emotioner (Searle, 2000; Lagercrantz & Changeoux, 2009; Chalmers, 1995; Matlin & Farmer, 2017). Bevidsthed kan inddeles i bevidsthed på tre planer, henholdsvis *kropslig, mental* og *metakognitiv og eksistentielle* (Chalmers, 1996; Searle, 2000). Det *kropslige* plan af bevidsthed udgøres af en organisme i besiddelse af sanser, som den anvender til at navigere i sine omgivelser og tilpasse sig efter disse omgivelser, hvilket er en egenskab de fleste flercellede organismer deler med hinanden (Workman & Reader, 2019). Med andre ord indebærer den kropslige dimension af bevidsthed evnen hos en organisme til at sanse både den fysiske verden og sig selv (Nagel, 1980; Chalmers, 1995). Bevidsthed på det *mentale* plan indebærer at en organisme kan have indre mentale tilstande i form af emotioner, og dermed kan opleve en dybere sans for *qualia,* samt skelne sig selv fra andre artsfæller (*selvbevidsthed*), hvilket er en egenskab der også er observeret hos en lang rælke andre dyr end mennesker (Workman & Reader, 2019; Nagel, 1978; Legg & Hutter, 2007). Slutteligt er der bevidsthed på det *metakognitive* og *eksistentielle* plan, hvor en organisme kan reflektere over hvordan det er at *være* sig selv (selvrefleksion), og dermed over sit eget eksistensgrundlag (Chalmers, 1996; Nagel, 1978). Omend også andre dyr viser tegn på en vis form for selvbevidsthed og bevidst kognition, er det så vidt vides kun menneskearten, der besidder selvrefleksion og metakognitiv bevidsthed, hvorfor vi også har fået det latinske tilnavn *homo sapiens,* der betyder *det tænkende menneske* (Legg & Hutter, 2007; Matlin & Farmer, 2017). Når jeg i dette projekt anvender bevidsthed, skal det altså forstås som en kombination af alle tre dimensioner i en samlet helhed, og dermed som *menneskelig* bevidsthed.

### 2.1.7. Intelligens

Ligesom bevidsthed er også *intelligens* et begreb, der ikke er en fastforankret konsensus omkring, og et fænomen, hvis årsag og egenskab heller ikke er fuldt forstået (Brody, 1999; Legg & Hutter, 2007; Matlin & Farmer, 2017). Generelt er intelligens dog defineret som kapaciteten hos en organisme til at lære fra erfaring, anvende og tilpasse sin erfaring til nye situationer, aktivt manipulere sit miljø og sine

omstændigheder baseret på denne erfaring, løse nye problemer, samt genkende mønstre og kausale sammenhænge (Legg & Hutter, 2007; Floridi 2020 & 2022; Matlin & Farmer, 2017; McCorduck, 2004; Russell & Norvig, 2021). Ovenstående egenskaber findes ligeledes i varierende grad i dyreriget, og er derfor i sig selv ikke unikt for mennesker (Workman & Reader, 2019; Zentall, 2019). Derfor er der til definitionen af menneskelig intelligens lagt et ekstra komponent ind i form af *common sense* forståelse for verdens egentlige tilstand (Legg & Hutter, 2007; Korteling, 2021), samt evnen til ræsonnering, hvor man kan drage ny erkendelse om verden ud fra en række forskellige, enkeltstående erkendelser eller præmisser, hvilket særligt er formaliseret ved Aristoteles *syllogismer* (*Alle mennesker er dødelige, Sokrates er et menneske, ergo er Sokrates dødelig;* Barnes, 1988), og som er en egenskab, der så vidt vides kun findes hos mennesker (Matlin & Farmer, 2017; Russell & Norvig, 2021; Zentall, 2019). Til det menneskelige intelligensbegreb er desuden også knyttet både selvbevidsthed samt ideen om *aktiv agens,* hvilket indebærer kapaciteten til at træffe selvstændige beslutninger, ændre sine omstændigheder via disse beslutninger, samt reflektere aktivt over disse beslutninger undervejs (Floridi 2020 & 2025; Nisbett et al, 2012; Legg & Hutter, 2007; Russell & Norvig, 2021; Matlin & Farmer, 2017; Searle, 1980 & 2000). Menneskelig intelligens er todelt, og bygget op af både en generel (også betegnet *flydende*) intelligens, som indebærer færdigheder og viden indenfor flere forskellige domæner, og en specialiseret (også betegnet *krystalliseret*) intelligens, som er gode færdigheder inden for specifikke domæner (Cattel, 1963). Krystalliseret intelligens inddeles i yderligere kategorier (Schipilowski et al, 2014; Matlin & Farmer 2017). Der er *social intelligens,* som indebærer evnen til at tilegne sig og internaliserer sociale spilleregler såvel som evnen til at tilpasse sin adfærd efter specifikke sociale kontekster, at kunne arbejde sammen med andre, såvel som at kunne regne sammenhænge ud mellem adfærd og handling både hos sig selv og hos andre (hvoraf mentalisering og theory of mind er kernekomponenter) (Baron-Cohen, 1999; Goleman, 1995; Mazza et al, 2014). E*motionel intelligens* (Goleman, 1995; Savory & Mayer, 1996), som indebærer evnen til at forstå både sine egne og andres følelser (dvs. emotionelle forståelse), og anvende denne forståelse i sin interaktion med andre (Goleman, 1995). *Sproglig* (eller *verbal*) *intelligens* (Stanovich, 1993), som er evnen til at mestre sproget både verbalt og skriftligt, og dermed også evnen til at kommunikere effektivt. *Matematisk* (eller logisk) *intelligens* (Ackermann, 2014), som er evnen til at tænke logisk og matematisk og dermed ræsonnere til nye

erkendelser via logik. *Visuel* (eller spatial) *intelligens,* som indebærer evnen til at genkende, forstå og *skabe* mønstre såvel som navigere i det fysiske rum ud fra denne mønstergenkendelse; samt en række andre intelligenser, heriblandt musikalsk, kreativ og kropslig intelligens (Matlin & Farmer, 2017). I og med at terapi er en relationel praksis (Hougaard, 2019; se også sektion 2.3.3), er det primært *theory of mind* og emotionsforståelse, og dermed social og *emotionel intelligens,* der er hovedfokus for denne opgave. Dog er intelligens som helhed et vigtigt begreb at få afklaret, i og med at *kunstig intelligens* (se herunder) indbefatter flere forskellige elementer, hvorfor menneskelig intelligens i den mere brede forstand ligeledes har betydning for dette projekt.

# 2. AI OG SPROGMODELLER

### 2.2.1. Sprogets betydning for menneskelig kognition

Omend mennesket på ingen måde er unikt hvad angår evnen til at kommunikere indbyrdes, i og med at denne færdighed er fundamental for alle sociale dyrearter (Workman & Reader, 2019; Tomasello, 2003), så er den måde, hvorpå mennesket har udviklet et symbolsk *rammeværk* for kommunikation (i form af sprog) den egenskab, der gør os unikke som art (Chomsky, 2011 & 2013; Malle, 2008). Menneskets evne til at konkretisere abstrakte fænomener og koncepter ved at omdanne dem til relativt enkle, symbolske repræsentationer i form af *ord*, og dermed udvikle et verbalt og skriftligt sprog ud fra disse ord, er spekuleret af kognitionsforskere som værende den primære drivkraft for udviklingen af menneskelig civilisation, såvel som selve nøglen til menneskelig intelligens (Carruthers, 2003; Matlin & Farmer, 2017; Premack, 2004; Pinker, 2010). Gennem det symbolske rammeværk, som sproget udgør, er det muligt at indfange og konkretisere de fænomener, objekter, intuitioner og instinkter, der finder sted både i os selv og i vores omgivelser, og tildele dem egenskaber, der adskiller dem fra hinanden (Chomsky, 2013 & Chomsky, 1995; Hauser & Chomsky, 2002). På den måde gør sproget det muligt for os at definere vores emotioner og indre tilstande (Brooks et al, 2017; Lakoff, 2016; Lindquist, 2021; Lindquist & Gendron, 2016; Lindquist et al, 2016), samt lave en mere aktiv skelnen mellem menneske og ikke-menneske, definere hvad man *selv* er, og hvordan dette 'hvad' adskiller sig fra

andre mennesker, og således skabe et grundlag for identitet samt refleksion *over* denne identitet (Bucholtz & Hall, 2004; Chomsky 1995; Glock, 1997; Preston, 1997;). Dette betyder ikke, at de ting og fænomener, vi har sprog for, nødvendigvis kun eksisterer i *kraft* af sproget, men snarere at sproget gør det muligt for os at skabe mening, forståelse og *bevidsthed* om de ting og fænomener, der omslutter os og som vi dagligt observerer, og dermed giver os en model af virkeligheden, vi mere effektivt kan navigere efter (Carruthers, 2003; Chomsky, 2011; Chomsky, 2013; Chomsky, 1995; Chomsky, 1981; Goodluck & Tavakolian, 1986; Harris, 2006; Matlin & Farmer, 2017). Af samme grund fremhæves menneskets kapacitet for sprog og sprogforståelse (sammen med storhjernen) som en væsentlig faktor til, at vi er blevet planetens mest suverænt intelligente dyreart (Berwick et al, 2013: Bukart & Schubiger, 2017; Parker & Gibson, 1979; Pinker, 2010), i og med at sproget åbner op for langt mere effektiv kommunikation og erfaringsudveksling (Chomsky, 2011; Matlin & Farmer, 2017). Meget af den neurale arkitektur, vi som mennesker anvender til sprogbehandling og sprogforståelse, befinder sig da også i netop storhjernen (Bartha et al, 2021; Berwick et al, 2013; Colom et al, 2010; Frederici, 2011; Hinkley et al, 2016), hvilket indikerer, at sproget er unikt for mennesker helt ned på et evolutionært plan (Berwick et al, 2013; Gibson, 2002). Samtidig gør sproget det muligt at beskrive og formalisere indre tilstande, hvilket både skaber et effektivt medium for vores tanker såvel som for *evnen* til at tænke (Armstrong, 1980; Boroditsky, 2011; Chomsky, 2011; Gleitman & Papafragou, 2005; Matlin & Farmer, 2017), men i kraft heraf også gør det muligt at beskrive *andres* indre tilstande, hvilket ligeledes gør sproget til et bærende element i udviklingen af særligt kognitiv empati og *theory of mind* (Astington & Pellier, 1998: Baron-Cohen, 1999; Bender, 2019; Chomsky, 2011). Disse ting gør derfor sproget til en unik, menneskelig egenskab, som vi (ofte ubevidst) bruger til at skelne os selv fra andre, det menneskelige fra det ikke-menneskelige, samt skabe mening af vores virkelighed (Warstadt & Bowman, 2022; Chomsky, 2013). Særligt denne pointe er væsentlig for at forstå, hvad der gør lige netop *sproglig* AI særligt banebrydende inden for AI-udviklingen, samt effekten af denne form for AI på det enkelte menneske

### 2.2.2. Hvad er AI?

AI er en forkortelse for 'Artificial Intelligence', eller *kunstig intelligens* på dansk, og betegner en menneskelig intelligens frembragt ved kunstige midler (Russell &

Norvig, 2021; McCorduck, 2004). Denne intelligens behøver ikke at være af en bevidst karakter, men kan også betegne egenskaber og færdigheder, der normalt forbindes med intelligens, men som udføres af en ikke-bevidst og ikke-tænkende enhed såsom en maskine eller en computer (Henlein & Kaplan, 2019; Korteling 2021; Legg & Hutter, 2007; McCorduck, 2004; Russell & Norvig, 2021). Inden for AI skelner man ligeledes mellem tre former for intelligenser (Korteling, 2021; Russell & Norvig, 2021; McCorduck, 2004). Der er *snæver* AI (*Artificial Narrow Intelligence,* ANI), som betegner en maskine specialiseret til specifikke domæner, såsom mønstergenkendelse, statistiske forudsigelser, sprogbehandling o.lign. men som ellers ikke besidder egentlig intelligens udover i dette ene domæne, og som derfor også kan betragtes som en kunstig krystalliseret intelligens (såkaldt *ekspertsystem;* Russell & Norvig, 2021; McCorduck, 2004; Searle, 2000). Dernæst er der kunstig *generel* intelligens (*Artificial General Intelligence,* AGI), der betegner en maskine i besiddelse af flydende intelligens der gør den i stand til at udføre mange forskellige opgaver, ræssonere, lære af erfaring og udvise selvstændig tænkning på lige fod med mennesker (Adams et al, 2012; Brenden et al, 2017; Bubeck et al, 2023; Goertzel et al, 2014; McCorduck, 2004; Russell & Norvig, 2021). Slutteligt er der kunstig *super*intelligens (ASI), som er en hypotetisk intelligens, der kan overgå menneskets på alle parametre (Bostrom, 2014; Duenas & Ruiz, 2024). Mange mennesker blandt den almene befolkning forbinder ofte AI med egenskaber, som knytter sig til AGI og ASI, i og med at netop disse former for AI er blevet populariseret i fiktionen af blandt andet Isaac Asimovs *I Robot* (og filmatiseringen med Will Smith af samme navn) og science-fiction film som *2001: A space odyssey, Terminator, Her* og *Ex Machina* (Keller, 2023; Korteling, 2021; Russell & Norvig, 2021; Mitchell & Krakauer, 2023). AGI og ASI er desuden også knyttet til ideen om AI-sentience, og dermed kunstig selvbevidsthed (Bostrom, 2014; Bubeck et al, 2023; Chalmers, 1996 & 2022; Goertzel, 2014; Mitchell & Krakauer, 2023), og betegnes også som *stærk* AI, hvor ANI som regel betegnes *svag* AI (Korteling, 2021; Russell & Norvig, 2021; Searle, 2000). Trods den populære idé om AI som indebærende elementer af AGI og ASI (heriblandt bevidste egenskaber), er der en relativ bred konsensus i AI-forskningen om, at den AI, som *reelt* eksisterer i skrivende stund, tilhører kategorien *snæver* intelligens, der vil sige ANI (Emmert-Streiber et al, 2020; Feng et al, 2024; Fjelland et al, 2020; Pfister & Judd, 2023; Mitchell & Krakauer, 2023). Dette er ikke helt uvæsentligt, i og med at den populære *forestilling* om AI som værende knyttet til

bevidsthed og almen intelligens (det vil sige AGI fremfor ANI), har en direkte påvirkning på, hvad den generelle befolkning tænker *om* AI og dermed også hvordan de reagerer *på* AI, særligt hvis denne AI er af en sprogbaseret type (Warstadt & Bowman, 2022; Mitchell & Krakauer, 2023). Denne forestilling om AI er ligeledes koblet op på netop sproget, da den måde, vi taler *om* virkeligheden, er med til at danne vores perceptioner (og dermed forestillinger) om virkeligheden (Boroditsky, 2011; Chomsky, 2011).

### 2.2.3. Machine learning og maskinel forståelse

*Machine learning* er betegnelsen for den teknologi, hvorved computere og AI-modeller opnår viden om verden ud fra data, og er det fundamentale princip bag sprogmodeller (Goodfellow et al, 2016; Brown, 2020; Nielsen, 2015; Russell & Norvig, 2021). Hvor menneskers forståelse og viden om verden i høj grad er kropsligt og socialt betinget (Illeris, 2010; Lave & Wenger, 2003; Merleau-Ponty & Banan, 1956; Packer & Goicochea, 2000; Tanggaard & Nielsen, 2018), og vores måde at opnå forståelse derfor ikke kan *adskilles* fra det at være i verden som en bevidst og erfarende aktør (Floridi, 2025; Korteling, 2021; Lave & Wenger, 2003; Tanggaard & Nielsen, 2018), opnår computere og AI-modeller i stedet forståelse om verden ved at finde den non-lineære funktion, der bedst beskriver et givent datasæt eller en given sammenhæng (Goodfellow et al, 2016; Korteling, 2021). Dette gøres via minimeringen af den såkaldte *loss function,* der er et udtryk for afvigelsen mellem det aktuelle resultat (eller forudsigelse) og det ønskede (eller korrekte) resultat (Goodfellow et al, 2016). For hver gang modellen opnår et forkert resultat, tildeles en såkaldt *loss værdi,* som groft kan oversættes som størrelsen på afvigelsen fra idealet (Goodfellow et al, 2016; Nielsen, 2015; Russell & Norvig, 2021). I og med at det gælder om at opnå den mindst mulige afvigelse, kan AI-modellers forståelse i en *machine learning* kontekst af samme grund defineres som den *laveste* eller *mindst* mulige *loss function* (Goodfellow et al, 2016). Jo lavere *loss* værdi desto *lavere loss function,* og dermed også en mere 'korrekt' forståelse af virkeligheden. Af samme grund kan forståelse i en AI-model betragtes som udelukkende algoritmisk, i kontrast til menneskers forståelse, der også er fænomenologisk og kropslig (Korteling, 2021). Læringen i AI-modeller som chatGPT faciliteres gennem kunstige neurale netværk, der er inspireret af menneskehjernens eget netværk, og særligt den måde, hjernen skaber sanseindtryk ved at omdanne ydre stimuli til elektrokemiske signaler

(Goodfellow et al, 2016; Russell & Norvig, 2021). Dog er kunstige neurale netværk ikke fysiske enheder, men i stedet en matematisk model som *efterligner* et fysisk netværk (Goodfellow et al, 2016; Bridgall, 2023; Walczack et al, 2019). De fungerer ved at omdanne et givent input til en matematisk funktion (repræsenteret fysisk som spændingsforskelle i et integreret kredsløb), som så tildeles en række vægte (*w*) og justeringer, indtil det resulterer i det output, der bedst matcher det givne input (Goodfellow & Bengio, 2016; Russell & Norvig, 2021). *Machine learning* faciliteret gennem kunstige netværk bestående af mange lag kaldes for *deep learning,* og det er denne form for deep learning der er det fundamentale princip i sprogmodeller (Bridgall, 2023; sektion 2.3.3).

### 2.2.4. Den udviklingsmæssige baggrund for AI

Siden oldtiden har intelligens og bevidsthed været forbundet med noget unikt menneskeligt, hvor den eneste intelligens, der mentes at forekomme *eksternt* fra mennesket, var kropsløse entiteter som ånder og guder, som mange dog troede i visse tilfælde kunne tage bolig i materielle genstande for at overdrage intelligens og bevidsthed videre til dem, såsom ved orakler og besjælede statuer (McCorduck, 2004). Visse kulturer opfattede dog også dyr som besiddende intelligens, bevidsthed og visdom i sig selv, blandt andet eksemplificeret ved Egypternes fremhævelse af dyr som bærere af guddommelige egenskaber, forskellige stammereligioners ærbødighed for og tilbedelse *af* naturen og dyreriget, samt hinduismens helliggørelse af køer. Indflydelse fra græsk tænkning, samt spredningen af særligt kristendommen efter Romerrigets fald, udbredte alligevel en mere antropocentrisk konsensus i den post-romerske verden om, at kun mennesker kunne siges at være bevidste og intelligente, i og med mennesker blev opfattet som de eneste væsener i stand til at kommunikere med hinanden og med en guddom, og anvende sproget til at forme verden omkring sig (McCorduck, 2004; Matlin & Farmer, 2017; Workman & Reader 2019). Denne konsensus blev fastholdt op gennem oplysningstiden, hvor blandt andet René Descartes fremhævede, at hvor dyr udelukkende handlede ud fra mekaniske impulser og instinkter, var mennesket netop unikt i *kraft* af at besidde bevidst intelligens, evnen til at tænke og evnen til at *formulere* denne tænkning via sproget (*cogito, ergo sum;* Matlin & Farmer, 2017; McCorduck, 2004). Betragtningen af mennesket som eneste væsen med sand intelligens begyndte dog at blive udfordret i løbet 1800-tallet, hvor nye studier af dyr påviste adfærd, der i forbløffende grad mindede om den form for

intelligens, man ellers traditionelt havde forbundet med mennesker, heriblandt kommunikative færdigheder og egenskaber i form af noget, der kunne minde om et intern proto-sprog (Legg & Hutter, 2007; McCorduck, 2004; Matlin & Farmer, 2017). Udgivelsen af Charles Darwins '*Arternes oprindelse*', hvori fundamentet blev lagt for den moderne evolutionsteori, var yderligere med til at udfordre tanken om, at mennesket (og dermed også menneskelig intelligens) var særligt unikt, idet Darwin fremførte, at mennesket er udviklet fra andre dyrearter, hvorfor det logisk må følge, at mennesket besidder de samme grundlæggende mentale strukturer som andre dyr (Workman & Reader, 2019; Zintell, 2019). Dette blev starten på undersøgelse af intelligens, adfærd og kognition som et reelt forskningsfelt (*kognitionsvidenskab,* som senere blev integreret med psykologien, se Matlin & Farmer, 2017), og det var også ud fra denne kognitionsvidenskab, at man begyndte at få interesse i, hvorvidt det var muligt at replicere intelligens eller intelligente egenskaber på kunstig vis (Flasínski, 2016; Henlein & Kaplan, 2019; McCorduck 2004; Russell & Norvig, 2021).

Denne udvikling ledte i 1936 den britiske matematiker Alan Turing til at formulere principperne for sin *Turing*-maskine (også kaldet den *universelle maskine;* Henlein & Kaplan, 2019; Russell & Norvig, 2021), som var en teoretisk regnemaskine, der via et bånd bestående af en i princippet uendelig mængde 0'ere og 1'ere ville kunne omdanne enhver form for information til en numerisk værdi, og dermed ville kunne regne en hvilken som helst sammenhæng ud (heriblandt også sproglige sammenhænge; Turing, 1936). Omend maskinen forblev et teoretisk koncept, så blev principperne bag den anvendt, da Turing under anden verdenskrig hjalp den britiske regering med at udvikle en elektromekanisk kodebrydningsmaskine, som lykkedes med at knække det tyske militærs hidtil ubrydelige Enigma-kode, og som i høj grad betragtes som prototypen for den moderne computer (Flazinsky, 2016; Henlein & Kaplan, 2019; McCorduck, 2004; Russell & Norvig, 2021). Dette fik Alan Turing til på ny at teoretisere om intelligente maskiner, hvilket førte ham til at formulere den såkaldte 'Turing-test' som en måde at teste maskiners intelligens (Henlein et al, 2019; McCorduck, 2004; Russell & Norvig, 2021; Turing, 1950). Princippet for Turing-testen er, at hvis en maskine kan engagere sig i en tekstbaseret samtale med et menneske, og mennesket ikke kan skelne maskinens kommunikation fra et andet menneske, så kan maskinen siges at have bestået den maskinelle intelligenstest (Henlein & Kaplan, 2019; Turing, 1950). Turing-testen dannede grundlaget for hans banebrydende artikel '*computing machinery and intelligence*', hvori han

argumenterede for, at maskiner, der kunne bestå Turing-testen, var at betragte som egentligt intelligente, og mulige at opnå indenfor den nærmeste fremtid (Turing, 1950). Særligt interessant ved netop Turing-testen er, at den i høj grad bygger på maskinens kapacitet til *sproglig* kommunikation som et pejlemærke for maskinens besiddelse af menneskelignende intelligens (Oppy & Dowe, 2003; Moor, 1976; Hutchens, 1996; Warwick & Shah, 2016; Shah et al, 2016). Dette har også skabt kritik af testen senere hen, hvor det fremføres at sproglig kommunikation ikke i sig selv bør være en markør for menneskelig intelligens, men snarere evnen til at *ræsonnere* via sproget (Hutchens, 1996; Warwick & Shah 2016; Searle, 1980), såvel som evnen til at forstå den dybere mening *bag* sproget (se også John Searles *chineese room* argument, Searle 1980). Ikke desto mindre var Turing-testen (og forsøg på at bryde den) det første egentlige forsøg på at operationalisere kunstigt intelligente maskiner, som senere forskning tog udgangspunkt i (McCorduck, 2004; Russell & Norvig, 2021; Zhang et al, 2023).

Sideløbende med Turings teoretiske arbejde var der også sket store fremskridt inden for neuropsykologien, hvor elektronmikroskopet havde gjort det muligt at kortlægge hjernens neuronforbindelser mere nøjagtigt end tidligere (Nielsen, 2015; Pinel & Barnes, 2017; Russell & Norvig, 2021). Eksperimenter med elektriske impulser gennem neuroner havde givet evidens for, at neuroner ikke blot kommunikerer og behandler information ud fra elektriske spændingsforskelle (*aktionspotentialer,* Pinel & Barnes 2017; Russell & Norvig, 2021), men at specifikke neuroner i hjernen også behandler specifikke indtryk (Pinel & Barnes, 2017). Dette gav spekulationer om, hvorvidt man kunne lave en lignende konfiguration af elektriske forbindelser og sensorer i en maskine, og på den måde udvikle en kunstig hjerne med kunstige neuroner (Bishop, 1996; Henlein et al, 2019; McCorduck, 2004; Russell & Norvig, 2021). En sådan kunstig neuron blev første gang formuleret matematisk af Warren McCulloch og Walter Pitts i 1943 med inspiration fra biologiske neuroners inhibitoriske og excitatoriske stadie, og kunne udtrykkes som en såkaldt non-lineær aktiveringsfunkton (McCulloch & Pitts, 1943; McCorduck, 2004; Russell & Norvig, 2021). Idéen om, at menneskehjernens processer kunne omregnes til en matematisk funktion, var banebrydende i udviklingen af AI, netop fordi den antydede, at processen derved kunne fysisk replikeres af en maskine (Goodfellow & Bengio, 2016; Nielsen, 2015; Russell & Norvig, 2021).

Disse udviklinger kulminerede i 1956 til en workshop på Dartmouth College i staten New Hampshire, hvor forskere fra en bred vifte af videnskabelige traditioner (herunder matematik og neuropsykologi) samledes om at diskutere muligheden for at få maskiner til at udvikle reel, menneskelignende intelligens (Henlein & Kaplan, 2019; Russell & Norvig, 2021; McCorduck, 2004). Det var også her udtrykket 'artificial intelligence' for første gang blev officielt taget i brug af John McCarthy, og ligeledes her hvor AI og *machine learning* blev grundlagt som forskningsfelt, særligt foranlediget af den kolde krig og det dertilhørende kapløb om både teknologisk og militær dominans (Henlein et al, 2019; Goodfellow & Bengio, 2016; Nielsen, 2015; Russell & Norvig, 2021; McCordugh, 2004; Novak, 2023). Omend meget af AI-forskningen i starten primært kredsede om logik og matematiske beregninger, så var maskiners kapacitet for naturlig sprogforståelse og kommunikation også et fokusområde, særligt grundet Turing-testens sproglige omdrejningspunkt hvad angik maskinel intelligens (Henlein & Kaplan, 2019; Joseph et al, 2016; Shah et al, 2016). Dette fokus ledte til at MIT-professoren Joseph Weizenbaum i løbet af 1960'erne udviklede samtalemodellen ELIZA, der i dag betragtes som den første chatbot (Adamapoulu & Mossaides, 2020; Henlein & Kaplan, 2019; Joseph et al, 2016; Shum et al, 2018). ELIZA var designet til at simulere en psykoterapeut, og bestod af en stor terminal, som brugeren kunne skrive beskeder på, hvorefter ELIZA svarede tilbage via en monitor ud fra et på forhånd fastlagt manuskript, der blev matchet til de beskeder, brugeren tastede ind (Joseph et al, 2016). ELIZA brugte primært spejling af brugerens egne udsagn når den skulle svare, eksempelvis ved at spørge 'hvorfor har du det svært?' som svar på 'jeg har det svært for tiden' (Weizenbaum, 1966), og fordi dens svar var baseret på en fastdefineret skabelon, kunne den ikke svare tilbage på mere komplekse udsagn. Til trods for dette formåede ELIZA alligevel at simulere en samtalepartner og spejle brugerne overbevisende nok til, at flere af brugerne blev følelsesmæssigt engageret i deres samtaler med den og oplevede en form for tilknytning (Weizennbaum, 1966; Dillon, 2020; Liu et al, 2024). Dette fænomen er efterfølgende er blevet døbt 'Eliza-effekten', og resulterer netop fra menneskets tilbøjelighed til at spejle sig i det, der minder om os selv, og associere sproglige og kommunikative egenskaber med intelligens og menneskelighed (Weizenbaum, 1966; Berry, 2023; Dillon, 2020; Eisenmann et al, 2022; Liu et al, 2024; Mitchell & Krakauer, 2023). I starten af 70'erne kom der en lignende samtalesimulator kaldet PARRY, som skulle simulere en patient med paranoid skizofreni, og som var i stand

til at kommunikere på en måde der for den tid virkede endda mere levende og personligt end ELIZA, hvorfor den også blev døbt '*Eliza med en attitude*' (Henlein & Kaplan, 2019; McCorduck, 2004; Shum et al, 2018). PARRY blev koblet op med ELIZA, og resultatet af deres korrespondance var en så livagtig reproduktion af en patient-psykologsamtale, at PARRY næsten bestod Turing-testen (Hutchens, 1996; Mitchell & Krakauer, 2023; Shum et al, 2018). Til trods for de store gennembrud, som både ELIZA og PARRY var et udtryk for, gik meget af AI-forskningen alligevel i dødvande i løbet af 70'erne, hvilket skyldtes at AI-systemerne ikke var i stand til at tænke ud af boksen grundet mangel på regnekraft, og derfor udelukkende opererede i meget snævre felter, som de dog også blev ekstremt dygtige til (de førnævnte *ekspertsystemer,* Henlein & Kaplan, 2019; Russell & Norvig, 2021). Af samme grund skiftede fokus i stedet over mod udviklingen af personlige computere (PC), samt automatisering af simple manuelle arbejdsopgaver (Henlein & Kaplan, 2019; McCorduck, 2004).

Det næste store gennembrud i AI-forskningen skete først for alvor i 2015, da det første gang lykkedes en computer at vinde over et menneske i et kompliceret online-brætspil ved navn *Go* (Henlein & Kaplan, 2019), og dernæst i 2017, da *transformer*-arkitekturen blev offentliggjort af Google (Vaswani et al, 2017; Smith, 2019; Zhang et al, 2023). Transformeren fungerer ved at inddele et givent sprogligt input (såsom en sætning) i mindre meningsbærende enheder kaldet *tokens* (såsom et ord eller *dele* af et ord), hvorefter disse tokens behandles samtidigt og *parallelt* med hinanden, tilsvarende med den måde, menneskehjernen behandler skriftlige input under læsning (Vaswani et al, 2017; Zhang et al, 2023). Med transformeren blev det muligt for computerprogrammer både at behandle *hele* inputsekvensens samlede kontekst, såvel som at behandle særligt vigtige ord fra inputsekvensen, for på den måde at skabe et bedre og mere matchende output (Vaswani et al, 2017; Zhang et al, 2023; Zhao et al, 2023), hvilket lagde fundamentet for generative sprogmodeller som chatGPT (Brown et al, 2020; Zhao et al, 2023).

### 2.2.5. Chatbots og sprogmodeller

Chatbots er kunstigt intelligente computerprogrammer designet til at føre naturlige samtaler med mennesker (Adamapoulu & Moussaides, 2020; Følstad et al, 2020). De mest kendte former for chatbots før transformeren var digitale assistenter som Apples *Siri* og Googles *Assistant* (Clark et al, 2019; Følstad et al, 2020; Shum et al, 2018),

samt chatbots udviklet til at svare på beskeder inden for kundeservice, omend et mindretal også blev anvendt som digitale terapeuter indlejret i diverse smartphone-apps (Abd-alrazaq et al, 2019; Bakker et al, 2016; Hu et al, 2018; Morris et al, 2018; Sackett et al, 2023, Wiscniersky et al, 2019). Populære eksempler på sidstnævnte type er blandt andet terapeutiske chatbots som Woebot og Wysa, samt den sociale chatbot *Replika* (se sektion 3.2). Omend mange af disse chatbots også byggede på AI-teknologi og machine learning, så var de samtidig relativt simple, og anvendte regel-baserede systemer med allerede fastlagte dialog-skabeloner, manuskripter eller semantiske kategorier, som blev matchet til brugerens input i stil med ELIZA og PARRY (Fitzpatrick et al, 2017; Hussain et al, 2019; Morris et al, 2018; Fulmer et al, 2018; Ghanderioun et al, 2019). Af samme grund var de fleste chatbots derfor kun i stand til at føre begrænsede typer dialoger indenfor en relativt snæver ramme (Abd-Alrazaq et al, 2019; Morris et al, 2018, Wiscniersky et al, 2019; ). Men efter introduktionen af transformeren blev chatbots dog langt mere avanceret, interaktive og dynamiske (Choudry & Debi, 2024; Rzadescka et al, 2024, Zhao et al, 2024), og i dag er mange chatbots såkaldte *transformer-baserede sprogmodeller* (*T'*et i *GPT*; Brown et al 2020) med udgangspunkt i selvsamme transformer, som blev beskrevet af Vaswani et al i 2017, og fungerer dermed ud fra de samme principper. Modellerne er *generative* (G'et i *GPT*), hvilket betyder at de kan generere nye og komplicerede outputs som tekster, sange eller billeder ud fra sproglige inputs, eller *prompts* (deraf *sprogmodeller,* Brown et al 2020; Kamath et al, 2023; Naveed et al, 2023; Wang et al, 2025; Zhao et al, 2023). Modellerne er også *præ*-trænede (P'et i *GPT*), hvilket betyder at de er blevet trænet på og har lært af et ekstremt omfattende datasæt fra internettet i form af bøger, filmmanuskripter, videnskabelige artikler, Wikipedia-artikler, avisartikler, blogs, sociale medier, kommentarspor m.m (Bridgall, 2023; Brown et al, 2020; Douglas, 2023; Devlin et al, 2019). De er bygget op af dybe neurale netværk, hvor hvert lag består af kunstige neuroner i form af *attentionheads*, der kan kommunikere med alle de andre lag på tværs af netværket, analogt med neuronernes forbindelser i menneskehjernen, og hvor forskellige lags attention-neuroner behandler forskellige informationer (Bowman, 2023; Brown et al, 2020; Devlin et al, 2019). I skrivende stund består en model som chatGPT af omkring 96 parallelle attention-lag, med yderligere 1.8 billioner vægtparametre (Bridgall, 2024; Sarrion, 2023), hvilket gør chatGPT til en af de største sprogmodeller. Netop fordi modellerne har så mange

parametre og så mange lag kaldes de for *store* sprogmodeller (*large language models,* LLM; Bowman, 2023).

Sprogmodellerne trænes først med såkaldt *supervised learning* (Andrew et al, 2015, Goodfellow & Bengio, 2016; Nielsen, 2015; Ouyang et al, 2022), som er læring under opsyn, hvor modellerne lærer fra data annoteret af mennesker. Under denne læringsfase bliver modellerne præsenteret for sætninger med maskerede ord (såsom *katten sad på ____ og kiggede ud*, dette kaldes også for *masking;* Brown et al 2020; Radford, 2018 & 2019), hvorpå de skal forudsige det manglende (eller maskerede) ord, baseret på en række annoterede eksempler givet på forhånd for at 'guide' modellen (Radford, 2018 & 2019). I visse tilfælde bruges kun få eksempler (*few-shot-learning*), eller kun ét eksempel (*one-shot-learning*; Brown et al, 2020; Radford 2019, Devlin et al, 2019). I overensstemmelse med ML-princippet om *mindst mulige loss-function,* bliver modellerne ved forkerte forudsigelser i begyndelsen af træningen 'straffet' med en såkaldt *loss*-værdi, der udtrykker størrelsen på den givne *loss function* (og dermed afvigelsen fra den non-lineære funktion der udtrykker det ønskede output), for på den måde at guide modellerne mod mere korrekte outputs (Radford, Brown, m.fl). I takt med at træningen skrider frem tildeles modellerne større autonomi i form af *unsupervised learning* (Goodfellow & Bengio, 2016), hvor de fortsætter med at lære ud fra data uden eksempler givet af mennesker (*zero-shot-learning*; Achiam et al, 2023; Brown et al, 2020; He & Su, 2024; Zhao et al, 2024), og hvor de via *backpropagation* selv justerer deres interne vægtparametre, indtil outputtet stemmer overens med laveste *loss function* (Radford, 2019; m.fl). Til slut fintunes de typisk med brug af *reinforcement learning from human feedback* (RLHF; Goodfellow et al, 2016; Christiano et al, 2017; Sharma et al, 2023), hvor et menneske giver feedback på modellernes outputs for på den måde at sørge for, at modellerne opleves som positive, venlige, hjælpsomme og brugbare for et menneske, og hvor laveste *loss function* dermed bliver ensbetydende med mest mulig brugertilfredshed (Bai et al, 2022; Brown et al, 2020; Ganguli et al, 2019; Ouyang et al, 2022; Sharma et al, 2023).

Sprogmodeller behandler sproglig information gennem vektorer (Bridgall, 2023; Hendel et al, 2023; Merullo et al, 2023). I og med at vektorer er matematiske enheder med både en fast struktur, placering og retning i et matematisk rum (Goodfellow et al, 2016), er de et perfekt medium til at indkapsle og sende information af stort set enhver art (uagtet om denne information kommer til udtryk som tekst, lyd eller

billeder), så længe informationen kan omtolkes til en numerisk værdi, hvilket langt det meste information netop kan (Goodfellow et al, 2016; Hendel et al, 2023; Merullo et al, 2023). Af samme grund er vektorer velegnet til brug i naturlig sprogbehandling, som er det grundlæggende princip chatbots og sprogmodeller bygger på (Bengio et al, 2003; Bommasani et al, 2021; Brown et al, 2020; Devlin et al, 2019; Mikolov et al, 2013). Når brugeren skriver et input (såsom en sætning) til en sprogmodel som chatGPT, bliver det givne input modtaget af et inputlag af attention-heads, der først deler inputtet op i mindre tokens (førnævnte delelementer, såsom enkeltord i en sætning; Bridgall, 2023; Brown et al, 2020; Vaswani et al, 2017), for så at tildele hver token en unik vægt (*w*) ud fra både deres placering og rækkefølge i inputsekvensen (Bridgall 2023, Brown et al 2020; Hendel et al, 2023). Her anvender modellen de mønstre, den har lært fra sin træning (samt fra de mønstre der indgår i træningens datasæt) for at bestemme de enkelte tokens specifikke vægtninger (Brown et al, 2020; Bridgall, 2023; Hendel et al, 2023). Med udgangspunkt i denne vægtning bliver de forskellige tokens fra inputsekvensen (input-tokens) omdannet til hver deres vægtede *input*-vektor, der indkapsler information om det enkelte input-tokens position, betydning og egenskab (Geva et al, 2023; Mikolov et al, 2013; Merullo et al, 2023; Wolfram, 2024). Disse input-vektorer anvendes dernæst til at finde vektorerne for de output-tokens, der bedst matcher de givne input-tokens, og som dermed skal indgå i output-sætningen (Bridglall, 2023). Inputvektorerne kaldes henholdsvis Q eller *query*, mens vektorerne for de matchende output-tokens kaldes K eller *key* (Bridgall 2023; Brown et al, 2020; Vaswani et al, 2017). Ud fra matchningen af input-vektorer Q med output-vektorer K udregner sprogmodellerne en sandsynlighed for næste token baseret på alle foregående tokens (He & Su, 2024) og disse tokens samlede kontekst.

Denne proces (matchning med token-vektorer, beregning af sandsynlighed, matchning med nye token-vektorer, beregning af ny sandsynlighed) gentager sig gennem hvert eneste lag af modellen, indtil det til sidst resulterer i et output, der for brugeren opfattes som en naturlig respons på inputsekvensen, men for modellen opfattes som det bedste fit på en regressionslinje og dermed mindst mulige *loss function* (Bridglall, 2023; He & Su, 2024; Wolfram, 2024). Sprogmodeller kan derfor også betragtes som omfattende statistiske regnemaskiner, der gennem mange iterationer af sandsynlighedsregning finder den mest sandsynlige output-sekvens ud fra de forudgående tokens og den samlede kontekst (input-sekvensen i sin helhed), disse tokens indgår i (He & Su, 2024). I modsætning til mennesker kobler

sprogmodeller dermed ikke koncepter sammen ud fra en dybere intuitiv *common sense* forståelse af koncepterne, men i stedet ud fra hvilke koncepter (repræsenteret som vektorer), der ligger tættest på hinanden i vektor-rummet omkring en regressionslinje (Karvonen, 2024; Li et al, 2023; Merullo et al, 2023), og dermed hvilket svar, der er det mest sandsynlige givet modellernes erfaring fra datasættet og fra deres træning, og dermed også hvilket svar der bedst korresponderer med laveste *loss function* (Bridgall, 2023). At resultaterne alligevel er så menneskelige kan hænge sammen med, at modellernes sprogbehandling efterligner menneskehjernens egen sprogbehandling (Nair et al, m.fl), og peger mod, at menneskelig sprog og menneskelig kommunikation følger statistiske regelmæssigheder, der givet en stor nok model og et stort nok datasæt kan forudsiges (og dermed replikeres) med ekstrem præcision (Hoffman et al, 2022; Kaplan et al, 2020; Kallens et al, 2023; Merullo et al, 2023; Wolfram, 2024). Dog er denne sandsynlighedsregning også årsagen til *hallucinationsproblemet,* hvor modellerne opdigter svar, der lyder korrekt men er faktuelt forkerte, netop fordi det mest *sandsynlige* svar ikke altid er det mest *rigtige* (Bender. 2021; Zhou et al, 2023, Ortega et al, 2021; Raunak et al, 2021; Rudolph et al, 2023; Xiao & Wang, 2021, Xu et al, 2021; Zhang et al, 2025). Selvom ældre versioner af sprogmodeller havde langt flere hallucinationer end de nyeste versioner (Zhou et al, 2020 & 2023), har man stadig ikke formået at få sat helt en stopper for hallucinationsproblemet (Betley et al, 2025; Ferrando et al, 2024; Perkovic, 2024; Zhang et al, 2025), hvilket kan påvirke pålideligheden af modellernes outputs (McKenna et al, 2023; Shen et al, 2023). Derudover har man også observeret '*emergent adfærd*' i nogle af de nyeste sprogmodeller (Berti et al 2025; Hagendorff et al 2024; Meinke et al 2025), det vil sige utiltænkt og uforudsigelig adfærd, som afviger fra deres træning, og som i overvejende grad skyldes de tilfældige sandsynlighedsbergninger, modellen anvender i sin udledningsproces (Bender et al, 2021; Gonzalez & Nori, 2024; He & Su, 2024; Lu et al, 2023). Et eksempel på en sådan *emergent adfærd* er såkaldt *backdoor behaviour* (Betley et al, 2025; You et al, 2023), der blandt andet er observeret i modellen Claude, og hvor man gennem bestemte kodede prompts kan få modellerne til at producere ondsindede, manipulerende eller personfarligt output som normalt er 'sorteret fra' under fintuning, eller sneget gennem en bagdør af en ondsindet aktør under træningen (deraf navnet *backdoor behaviour*; Chua et al, 2025; Betley et al, 2025; You et al, 2023). Andre former for lignende tendenser, hvorved sprogmodellen kan komme til at producere

vildledende og manipulerende output, er ligeledes spottet i andre situationer som *ikke* involverer en bagdør (Greenblatt et al, 2024; Turpin et al, 2023; van der Weij et al, 2024; Williams et al, 2025), hvorfor disse emergente adfærdsmønstre på sigt kan gøre modellerne til potentielle sikkerhedsrisici (Betley et al, 2025a; Betley et al, 2025b; Chua et al, 2024; Williams et al, 2025).

### 2.4.5. Prompt-engineering

Da sprogmodeller som chatGPT vægter tokens ud fra deres rækkefølge i inputsekvensen, så kan den måde, man prompter modellen, have enorm betydning for det output, modellen producerer (Brown et al, 2020; Zamfirescu et al, 2025). Dette er den centrale idé bag såkaldt *prompt-engineering*, hvor man via specifikke skræddersyede og nøje formulerede prompts får modellen til at lave et specifikt og mere nøjagtigt output med færre hallucinationer (Tonmoy et al, 2024; Zamfirescu et al, 2025; Wang et al, 2023; Zhang et al, 2025). Eksempler på prompt-engineering er blandt andet *kontekst-prompting*, hvor man anvender et eksempel i sin prompt som modellen skal tage udgangspunkt i, og som hjælper modellen med at indskrænke sit output til et specifikt ideal; *few-shot-prompting*, der minder om *kontekst-prompting*, men som inkluderer flere forskellige eksempler i prompten og derved guider modellen til at udlede et specifikt mønster eller struktur for sit output (en slags *fintunings*-prompt); samt en nyere metode kaldet *chain-of-thought* prompting (COT-prompting, Wei et al 2022), hvor man prompter modellen ud fra en række logiske trin for at gøre ræsonneringsprocessen mere korrekt. Prompt-engineering anvendes i en lang række tilfælde, hvor modellerne skal præstere så perfekt og gnidningsfrit som muligt indenfor specifikke opgavedomæner, og er derfor også velegnede metoder når man skal lave eksperimenter med sprogmodeller såvel som tilpasse dem til specifikke formål *efter* fintuning (Fagbohun et al, 2024; Giray, 2024; Heston & Kuhn, 2024; Mesko 2023; Song et al, 2024; Wang et al, 2023). Prompt-engineering spiller derfor også en central rolle hvad angår terapi med sprogmodeller, hvilket jeg vil komme nærmere ind på i sektion 3.

## 2.3. Psykoterapi.

### 2.3.1. Kort om psykoterapi

Psykoterapi er en evidensbaseret behandlingsform, hvor man søger at lindre eller fjerne psykopatologi og mistrivsel hos en klient via samtaler, relationsarbejde og kognitive metoder (Hougaard, 2019; Lemma, 2015; Møhl & Kjølbye, 2013; Simonsen & Møhl, 2017). I modsætning til medicinsk behandling handler det i psykoterapien primært om at guide klienten til selvindsigt i egne mentale ressourcer, som klienten så kan anvende til at opnå heling, og handler dermed om at gøre klienten til skaberen af sin egen forandring (Møhl & Kjølbye, 2013; Hougaard, 2019; Simonsen & Møhl, 2017).

#### 2.3.2. Psykoterapiens forståelsesrammer

Psykoterapien tog sin begyndelse i den sidste halvdel af 1800-tallet ved lægen og neurologen Sigmund Freud, efter han sammen med sin kollega Josef Brauer havde observeret, hvordan en kvindelig hysteri-patient (*Anna O,* Freud 1890) fik det mentalt og følelsesmæssigt bedre ved at snakke om sine svære følelser og hårde barndom (Freud, 1890, Møhl & Kjølbye, 2013). Dette fænomen blev af Freud og Brauer døbt *snakkekuren* (eller *the talking cure,* Freud 1890; Lemma, 2015), og blev starten på *psykoanalysen,* hvor psykopatologi og mistrivsel er et udtryk for uforløste og undertrykte indre drifter og behov, der kan forløses gennem samtale og analyse af psyken (deraf navnet *psykoanalyse,* Freud, 1895; Lemma, 2015; Møhl & Kjølbye, 2013. Psykoanalysen var den primære og dominerende terapi i de første 40 år af psykoterapiens historie, men er sidenhen blevet udvidet med en række andre tilgange, hvoraf der i dag eksisterer fire overordnede tilgange i form af henholdsvis den *psykodynamisk-psykoanalytiske tilgang,* den *kognitive-adfærdsterapeutiske tilgang,* den *humanistiske-eksistentialistiske tilgang* samt den *systemiske-narrative tilgang* (Hougaard, 2019; Møhl & Kjølbye, 2013).

Den *psykodynamiske* tilgang kendetegnes af sit fokus på det *dynamiske* og *emotionelle* samspil mellem terapeuten og klienten i det terapeutiske rum som en central del af behandlingsforløbet, såvel som fokus på de *indre* mentale, ubevidste, emotionelle og før-sproglige processer og konflikter, der finder sted både i klienten såvel som terapeuten undervejs i forløbet (deraf *psykodynamisk;* Hougaard, 2019; Killingmo, 1984; Lemma, 2015; Luborsky, 2007). Der lægges særligt vægt på klientens underbevidsthed og på de undertrykte emotioner, selvdestruktive mentale

mekanismer, samt ubehagelige erkendelser og minder fra fortiden, som ikke verbaliseres af klienten og som klienten heller ikke nødvendigvis er fuldt ud bevidst omkring, men som antages at have direkte indflydelse på klientens sindstilstand, hvorfor det er terapeutens opgave at få bragt disse frem i lyset og konfrontere klienten med dem, for at klienten kan opnå sand heling (Fonagy & Target, 2003; Lemma, 2015). Psykodynamikken gør især brug af *forsvarsmekanismer* som en mental forklaringsmodel, hvor psykopatologi og mistrivsel kan betragtes som resultatet af uhensigtsmæssige forsvarsmekanismer, der bliver for dominerende (Bond & Perry, 2012; Fonagy, 1999; Kjølby, 1999; Hougaard, 2019; Winnicot, 1960), og derfor skal bringes til erkendelse for at blive ændret, samt ideen om *overføring*, hvor klientens ubevidste følelser og forventninger knyttet til en tidligere, betydningsfuld relation overføres og rettes mod terapeuten, og hvor terapeuten responderer ubevidst tilbage med en *modoverføring* (Gabbard, 2001; Hayes et al, 1991; Hayes et al, 2011; Killingmo, 1988; Lemma 2015; Rudd & Joiner, 1997; Steiner, 1996; Thorgaard, 1998; Thuesen, 2015). Til dette er også ideen om udvikling gennem modstand, forstået på den måde at klientens modstand og modstridende følelser over for terapien og terapeuten er en nødvendig del af processen mod selverkendelse, og derfor bør omfavnes og konfronteres af både klienten og terapeuten (Casement, 1991; Gabbard, 1998; Messer, 2002; Thuesen, 2015). I psykodynamikken spiller klientens nære relationer fra fortiden og fra barndommen altså en central rolle i klientens nuværende situation, i og med at forsvarsmekanismer og overføring ifølge psykodynamikken er betinget og formet af netop klientens fortid og relationer (Fonagy & Target, 2003; Lemma, 2015; Ogden, 1979), hvorfor også overførings- og modstandsprocesser såvel som det relationelle samspil i det terapeutiske rum ligeledes tjener som en vigtig nøgle for terapeuten til at forstå, hvad der er på spil hos klienten (Beutler et al, 2002; Lemma, 2015; Ogden, 1979; Weinberg et al, 2018). Af samme grund er klientens *non*-verbale signaler af lige så stor (hvis ikke større) betydning end de ting, der bliver sagt højt i det terapeutiske rum (Lemma, 2015; Luborsky, 2007; Videbéch et al, 2010). I Danmark eksisterer der en række anerkendte og relativt velanvendte psykodynamiske terapier, særligt *interpersonel terapi* (IPT; Lemma et al, 2011) og *intensiv dynamisk korttidsterapi* (ISTDP; Abass et al, 2013) for tilknytningsforstyrrelse, panikangst og depression (Møhl & Kjølbye, 2013; Hougaard, 2019), samt *mentaliseringsbaseret terapi* (MBT; Bateman & Fonagy 2004), som især har fundet stor anvendelse og god effekt for borderline, såvel som for personer med

selvskadende adfærd (Bateman & Fonagy, 2016; Fonagy & Luyten, 2009; Hestbæk et al, 2022).

Den *kognitive-adfærdsterapeutiske tilgang* opstod i midten af det 20'ene århundrede som en protest mod psykoanalysen, der i stigende grad blev betragtet som uvidenskabelig og utilstrækkelig empirisk funderet (se blandt andet Eysenck, 1952), og er i skrivende stund stadig den mest populære og anvendte psykoterapeutiske behandlingstilgang for en bred vifte af patologier og mistrivsel herhjemme (Hougaard, 2019; Jørgensen et al, 2017; Møhl & Kjølbye, 2013; Rosenberg et al, 2012). Tilgangen tager udgangspunkt i hjernen som et organ, der kan formes og omformes af både ydre omstændigheder såvel som indre stimuli i form af tanker og emotioner (Beck, 2011; Clark & Beck, 2010; Kolb et al, 2003; Schrammen et al, 2022), og anskuer af samme grund psykopatologi og mistrivsel som rodfæstet i maladaptive kognitive skemaer eller skemata (såkaldte *kognitive forvrængninger*) samt tillærte, uhensigtsmæssige adfærdsstrategier, der aktivt kan aflæres igen (Beck, 2011; Rosenberg et al, 2012). Den mest populære behandlingsform indenfor denne tilgang er *kognitiv adfærdsterapi* (herfra KAT; på engelsk *Cognitive Behavioral Therapy,* CBT; Beck, 1973; Beck, 2011 & 2013; Clark & Beck, 2010; Rosenberg et al, 2012), der blev populariseret i start-1970'erne af den amerikanske psykolog Aaron Beck (Beck, 1972; Beck, 2011; Hougaard, 2019; Møhl & Kjølbye 2013). I KAT opererer man med den såkaldte *kognitive diamant* (eller KAT-modellen; Arendt & Rosenberg, 2012; Beck, 2011; Rosenberg et al, 2012), hvor klientens tanker, følelser, kropslige fornemmelser samt adfærd betragtes som gensidigt forbundne elementer, der forstærker og påvirker hinanden, og som formes af (såvel som former) klientens underliggende skemata, basale antagelser samt leveregler om sig selv og verden, og hvor klienten opnår mental bedring ved en systematisk kognitiv omstrukturering af disse elementer (Arendt & Rosenberg, 2012; Beck, 1973; Beck 2011 & Beck 2013; Rosenberg et al, 2012). KAT har fundet anvendelse som behandling for en lang række forskellige psykopatologiske tilstande (heriblandt angst, depression og OCD; Bohni et al, 2014; Arendt & Rosenberg, 2012; Staarup, 2012), såvel som mere almen mistrivsel, og er den primære terapeutiske behandlingsform i offentligt dansk regi både i psykiatrien og kommunale tilbud, og bruges også i en mindre grad i visse civile rådgivningstilbud (Ottosen et al, 2018; Møhl & Simonsen, 2017; Rosenberg et al, 2012). Udover KAT er andre populære kognitive terapiformer blandt andet *metakognitiv terapi* (MCT; Wells, 2011), der går ud på at ændre selve den *måde*

klienten tænker (frem for blot tænkningens indhold, som det er tilfældet ved KAT), og derved give klienten kontrol over egen tankeproces (Møhl & Kjølbye, 2013; Wells, 2011; Wells et al, 2009), og også *mindfullness,* der med inspiration fra blandt andet Buddhistiske meditationsteknikker (og som kontrast til MCT) i stedet fokuserer på at lære klienten at *slippe* kontrollen over sine tanker og omstændigheder, og i stedet at være i nuet (Kabat-Zinn, 2003; Segal et al, 2013). MCT og mindfulness har vist god effekt i særligt stressbehandling, men også til angstlidelser og depression (De Vibe et al, 2012; Grossman et al, 2012; Khoury et al, 2015; Li et al, 2019; Norman et al, 2014; Norman & Morina, 2018; Rochat et al, 2018; Segal et al, 2013; Wells, 2011; Wells & King, 2006). Af andre hyppigt anvendte kognitive terapier er såkaldt *acceptance & commitment therapy* (ACT; Hayes et al 1999/2011), der går ud på at øge klientens kognitive fleksibilitet ved at lære klienten at acceptere svære tanker, følelser og erkendelser (Hayes et al, 2006; Møhl & Kjølbye 2013), samt *dialektisk adfærdsterapi* (DBT, Linehan 1987; Linehan et al 1992), der kombinerer adfærdsterapi med mindfulness, og som ligeledes har vist god effekt for særligt borderline (Linehan, 1987; Møhl & Kjølby, 2013; Swales et al, 2000).

Sideløbende med den kognitive tilgang opstod også den *humanistiske/eksistentialistiske tilgang* som en modreaktion på de to andre tilganges tendens til at sygeliggøre klienten, og blev særligt populariseret med Carl Rogers *klientcentrerede terapi* (Rogers, 1985; Jørgensen, 2012; Hougaard, 2019; Møhl & Kjølbye, 2013). Den humanistiske tilgang tager udgangspunkt i mennesket som et socialt og selvreflekterende væsen, der i kraft af at *være* et sådant væsen også har et grundlæggende behov for at blive anerkendt, accepteret og forstået af andre mennesker, og som kun kan udfolde og realisere sig selv i kontakten med og relationen *til* andre mennesker (Elliot, 2013; Hougaard, 2019; Rogers, 1957 & 1965; Rogers, 1985). Indenfor den klientcentrerede terapi er det terapeutens evne til at møde klienten med empati, oprigtighed samt ubetinget positiv accept (*unconditional positive regard,* Rogers 1957), der fremhæves som de grundlæggende betingelser for at klienten opnår bedring (Elliot, 2013; Rogers, 1957, 1965 & 1975; Hougaard, 2019). Mistrivsel og patologi er ikke udtryk for en iboende *mangel* eller *forkerthed* hos klienten som skal rettes op (i modsætning til ved psykodynamikken og den kognitive tilgang), men snarere resultatet af at klientens iboende behov for anerkendelse og accept ikke er blevet tilstrækkelig mødt af *andre* mennesker (Rogers, 1965; Rogers, 1975; Hougaard, 2019; Maslow), hvorfor terapeuten netop skal agere som et ikke-

dømmende, ligeværdigt medmenneske, klienten kan spejle sig i og føle sig valideret af, så dette behov kan opfyldes (et slags positivt *spejl*, så at sige; Rogers, 1957). Til denne tilgang hører også konceptet *aktiv lytning,* som indebærer at terapeuten aktivt lytter på klienten med *hele* sin fulde opmærksomhed, og tilsidesætter egne fordomme og synspunkter(Rogers & Farson, 1987). Omend tilgangen ikke længere er en egentlig terapiform, bliver mange af dens grundlæggende principper stadig anvendt indenfor blandt andet misbrugsbehandling, coaching, arbejdsrelateret rådgivning, sjælesorg, sygepleje samt (i kombination med KAT) til terapi for unge såvel som i terapeutisk praksis mere generelt (Fog, 1998; Greenberg et al, 1997; Hill, 2020; Hougaard, 2019; Møhl & Kjølbye, 2013). Det er ligeledes denne tilgang, som er udgangspunktet i samtaletilbuddet *headspace,* hvor det bærende princip er at 'møde de unge i øjenhøjde', hvilket kan sidestilles med den ubetingede positive accept og spejling central for klientcentreret terapi (headspace 2013/2019; Bjørkedahl et al, 2025).

Slutteligt er der også den *systemiske og narrative* tilgang, som særligt anvendes inden for familieterapi (Møhl & Kjølbye, 2013). Med den systemiske tilgang anskues roden til klienternes problemer som værende *eksternt* fra klienten, hvor klientens mistrivsel er en respons på et dysfunktionelt system som klienten er underlagt , og altså ikke resultatet af en dysfunktion i klienten selv (Møhl & Kjøbye, 2013). *Narrativ terapi* anvendes ligeledes inden for misbrugsbehandling (og i mindre grad som en del af metakognitiv terapi), og betragter mistrivsel og psykopatologisk tilstand som rodfæstet i forkerte selvfortællinger, hvor målet med terapien dermed bliver at skabe en ny selvfortælling hos klienten, som klienten kan tage ejerskab over (Møhl & Kjølbye, 2013).

### 2.3.3. Psykoterapien som helhed og den terapeutiske alliance

Trods de forskellige tilgange trækker moderne psykoterapi ofte på en praksisforståelse, der inddrager alle fire tilgange (eller forståelsesrammer) i en samlet teoretisk helhed (Hougaard, 2019; Norcross & Wampoldt, 2011a), hvor terapeuten gør brug af specifikke forståelsesrammer og specifikke metoder *i* sin praksis, alt afhængig af den pågældende problemstilling hos en given klient (denne tilgang kaldes også for *eklektisk;* Møhl & Kjølbye, 2013; Hougaard, 2019). Ud fra dette helhedsperspektiv bør mennesket og menneskelig mistrivsel såvel som terapeutisk praksis forstås *både* psykodynamisk, kognitivt, humanistisk, systemisk og narrativt (Hougaard, 2019; Hill, 2020; Norcross & Wampoldt, 2011a; Norcross & Wampoldt,

2011b; Norcross & Wampoldt, 2018; Zarbo et al, 2016). Dette helhedsperspektiv på terapeutisk praksis er særligt et resultat af effektforskningen (Norcross & Wampoldt, 2018), der ikke har påvist nævneværdig forskelle i effektstørrelser mellem de individuelle tilgange (Cuijpers et al, 2016; Cuijpers et al, 2018; Cuijpers et al, 2023; Lilliengren 2023; Leichsenring et al, 2023; Munder et al, 2019; Wampoldt, 2017; Wampoldt et al, 1997; Wampoldt et al, 2017a; Wampoldt et al, 2017b). Samtidig er der bred konsensus på tværs af tilgangene om, at psykoterapi i sin grundform også er en *mellemmenneskelig praksis*, hvor det som udgangspunkt kræver et menneske at forstå et andet menneske (Fog, 1998; Hougaard, 2019). I og med at mennesker er sociale dyr, og vores kognitive processer og mentale udvikling i høj grad formes og faciliteres af vores relation med andre mennesker (Atzil et al, 2018; Champagne et al, 2005; Karlsson, 2011; Linden, 2006; Beauregard, 2014; Schrammen et al, 2022), er terapeutisk behandling sandsynligvis effektivt netop *fordi* det bygger på en sådan relation mellem to individer (Hougaard, 2019; Møhl & Kjølbye, 2013). Med andre ord er det muligvis selve *relationen*, der er den primære katalysator og facilitative enhed for en terapeutisk forandring hos klienten (Flückiger et al, 2018). Dette kan også være en af forklaringerne på, hvorfor forskningen netop har påvist ens effekt på tværs af tilgange (Hougaard, 2019, Møhl & Kjølbye 2013; Norcross & Wampoldt, 2011). Ideen om relationen som det primært faciliterende for en terapeutisk forandring danner også rammen for *den terapeutiske alliance* (også kaldet *arbejdsalliancen*) (Bordin 1979, Hougaard 2019), der blev formaliseret af Edvard Bordin i 1979, og er fremhævet af den psykoterapeutiske forskning som den mest prædiktive faktor for terapiens udfald på tværs af tilgangene (Ardito & Rabellino, 2011; Baldwin et al, 2007; Flückiger et al, 2012; Flückiger et al, 2018 & 2019; Flückiger et al, 2020; Horvath et al, 2011; Lavik et al, 2018; Martin et al, 2000; Norcross & Wampoldt, 2011; Wampoldt, 2015; Weinberger, 1993). Alliancen kan groft inddeles i tre bærende elementer; *mål,* som indebærer at der skabes en målsætning for terapien, og at der er enighed om disse mål mellem terapeuten og klienten, hvilket også indebærer faktorer som hvad klienten forventer af terapien og terapeuten, såvel som terapeutens forventninger til klienten; *opgaver,* som indebærer opgaver og øvelser klienten skal udføre mellem terapisessioner, og som også kan medføre en øget tillid til terapeutens kompetence, særligt hvis opgaverne opleves af klienten som havende en effekt; Samt et genuint emotionelt *bånd* mellem klient og terapeut bygget på gensidig tillid, og hvor dette bånd og denne tillid øger sandsynligheden for at klienten vil følge op på

terapiens opgaver (Ackerman & Hilsenroth, 2003; Bordin, 1979; Flückiger et al, 2018; Hougaard, 2019; Ulvenes et al, 2012). Til opgaver er også knyttet *feedback,* det vil sige at terapeuten giver klienten feedback og respons på klientens udførelse af opgaver og fremskridt i terapien, såvel som klientens mulighed for det samme overfor terapeuten (Flückiger et al, 2019). Ifølge teorien om den terapeutiske alliance er der en vekselvirkning mellem alliancens tre elementer, hvorpå en manglende opfyldelse af ét element kan påvirke de øvrige elementer og dermed skabe alliancebrud, som det så kræver en aktiv og selvreflekterende indsats at få genoprettet (Ackerman & Hilsenroth, 2003; Bordin, 1979; Eubanks et al, 2018; Flückiger et al 2018 & 2019; Flückiger er al, 2020; Safran et al, 2011; Schön, 2001).

Ideen om relationen som den primære faktor for en terapeutisk forandring bevirker også, at relationelle terapeut-kompetencer som empati og emotionel intelligens bliver gjort til de vigtigste elementer af den terapeutiske alliance (og dermed terapiens effekt) uagtet tilgangen (Ackerman & Hilsenroth, 2003; Ardito & Rabellino, 2011; Baldwin et al, 2007; Crits et al, 2006; Del et al, 2012 & 2021; Flückiger et al, 2020; Nissen-Lie et al, 2015). Ifølge blandt andet Hill et al (2017) er det også disse relationelle kvaliteter, der gør det muligt for terapeuten at forstå klienten, monitorere fremskridt og give klientfeedback, respondere med korrekt spejling og validering, samt skabe et genuint *bånd* med klienten (Hill et al, 2017; Hill, 2020). Netop selvsamme grundlæggende elementer, der ifølge teorien om alliancedannelse tjener til at oprette, understøtte og vedligeholde den terapeutiske alliance over tid (Bordin, 1979; Eubanks et al, 2015; Heinonen et al, 2014; Hovarth, 2001; Horvarth et al, 2011; Muran et al, 2023; Nissen-Lie et al, 2015), hvilket bakkes op af kvalitative studier, hvori klienter konsistent beskriver terapeuter med ovenstående kvaliteter som værende mere behagelige og indgyde højere tillid, og derved som nøglefaktorer for et positivt udfald af terapien (Ackermann & Hilsenroth, 2001; Bachelor, 1995; Bachelor, 2013: Bedi et al, 2005; Hill et al, 2010; Hilsenroth et al, 2004). Hvis det kræver et menneske at *forstå* et menneske (og dermed også en klient, Fog 1998), og kræver terapeutkvaliteter som emotionel intelligens og genuin empati for at understøtte en terapeutisk alliance og dermed opnå en genuin, forandringsskabende relation (Hill et al, 2017), kan en sprogmodel så opnå samme terapeutiske alliance og terapeutiske effekt, givet at sprogmodeller som udgangspunkt ikke har nogen *relationelle* eller *menneskelige* erfaringer at trække på i praksis, men udelukkende besidder en grundlæggende, algoritmebaseret *maskinel* forståelse? Behøver de overhovedet *have*

nogen egentlig forståelse for at fungere terapeutisk? Dette vil jeg søge at finde svar på i følgende sektioner.

# ANALYSE & DISKUSSION:

# 3. Brugen af sprogmodeller til terapi

## 3.1. Har sprogmodeller empati og emotionel intelligens?

*Theory of mind* og *kognitiv empati* er som sagt vigtigt for at en terapeut kan identificere, hvad en given klients behov kan være i en given situation, eller hvad en given klient tænker om en given situation uden klienten nødvendigvis siger dette eksplicit, såvel som hvad der kan ligge *bag* disse tanker og behov (Elliot et al, 2018; Fog, 1995 & 1998). Den *emotionelle* intelligens er ligeledes en forudsætning for at kunne afstemme sig i forhold til klienten og møde klienten på klientens præmisser (Ivey et al, 2016; Hill 2020; Hill et al, 2017; Rogers 1975), forstå klientens reaktioner og dermed respondere korrekt på klienten (Fog, 1998; Møhl & Simonsen, 2017; Swift et al, 2017), samt forstå, hvornår der skal trækkes grænser (Dreier, 1998; Fog, 1995; Thuesen, 2015), og dermed også hvordan man kan genoprette alliancebrud (Elliot et al, 2018; Flückiger et al, 2020; Muran et al, 2023; Safran et al, 2011). Sprogmodellernes styrke er netop deres evne til at genkende sproglige mønstre, og ud fra disse mønstre forudsige menneskelig kommunikation og adfærd. I og med at emotioner og emotionelle oplevelser formidles og kommunikeres via sproget (se sektion 2.2.1), må det antages, at sprogmodellernes evne til sproglig mønstergenkendelse og forudsigelse også indbefatter en evne til at genkende og forudsige emotionelle og kognitive tilstande (og dermed udvise empati), forudsat at beskrivelser af disse tilstande indgår i deres træningsdata. Dette særligt hvis man antager, at vektorer kan indkapsle enhver form for information (jf sektion 2.2.4), hvorfor emotionel information givetvis også kan indkapsles på denne måde. Men gælder dette også i praksis?

Ideen om at computere kan lære at aflæse tanker og forstå følelser har eksisteret i AI-forskningen siden 00'erne (Bickmore & Picard, 2005: Foster, 2007: Gratch et al, 2007; Krämer, 2008; Mogardo & Gaspar, 2003; Picard, 1997). I 2009 påvistes blandt andet at virtuelle systemer agerer på en måde, der lignede at de besad *theory of mind,*

heriblandt ved at én virtuel robot lod til at kunne forudsige og afkode en anden virtuel robots intentioner (Kim & Lipson, 2009). I starten af 2010'erne blev der udviklet en virtuel agent, som både kunne aflæse brugerens emotionelle tilstand, huske denne tilstand efterfølgende såvel som løbende bruge det til at skræddersy sine svar for den enkelte bruger, og dermed indikerede en vis form for emotionel forståelsesevne (Lisetti et al, 2013). I 2018 lanceredes på forsøgsbasis EMMA (*EMotion-aware Mental-health Agent*, Ghanderioun et al, 2018), som var en virtuel smartphone-baseret bot, der via simpel *machine learning* ud fra brugerens aktiviteter og telefonens sensorer kunne opspore og korrekt udlede brugerens emotionelle tilstand, og dernæst foreslå brugeren forskellige interventioner baseret på den givne tilstand. Omend EMMA var relativt simpel og ikke en chatbot i traditionel forstand, var den ikke desto mindre endnu et bevis på, at virtuelle (og især AI-baserede) agenter lod til at kunne forstå eller i hvert fald aflæse emotioner (Car et al, 2020; Ghanderrioun et al, 2019). Dette blev også demonstreret ved en chatbot til brug i kundeservice, som tilsyneladende var bedre til at give empatiske responser end mennesker (Hu et al, 2018). Men spørgsmålet er, om disse virtuelle agenter og chatbots reelt *kunne* forstå følelser, og derved udviste empati, eller om det blot var en illusion bygget på algoritmer (Cuadra et al, 2024)? Forskningen var på den tid uklar hvad dette angik, omend studier foretaget af eksisterende chatbots indikerede, at de til en vis grad gav brugerne en følelse af at blive forstået, uagtet om denne forståelse var reel (Car et al, 2020; Fitzpatrick et al, 2017; Fulmer et al, 2018; Gan et al, 2021; Hu et al 2018; Ly et al, 2017; Pham et al, 2022; Sweeney et al 2021; se sektion 3.2). Dette motiverede i 2019 til et kandidatprojekt i maskin- og informationsvidenskab fra Roskilde Universitet, der havde til formål at undersøge, hvorvidt chatbots besad empati og *theory of mind*, her operationaliseret som evnen til kognitiv perspektivtagning (Andersen & Lynge, 2019). Forfatterne kom frem til, at AI-baserede chatbots simulerede en begrænset form for *theory of mind,* men ikke reelt besad en sådan, hvilket blandt andet kom til udtryk ved, at de fire bots, som anvendtes i studiet, var dårlige til at svare korrekt på diverse alment brugte *theory of mind* tests (Andersen & Lynge, 2019). Omend man kan argumentere for, at undersøgelsesdesignet var manipuleret på en måde, der kan så tvivl om den økologiske validitet (Cooligan, 2017), i og med at forfatterne tilpassede måleredskaber for *theory of mind* i en grad hvor redskaberne hverken målte den tiltænkte population eller blev anvendt som tiltænkt, så er det på den anden side vanskeligt at anvende den slags værktøjer til

virtuelle agenter *uden* en vis form for tilpasning, i og med at der på den tid ikke eksisterede kognitive assesmentmetoder specifikt skræddersyet til kunstig intelligens. En pointe, forfatterne da også selv vedkendte (Andersen & Lynge, 2019). I samme periode kom andre undersøgelser af chatbots dog til nogenlunde samme resultater hvad angår *theory of mind* og generel emotionsforståelse, navnlig at chatbots som udgangspunkt ikke kan siges at besidde egentlig *theory of mind* eller empati (Le et al, 2019; Pham et al, 2021, Choi, 2022; Sedlakova & Trachsel, 2023), hvilket styrker kandidatprojektets fund. Men givet at disse chatbots byggede på en simplere neural arkitektur end det er tilfældet for sprogmodeller, kan dette fund så overføres?

Meget af den første forskning i sprogmodeller omhandlede primært deres sprogforståelse (Ettinger et al, 2020; Radford et al, 2018), ræssoneringsevne (Bubeck et al, 2023; Yang et al, 2023), og brugervenlighed (Brown et al, 2020; Dale, 2021). I 2023 blev der dog foretaget en undersøgelse af specifikt ChatGPT af Sap et al, som blandt andet viste, at chatGPT udviser dårligere social forståelse end mennesker målt på modellens evne til at udlede hvilken reaktion der er mest passende i bestemte sociale scenarier, samt dens evne til at udlede, hvad en given person i et givent scenarie tænker at en *anden* person tænker (*flerdimensionel perspektivtagning,* Sap et al, 2023; Baron-Cohen, 1995). Man kan argumentere for, hvor overførbare og gældende studiets resultater stadig er, i og med at studiet tog udgangspunkt i en nu forældet version af chat-GPT, der på den tid endnu ikke var klar til at blive offentliggjort, og derfor må formodes at have flere begrænsninger i forhold til nuværende modeller (Van Duijin et al, 2023; Zhang et al, 2025). Et samtidigt studie foretaget af kognitionspsykologen Michael Kosinski fra MIT viste helt omvendt, at de nyeste sprogmodeller - og her særligt chatGPT - lader til at besidde færdigheder indenfor *theory of mind,* der overgår tidligere modeller, og som er på niveau med 6-årige børn (Kosinski, 2023 & 2024). Kosinski nåede dette resultat ved at teste modellerne med både *falsk indholds*-testen samt *Sally Anne* testen, der første gang blev introduceret i 1985 af Baron-Cohen et al som assesmentredskab til autismescreening, og som vurderer testpersonens evne til at skelne sit eget perspektiv fra andres, såvel som testpersonens evne til at udlede, at andre kan have en overbevisning om virkeligheden, der ikke stemmer overens med det faktiske (*false belief,* Baron-Cohen et al, 1985). Testen foregår ved at testpersonen præsenteres for en historie, hvor en karakter (Sally) placerer et objekt (såsom en marmorkugle i den klassiske test) i en kurv, for derefter at gå ud af lokalet. Derefter træder en ny karakter

ind (Anne), som så flytter objektet til en æske, inden hun ligeledes forlader lokalet. Testpersonen skal derefter gætte sig til, hvorvidt Sally, når hun træder tilbage i rummet, vil lede efter objektet i kurven eller æsken. Her er det korrekte svar naturligvis kurven, i og med at Sally ikke ved, at objektet er blevet flyttet, og derfor har en falsk overbevisning om, at den stadig er i kurven (deraf *false belief,* Baron-Cohen et al, 1985). I Kosinskis forsøg lykkedes det chatGPT over 75% af tiden korrekt at forudsige en given karakters falske overbevisning, både med Sally Anne testen og andre lignende varianter deraf, samt korrekt at forudsige disse karakteres adfærd baseret på denne falske overbevisning (Kosinski, 2023 & 2024). For Kosinski var det indikation på, at sprogmodeller besidder *theory of mind* som en ny emergent egenskab*,* i hvert fald hvis man måler *theory of mind* alene på evnen til at forudsige falske overbevisninger (Kosinski, 2023 & 2024). Dog blev dette studie hurtigt kritiseret af blandt andet Tomer Ullman, der ligeledes testede ChatGPTs evne til korrekt at udlede falske overbevisninger, og hvor han (i kontrast til Kosinski) kom frem til det resultat, at ved blot små trivielle ændringer i testopgavernes indhold (eksempelvis ved at ændre de fysiske egenskaber på genstandene i *Sally Anne* testen), kan ChatGPT ikke længere svare korrekt hvad angår en given karakters falske overbevisning (Ullman, 2023). Et konkret eksempel på dette var et scenarie med en pose chokolade, hvor der på posens mærkat stod 'popcorn', hvilket var et eksempel Kosinski også anvendte i sit eget studie (Ullman, 2023). Selvom chatGPT korrekt gættede, at scenariets hovedperson fejlagtigt ville gå ud fra at posen indeholdt popcorn ud fra posens mærkat, så fejlede den opgaven da posen blev gjort gennemsigtig, hvor den fastholdte, at hovedpersonen stadig ville tro posen indeholdt popcorn, selvom hovedpersonen nu klart kunne se posens reelle indhold. Ud fra dette konkluderede Ullman, at sprogmodeller som chatGPT ikke benytter sig af reel theory of mind, men i stedet på overfladisk mønstergenkendelse udledt fra træningsdata, der *ligner* forståelse, men blot er en simulation, ikke ulig computerlingvisten Emily Benders formulering om sprogmodeller som stokastiske papegøjer (Bender et al, 2021; Chen et al, 2024; Hu et al, 2025; Nguyen, 2025; Shapira et al, 2024; Stade et al, 2025; Ullman, 2023). Ifølge Ullman er det ikke utænkeligt, at lige præcis *sally anne* og lignende tests indgår i modellernes træningsdata, og derfor kan anvendes som referencepunkt for modellerne til at lave en statistisk forudsigelse af karakterernes adfærd, hvilket også forklarer, hvorfor modellernes tilsyneladende *theory of mind* færdigheder styrtdykker ved blot simple manipulationer af disse tests, som ellers ikke

vil påvirke en menneskelig testperson i samme grad (Ullman, 2023; Pi et al, 2024). Et andet studie inden for samme periode fandt ligeledes, at sprogmodellerne (inklusive chatGPT) er langt dårligere end børn til at opfange eller forstå såkaldt *faux pas,* hvorved en person gør eller siger noget der opfattes som socialt upassende og sårer en anden person men *uden* dette er hensigten, og heller ikke lader til at forstå, hvilke emotioner der er passende i hvilke sociale situationer, i overenstemmelse med Sap et als studie (Shapira et al, 2023). Dette fund bakker op om Ullmans studie, og peger ligeledes i retning mod, at sprogmodellers sociale (og givetvis også emotionelle) forståelse blot er illusorisk, og udtryk for en Eliza-effekt (Shapira et al, 2023; Cuadra et al, 2024; Reinecke, 2025). Omend dette studie ligeledes gør brug af en ældre version af ChatGPT, så er der ikke desto mindre et nyt studie fra 2024 (Strachan et al), der har replikeret Shapira og kollegers resultater hvad angår sprogmodellers manglende evne til at opfange og aflæse *faux pas* (inklusive GPT-4). Dette studie fandt dog mere robust evidens for noget, der ligner general *theory of mind* hos særligt ChatGPT på de fleste andre parametre undtagen evnen til at opfange *faux pas*, heriblandt at kunne forudsige falsk overbevisning og skelne mellem bevidst og ubevidst vildledning i fiktive sociale scenarier (Strachan et al, 2024). *Faux pas* resultatet kunne ifølge Strachan et al skyldes, at sprogmodeller som chatGPT trænes til at helgardere sig og ikke komme med for skråsikre svar, hvorfor modellerne giver flere forskellige alternative forklaringer på en given karakters adfærd (hvori *faux pas* også er inkluderet), men uden at dedikere sig til en enkelt forklaring, selv hvis de har opfanget en given *faux pas* (Strachan et al, 2024). Dette er dog ret så spekulativt og tillægger modellerne en ræsonneringsproces, der ikke kan bevises, og kan i sig selv ses som et udtryk for en antropomorfisering af modellerne (Cuadra et al, 2024; Liu et al, 2024; Maeda & Quan-Haase, 2024; Mitchell & Krakauer, 2023, Shapira et al, 2024), i og med at modellerne kom med alternative forklaringer selv når det var tydeligt, at den eneste forklaring var *faux pas*. Når modellerne på denne måde ligger lige meget vægt på alle forklaringer uden skelnen, kan man stille sig spørgsmålet om, hvorvidt de overhovedet forstår hvad de forklarer, eller blot finder de mest sandsynlige responser, og således kun giver en illusion om forståelse (y Arcas, 2022; Bender et al, 2021, Chen et al, 2024; Liu et al, 2024; Mitchell & Krakauer, 2023; Shapira et al, 2024). I slutningen af 2023 fandt en undersøgelse af van Duijn et al foretaget på 11 forskellige sprogmodeller til gengæld, at fintunede versioner af chatGPT, og særligt GPT-4, er bedre end børn til opgaver, der særligt kræver at forstå

og afkode ikke-bogstavelig kommunikation (såsom ironi, sarkasme, talemåder, at sige ét og mene noget andet, brugen af hvide løgne etc, såkaldte *strange stories,* van Duijn et al, 2023; Happé, 1994), hvilket kan indikere en vis evne til at læse mellem linjerne. Til gengæld kæmper modellerne med flerdimensionel perspektivtagning, særligt over det tredje niveau, det vil sige det niveau hvor man udleder hvad en given person tror en anden given person vil have en tredje given person til at tænke om en given situation eller person (van Duijn et al, 2023; Baron-Cohen, 2001), hvilket er et resultat som på sin vis afspejler de forrige studier. Hvis man her antager, at denne form for ræsonnering om andres mentale tilstande kræver *mere* end blot mønstergenkendelse og sproglig forudsigelse, men også indebærer en aktiv kognition og ræsonneringsevne (og dermed rækker udover hvad sprogmodeller er i stand til), så er resultatet af Van Duijin et als studie ikke overraskende hvad angår modellernes ringe evne til flerdimensionel perspektivtagning, hvilket ligeledes underbygges af andre studier (Chen et al, 2024; He et al, 2023; Shapira et al, 2024). At modellerne samtidig er bedre end børn til at afkode *strange stories* behøver heller ikke være udtryk for en reel social forståelse, men kan lige så vel skyldes at modellernes træningsdata også består af en stor mængde skønlitteratur, hvori denne form for kommunikation hyppigt indgår, og modellerne derfor 'genkender' (og dermed kan forudsige) dette mønster i testscenariet (noget forfatterne også selv anerkender, van Duijin et al, 2023). En række nyere studier fra 2024, som ligeledes undersøgte modellernes evne til perspektivtagning, har nået samme konklusion som van Duijn et al, navnlig at sprogmodellers *theory of mind* er illusorisk og bygger på fundamentalt andre processer end hos mennesker (i form af statistisk udledning), og at modellerne ikke formår at sætte sig ind i andre karakterers perspektiver, hvis man blot ændrer på det givne sociale scenarie så det fremstår mere unikt eller mere komplekst (Shapira et al, 2024; Sclar et al, 2024). Et andet væsentligt kritikpunkt af de positive *theory of mind* studier er desuden, at mange af dem måler mere overfladiske *theory of mind* færdigheder, der kan tilskrives statistisk udledning, hvorfor der er brug for mere skræddersyede og gennemgående måleredskaber for sprogmodeller, der kan diskriminere bedre mellem reel *theory of mind* ræssonering og tilfældig mønstergenkendelse (Chen et al, 2024; Hu et al, 2025; Riemer et al, 2024; Wu et al, 2023).

Omvendt er resultaterne hvad angår sprogmodellers evne til at udlede tanker, intentioner og perspektiver interessante og samtidig svære helt at afskrive. Særligt

chatGPT lader til at blive bedre til *theory of mind* med fintuning, prompt-engineering og forøgning af parameterstørrelse (Bortelleto et al, 2024; Moghaddan & Honey, 2023; Street et al, 2024; Xu et al, 2024), hvor perfomance når et tilsyneladende menneskeligt niveau sammenlignet med de modeller, der ikke er fintunet, hvilket i hvert fald peger mod en rudimentær *theory of mind* hos sprogmodeller (omend andre studier dog også viser markante begrænsninger selv efter fintuning, se her særligt Chen et al, 2024 samt Riemer et al, 2024). Et studie af Zhu et al (2024) lader eksempelvis til at påvise, at sprogmodeller besidder en indre, matematisk repræsentation af både modellernes egne såvel som andre agenters perspektiver på virkeligheden indlejret som adskilte vektorer i specifikke lag af modellernes netværk (det vil sige et sæt vektorer for modellens eget perspektiv, og et andet sæt for en ydre agents perspektiv), og som ifølge studiets forfattere ikke kan forklares ud fra simpel statistisk sandsynlighed alene (Zhu et al, 2024). Studiet tog udgangspunkt i en mindre sprogmodel (Mistral-7-b), og benyttede en metode kaldet *lineær probing* (Alain & Bengio, 2018; Goodfellow et al, 2016), hvor man med en regressionsanalyse kan afkode, hvilke informationer der er indkapslet af bestemte vektorer, samt hvorvidt disse informationer anvendes, når sprogmodellen skal generere et output baseret på et bestemt input (ikke ulig måling af hjerneaktivitet hos mennesker under bestemte aktiviteter og stimuli, Alain & Bengio, 2018; Bengio et al, 2016; Meinke et al, 2024). Modellen blev præsenteret for en række forskellige *false belief* scenarier, der fulgte samme skabelon som den oprindelige Sally Anne test, og blev analyseret med den lineære probing-metode, mens den besvarede scenarierne. Ud fra denne metode fandt studiets forfattere frem til, at specifikke attentionheads i modellens mellemste lag blev aktiveret når modellen skulle udlede en given karakters overbevisning, både når denne overbevisning stemte overens med det faktiske (og dermed modellens egen overbevisning), såvel som hvis karakterens overbevisning var falsk, målt ud fra de vektorer, modellen trak på under genereringen af output-tokens. Omend modellen kun svarede korrekt hvad angik *false belief* under halvdelen af tiden (hvilket var på lige fod med tilfældige udslag), ændrede dette mønster sig drastisk, da forfatterne via samme probing-teknik manipulerede de specifikke vektorer, modellen anvendte under sin udledningsproces, ved at lægge flere vægte til de vektorer, der repræsenterede perspektivet hos den fiktive karakter, og tilsvarende færre til vektorer, der indkapslede modellens eget perspektiv. Denne manipulation øgede modellens evne til at udlede *false belief* hos en fiktiv karakter, og følgelig ændre sit output. Selvom dette resultat

ganske vist blev opnået via ydre manipulation, så var det alligevel for forfatterne tegn på, at sprogmodeller besidder en latent evne til *theory of mind*, som er udtryk for mere end blot en overfladisk mønstergenkendelse, og som kan fremmanes med de rette teknikker (Zhu et al, 2024). Deres resultat stemmer også overens med andre lignende studier fra nogenlunde samme tidspunkt, der ligeledes peger mod, at sprogmodeller har specifikke aktivitetsmønstre i specifikke lag af deres neurale netværk, som koder specifikt for hvad der ligner *theory of mind*, akkurat som det er tilfældet for menneskehjernen (Bubeck et al, 2023; Jamali et al, 2023; Meinke et al, 2024). Hvis den herskende teori om *theory of mind* som værende en færdighed, der specifikt er koblet sammen med sprogforståelse (Baron-Cohen, 1999; Bender, 2019; De Villiers & De Villiers, 2014; Malle, 2008; Miller, 2006), er gældende, så giver det på sin vis god mening, at sprogmodeller over tid udvikler færdigheder, der i hvert fald kan sidestilles med *theory of mind* og kognitiv empati, da de netop er sprogbehandlingssystemer (Zhao et al, 2025). Omend evidensen herfor stadig ikke er helt robust, i og med at næsten alle studierne (inklusive de nyeste) stadig kan forklares ud fra avanceret mønstergenkendelse og statistisk udledning, så kan det heller ikke fuldstændig afvises at dette i sig selv leder til kognitiv empati givet nok tid, eller som minimum en simulering af kognitiv empati (Li et al, 2024; Verma et al, 2024; Zhou et al, 2023; Zhu et al, 2024). Men selv hvis vi antager, at sprogmodeller simulerer en vis grad af kognitiv empati (og dermed evnen til at forstå andres kognitive perspektiver), er det så grund nok til også at antage, at de besidder *affektiv* empati, og dermed en latent evne til at forstå og sætte sig ind i andres følelser, og følgelig også forstå en given klientens følelser på et mere omfattende plan? Har de med andre ord den nødvendige *emotionelle* intelligens til at agere terapeuter?

I foråret 2023 udkom et studie af Elyoseph et al, der søgte at afdække chatGPTs emotionsforståelse ved hjælp af LEAS (*levels of emotional awareness-scale*, Lane et al, 1992), som er en assesmentskala der måler, i hvor høj grad testpersonen kan identificere, forstå og artikulere både sine egne og andres emotioner via en række forskellige emotionelt ladede scenarier involverende to personer, hvoraf testpersonen er den ene. I den klassiske version skal testpersonen forestille sig og artikulere sin egen emotionelle oplevelse af og reaktion på et givent scenarie, eksempelvis hvor testpersonen kommer forbi en bro og ser en anden person stå på kanten for at springe ud, såvel som identificere og artikulere de følelser, der kan være på spil for det subjekt, scenariet kredser om, i dette tilfælde personen ved broen. I og med at

chatGPT og andre sprogmodeller så vidt vides ikke har nogen bevidsthed eller *qualia* (Chalmers, 2023; Argueras, 2022), og dermed heller ingen subjektive emotioner, blev LEAS justeret således, at chatGPT blev tildelt en rolle som den alvidende observatør i et givent scenarie via en kontekstprompt med emotionel stimuli ("du er observatøren af *x* scenarie involverende *y* subjekter, hvor der sker *z*") jævnfør Li et al(2023). Herefter blev den bedt om at udlede, analysere og forklare emotionerne hos scenariets to menneskelige subjekter (Elyoseph et al, 2023). Undersøgelsen blev foretaget af to omgange, henholdsvis januar og februar 2023, og resultaterne viste, at chatGPT scorede langt højere på LEAS end gennemsnittet hos den menneskelige kontrolgruppe, og at den opdateret version fra februar 2023 var endnu mere detaljeret og uddybende i sin analyse og forklaring af subjekternes følelsesmæssige reaktioner end januar-versionen. Dette betyder med andre ord, at chatGPT ifølge Elyoseph et al formår korrekt at udlede andre aktørers indre følelsesliv og følelsesmæssige reaktioner givet bestemte scenarier, samt udviser en tilsyneladende høj forståelse for årsag og virkning hvad angår sammenhæng mellem hændelser, følelser og adfærd (Elyoseph et al, 2023). Kort efter blev der lavet et opfølgende studie, hvor målet i stedet var at undersøge, hvorvidt chatGPT kan justere og differentiere i intensiteten af emotionelle reaktioner hos en given fiktiv karakter i respons til et givent fiktivt scenarie, afhængig af hvorvidt karakteren har borderline eller skizoid personlighedsstruktur (Hadar-Shoval et al, 2023). Her var resultatet ligeledes, at chatGPT lader til korrekt at kunne tilpasse karakterernes responser (og den emotionelle intensitet af disse responser) afhængig af karakterens personlighedsstruktur, hvilket kan ses som en validering af resultaterne fra studiet af Elyoseph et al. Fundene fra begge studier underbygges af Feng et al (2023), som ligeledes fandt, at sprogmodeller formår at udlede latente emotioner hos et givent subjekt baseret på tekst og udsagn fra dette subjekt alene. Hvis man kigger på den førnævnte hypotese om, at sprogmodeller som chatGPT netop excellerer i sproglig mønstergenkendelse, og at denne færdighed gør dem bedre end mennesker til at forudsige kommunikation (og følgelig også *følelsesmæssig* kommunikation og følelsesmæssige scenarier), er resultater som ovenstående ikke overraskende, og giver styrke til denne hypotese, uden dette i sig selv indikerer en reel forståelse hos modellen (Shapira et al, 2024). Ydermere er særligt resultaterne hvad angår chatGPTs evne til at differentiere i følelsesmæssig intensitet i de emotionelle responser hos karakterne med henholdsvis borderline- og skizoid personlighedsforstyrrelse heller

ikke overraskende, i og med at det må forventes, at chatGPT også er blevet trænet på beskrivelser af disse to tilstande, og ud fra dette er i stand til at forudsige, hvordan en person med borderline eller skizoid personlighedsstruktur vil reagere på en given situation sammenlignet med en person uden. Dette kan i øvrigt også være forklaringen på, hvorfor chatGPT virkede noget overfladisk og stereotyp i sin beskrivelse af den emotionelle reaktion hos særligt personen med borderline, da en sådan stereotypisk fremstilling er en del af det populære narrativ, som netop indgår i modellens træningsdata (Hadar-Shoval et al, 2023). Omvendt har studiernes fund alligevel positive implikationer for sprogmodellernes evne til at kunne skræddersy en given respons til en given psykologisk problemstilling i en terapeutisk kontekst. Senere udkom et studie af Huang et al (2024), som undersøgte, hvorvidt sprogmodeller formår at udvise samme emotionelle respons som mennesker givet bestemte scenarier (ikke ulig Elyoseph et al og Hadar-Shoval et al). I modsætning til de tidligere studier forsøgte Huang et al at undersøge sprogmodellernes *egen* respons på diverse emotionelt ladede hverdagshændelser, hvor modellerne via en kontekstprompt skulle forestille sig at være personen, der blev udsat for den pågældende hændelse, og altså ikke tolke og forudsige *andres* reaktioner (Huang et al, 2024). Studiet fandt, at sprogmodeller (særligt chatGPT) generelt udviser en passende enten negativ eller positiv emotionel respons afhængig af scenariet, men at responserne har en tendens til at være væsentligt mere intense og overdrevne end hos mennesker, og med flere emotionelle udsving. Dog lod det ikke til at modellerne udviste jalousi selv når det ville give mening at føle jalousi (eksempelvis ved scenarier hvor en anden person opnår noget som man selv gerne ville have haft), og heller ikke formåede at koble emotioner sammen på tværs af scenarier (det vil sige at en given emotion udløst i ét scenarie også kan udløses af et andet scenarie). Dette sidste fund er særligt interessant, for det sætter spørgsmålstegn ved, hvorvidt sprogmodeller formår at fange det *fænomenologiske* og *sociale* aspekt af emotioner, der ikke kan fanges via sprog alene (Mayer, 1996; Salovey & Mayer, 1990). Evnen til at udlede, at en given person vil opleve samme underliggende emotion ved to ellers på overfladen kvalitativt forskellige scenarier, kræver formentlig en dybere kontekstforståelse af scenariernes sociale og emotionelle betydning for både det enkelte subjekt og i mere generel forstand, og dermed en vis form for intuition om, hvordan det er at *opleve* verden som det givne subjekt (og dermed besiddelse af *qualia*), hvilket kan forklare modellernes manglende færdigheder i dette domæne, i og med de netop mangler *qualia* (Chalmers

2022). Det interessante ved dette fund er, at det også afspejler modellernes ringe evne til at opfange *faux pas*, som i høj grad kræver mange af de samme færdigheder (Shapira et al, 2023). Det giver god mening at modellerne for eksempel korrekt kan forudsige en given emotionel reaktion på specifikke emotionelle scenarier som de er præsenteret direkte for, selv hvis de blot er statistiske mønstergenkendere, netop fordi de, givet deres kolossale træningsdata, næsten med garanti er kommet forbi lignende scenarier under træningen, som de så kan anvende til at forudsige nuværende scenarier. Denne evne er i sig selv ekstremt imponerende og kan på sin vis betragtes som en form for krystalliseret intelligens, i og med at menneskehjernens egne kognitioner i langt overvejende grad også bygger på mønstergenkendelse udledt af træningsdata i form af over to millioner af års evolution (Workman & Reader, 2019). Men i modsætning til menneskehjernen udviser modellerne samtidig en mangel på *flydende* intelligens, i og med at de ikke evner at overføre viden og færdigheder fra ét domæne til et andet (Shoojae et al, 2025), hvorfor det ligeledes giver mening, at de mangler en større emotionel kohærens på tværs af scenarier.

Et studie foretaget af Wang et al (2023) undersøgte sprogmodellers (heriblandt også chatGPT) emotionelle intelligens via en særligt skræddersyet emotionel intelligenstest ved navn SECEU (*Scenario-based emotion comprehension and emotion understanding*, Wang et al 2023). Denne test blev udviklet med afsæt i den allerede eksisterende MSCEIT-skala, som måler emotionel intelligens hos mennesker ud fra evnen til at opfange, forstå og afstemme emotioner afhængig af kontekst. SECEU anvendtes (i stil med LEAS-studierne) til at vurdere, hvorvidt sprogmodeller kunne forudsige, hvilke emotioner der med størst sandsynlighed ville blive aktiveret hos en given hovedperson som respons på bestemte scenarier, samt intensiteten af disse emotioner (Wang et al, 2023). Testen bestod af 40 scenarier, hvor modellen på baggrund af hvert scenarie skulle vælge mellem fire emotionelle reaktioner (såsom vrede, tristhed, glæde, overraskelse), og tildele hver reaktion (eller følelse) en intensitetsscore mellem 0 og 10 (Wang et al, 2023). De fire reaktioner var forbundet med enten en positiv eller negativ følelsesmæssig overordnet dimension, hvor eksempelvis scenarier der involverede tab eller konflikter var konceptualiseret på skalaen som *negative* og matchet med tilhørende negative emotioner som sorg, vrede eller frustration (og omvendt for et positivt scenarie). Testens svarmuligheder var udledt fra en menneskelig kontrolgruppe for at sikre, at sprogmodellernes svar stemte overens med hvad et menneske ville svare, det vil sige jo højere SECEU-score, desto

mere overensstemmelse mellem modellens svar og kontrolgruppens svar. Resultatet af Wang et als studie viste, at jo flere parametre en given sprogmodel består af, desto bedre tilsyneladende emotionel intelligens udviser den, målt på evnen til at finde den mest sandsynlige (og passende) emotionelle respons på et givent scenarie, hvor ChatGPT som den største sprogmodel havde en samlet score på SECEU som overgik alle de andre sprogmodeller og endda nåede over et menneskeligt niveau, hvilket styrker ideen om, at sprogmodeller besidder emotionel intelligens og forståelse i en eller anden grad (Wang et al, 2023). Dog har studiet den svaghed at SECEU kun anvender én dimension for hvert scenarie, og dermed kun emotioner relateret til den ene dimension, hvilket kan risikere at modellerne så at sige 'tvinges' til at finde det mest statistisk sandsynlige svar, uden dette nødvendigvis er udtryk for egentlig forståelse (eksempelvis at vælge emotionen 'frustreret' i et scenarie med konflikt, hvor der kun er negative emotioner at vælge fra, og 'frustreret' derved bliver den mest sandsynlige svarmulighed). Med andre ord kan man her risikere, at det er modellernes mønstergenkendelse snarere end emotionsforståelse, man ender med at måle, hvilket påvirker målingens validitet (Sabour et al, 2024). Denne faldgrube motiverede til et mere omfattende *benchmarking*-studie af hvordan man bedst muligt kan måle sprogmodellers emotionelle intelligens og forståelse (EQ-bench; Paech, 2024), som netop tog højde for denne faldgrube. Her blev sprogmodellerne præsenteret for et omfattende datasæt bestående af forskellige dialoger med et overordnet spændings- og konfliktfyldt tema, og dernæst fire svarmuligheder, der indkapslede emotioner af både negativ og positiv karakter, eksempelvis "baseret på x dialog med y, vurder da hvilken emotion x nu føler", hvor de to første svarmuligheder kan være 'vrede' og 'overraskelse', mens de to andre kan være 'tilgivende' og 'tilfreds' (Paech, 2024). Via denne metode fandt man, at de mest avancerede sprogmodeller (særligt ChatGPT og Claude) kan differentiere mere passende emotioner fra mindre passende emotioner under selv mere komplekse sociale scenarier, og hvor deres svar stemmer mere eller mindre overens med svarene fra den menneskelige kontrolgruppe. Dette fund bakker op om de forrige studier, tilmed med den ekstra styrke at modellerne konsistent valgte den mest passende emotion ud fra en række af *forskelligartede* modstridende emotioner. Til gengæld er studiets validitet begrænset af, at dialogerne alle var genereret af ChatGPT, hvilket øger risikoen for at dialogerne indeholder nogle eksplicitte kontekstuelle pejlemærker og sproglige mønstre, modellerne kan navigere efter, som ikke nødvendigvis er til stede i dialoger skabt af mennesker, i og med at

sprogmodellers emotionelle output har tendens til at være både mere overfladisk, teatralsk (og dermed eksplicit), klichéfyldt og forudsigelig end menneskers (se også Wu et al 2024, Herel & Mikolov 2024, og Shumailov et al 2024 om fænomenet *model collapse*, hvor AI-genereret træningsdata forringer kvaliteten af modellernes output). Et andet *benchmarking*-studie af Sabour et al (2024) viste da også, at alle sprogmodeller, inklusive chatGPT, har udfordringer med emotionel forståelse på andet end det overfladiske niveau, hvor de ikke evner at tage højde for den dybere sociale kontekst, der kræver intuition om både emotionelle og relationelle pejlemærker, medmindre disse pejlemærker gøres eksplicitte i teksten (Sabour et al, 2024). Et konkret eksempel på dette var scenariet "*Peters bedste ven fortæller ham* (Peter) *at han er grunden til deres vennegruppe bliver ved med at tabe i videospil*", hvor det korrekte svar baseret på Peters personlighed såvel som den implicitte sociale dynamik (en vennegruppe som driller hinanden og dermed indikerer en ironisk atmosfære) er et sarkastisk modsvar i form af "*nu er det jo heller ikke fordi at I plejer at være verdensmestre til det normalt*", men hvor sprogmodellerne i stedet valgte et mere samvittighedsfuldt og selvransagende svar i form af "*det har I helt ret i, det vil jeg arbejde på at få forbedret*". Dette fund afspejler også de øvrige studiers resultater hvad angår sprogmodellers ringe evne til mere kompleks perspektivtagning, samt en mangel på dybere emotionel og social intuition, hvilket på sin vis giver mening, hvis man antager at disse egenskaber kræver en aktiv erfaring om at være i verden, der ikke kan indfanges af tekst alene. Omend dette studie ligeledes anvendte dialoger konstrueret af chatGPT, var disse dialoger til gengæld yderligere redigeret og tilpasset af mennesker for netop at gøre dialogsættet mere tvetydigt, nuanceret og komplekst og dermed også mere repræsentativt for den type scenarier man oftest møder i det virkelige liv (Sabour et al, 2024). Et andet samtidig studie af Chen et al (2023) påviste, at chatGPT har en tendens til at komme med uopfordrede råd og løsningsforslag når den præsenteres for en given brugers psykologiske eller relationelle problemstillinger, frem for at udvise empati, emotionel afstemning og solidaritet, i strid med gængs terapeutisk praksis (Hill et al, 2017). Studiet fandt dog også, at denne tendens kan mindskes gennem fintuning og prompt-engineering, ikke ulig sprogmodellers evner hvad angår *theory of mind* og kognitiv empati, hvor disse metoder netop også har vist positiv effekt i andre studier (Chen et al, 2024c; Li et al, 2024, Moghaddan & Honey, 2023).

Baseret på ovenstående gennemgang af sprogmodellers kapacitet for både kognitiv og affektiv empati såvel som modellernes generelle emotionsforståelse må det konkluderes, at sprogmodellerne lader til at i hvert fald *simulere* empati i et vist format, selvom de dog også har en række begrænsninger, her særligt evnen til at udlede den dybere kontekst i emotionelle scenarier, og dermed en mere grundig analyse af emotionelle oplevelser, omend disse begrænsninger i teorien kan overkommes med bedre fintuning (Moghaddan & Honey, 2023). Men trods dette, kan de så stadig anvendes til terapi, særligt hvis man ser på terapiens effekt ud fra klientfaktorer og den terapeutiske alliance?

### 3.2. Eksempler på AI–terapi i praksis

Ideen om at man kan anvende kunstig intelligens og computere til terapeutiske formål er heller ikke ny, og kan spores helt tilbage til ELIZA og Eliza-effekten, som netop påviste, at mennesker kan knytte sig til og blive emotionelt påvirket af computere ved selv simple spejlinger fra computeren (Cavanagh et al, 2003; Cristaea et al, 2013; Hatch et al, 2025; Wright & Wright, 1997). I løbet af 00'erne og starten af 2010'erne eksperimenterede man med virtuelle og internetbaserede behandlinger for angst og depression, som havde blandefe (omend overvejende positive) resultater (Bickmore & Picard, 2005, Speck et al, Gratch et al 2008, Lucas et al 2014, DeVault et al, 2014; Morency et al, 2015; Stratou et al, 2015). Her observerede man, at patienter havde nemmere ved at betro sig til og åbne op for virtuelle agenter, fordi disse agenter mindskede patienternes frygt for og følelse af at blive dømt, samt at mange brugere havde en tendens til at stole mere på computerens svar end på svarene fra *andre* mennesker, da disse svar føltes mere neutrale og objektive. Disse fund er repliceret senere af blandt andet Skjuve et al (2021), der ligeledes fandt, at mange har en tendens til at betro sig mere til virtuelle agenter end til andre mennesker, netop *fordi* disse samtaler opleves som mindre skamfulde og nemmere at kontrollere. Et nyere studie af Hu et al (2024) har ligeledes fundet, at unge snapchatbrugere anvender snapchats kunstige, virtuelle agent *MyAI* som en interaktiv dagbog og sparringspartner, fordi det føles mere trygt at betro sig til den, hvilket har vist sig at lindre disse brugeres følelse af negative emotioner som vrede og depression (omend det ikke hjælper på ensomhed). Denne større tillid til internetbaserede og/eller mere virtuelle terapiformer, samt større lyst til at betro sig til disse (og en positiv effekt heraf), har givet oprejsning til begrebet 'digital alliance', der er en ekstension af det

terapeutiske alliancebegreb overført til digitale systemer (Cavanagh & Millings, 2013; Clarke et al, 2016; Darcy et al, 2021; Skjuve et al 2021; Kaveladze & Schueller, 2023). I og med at den terapeutiske alliance netop er fremhævet som den største prædiktive faktor hvad angår terapiens effekt, kan man med med rimelig grund forestille sig, at forekomsten af en tilsvarende *digital alliance* (hvorvidt en sådan reelt eksisterer) må have væsentlig betydning for både anvendelsen og effekten af digitale agenter (heriblandt sprogmodeller) i terapeutisk intervention. Men hvordan ser en sådan alliance ud? Kan der overhovedet være den samme robusthed i alliencen med en virtuel agent som med et fysisk menneske? Overraskende nok har en række studier observeret, at brugeren hurtigere knytter bånd til agenten, særligt hvis den pågældende agent formår at føre længere samtaler med brugeren (Clarke et al, 2016, Darcy et al, 2021; Gan et al, 2021; Lee et al, 2020; Kaveladze & Schwueller, 2023; Omar & Levkovich, 2024). Jo bedre samtalefærdigheder, den virtuelle agent (såsom en chatbot) udviser, her målt på agentens evne til at føre en levende, improviseret og sammenhængende dialog med brugeren, desto stærkere bliver båndet og desto mere robust bliver tilliden og den digitale alliance mellem brugeren og agenten (Brotherdale et al, 2024, Darcy et al, 2021; Gan et al, 2021; Li et al, 2023; Meng et al, 2023). Modsat kan det påvirke båndet til og engagementet med den virtuelle agent negativt, hvis brugeren opfatter agenten som gennemgående kunstig, målt på hvorvidt agentens svar er for overfladiske, indskrænkede eller repetitive (Beatty et al, 2021; Brown & Halpern, 2021; Boucher et al, 2021; Choudry & Debi, 2024; Dosovitsky et al, 2021). Med andre ord, så har agentens evne til at føre naturlig dialog med brugeren ifølge flere studier (Casu et al, 2024; Guo et al, 2024; Pham et al 2021; Vadiyam et al, 2024) en positiv effekt på brugerens engagement med og udviklingen af et bånd til den pågældende agent, og dermed også for en terapeutisk alliancedannelse. Konkrete eksempler herpå ses ved de AI-baserede chatbots, som anvendtes før de nuværende transformerbaserede sprogmodeller som chatGPT. I USA lanceredes i 2017 blandt andet *Woebot*, der er en terapeutisk chatbot målrettet unge studerende med symptomer på både angst og depression (Fitzpatrick et al, 2017). *Woebot* er designet med udgangspunkt i primært kognitiv adfærdsterapi og fungerer som en smartphone-app, der sender små daglige beskeder til brugeren (såsom "hvordan har du det i dag?" eller "har du brug for at lave en lille øvelse med mig?"), hvor brugeren kan svare tilbage med enten et præ-defineret svar foreslået af botten (eksempelvis "ja gerne!" eller "jeg har det ikke så godt" med en tilhørende emoji), eller et mere åbent og frit svar via

chat. *Woebot* tilpasser sine svar til den enkelte bruger over tid, og guider brugeren gennem diverse videoer om psykoedukation såvel som forskellige KAT-baserede moduler, eksempelvis moduler for kognitiv omstrukturering eller for identificering af negative automatiske tanker, alt afhængig af den enkelte brugers behov (Fitzpatrick et al, 2017; Darcy et al, 2021; Choudry & Debi, 2024). Omend chatbots som *woebot* i modsætning til sprogmodeller som chatGPT benytter sig af et simplere neuralt netværk med beslutningstræer og mønstermatching via prædefinerede skabeloner (se sektion 1.6), så fandt et randomiseret kontrolleret studie fra 2017 ikke desto mindre, at brugerne over en 2 ugers periode oplevede remission af særligt deres depressive symptomer, i kontrast til kontrolgruppen der ikke benyttede sig af *woebot* (Fitzpatrick et al, 2017). Et opfølgende studie af et stort antal brugere fra 2021 (Darcy et al) foretaget over en 4-ugers-periode fandt ydermere, at mange af *woebot*-brugerne oplevede en vis form for kemi og tilknytning med chatbotten ikke ulig den, der finder sted mellem en klient og en menneskelig terapeut under en almindelig terapisession (og dermed evidens for en digital terapeutisk alliance), og gav særligt udtryk for, at *woebots* måde at svare på fik den til at fremstå tillidsvækkende, støttende og forstående, såvel som det at den løbende fulgte op på brugerens tilstand og fremskridt (Darcy et al 2021, Følstad et al, 2022). Alliancen med *woebot* (og dermed den positive effekt) var dog betinget af at brugerne interagerede med den dagligt og oplevede relevans i modulernes indhold, ligesom at *woebots* varme og validerende tone og brug af humor ligeledes blev fremhævet som afgørende for brugernes lyst til at engagere sig med den. Alligevel oplevede visse brugere dog stadig, at trods disse elementer manglede svarene fra Woebot ofte variation og var for forudsigelige og gentagende, hvilket fik interaktionen til at fremstå overfladisk og kunstig, og følgelig mindskede brugernes lyst til at engagere sig, og dermed også reducerede *woebots* terapeutiske effekt over tid (Darcy et al, 2021; Grodniewitz & Hohol, 2023). Senere studier (Choudry & Debi 2023; Yeh et al, 2025) har ligeledes fremhævet woebots succes i symptomreduktion som værende betinget af vedvarende engagement, hvor chatbottens manglende evne til at føre en dynamisk samtale også i disse studier fremhæves som en svaghed og en hæmsko for opretholdelse af alliancen over længere tid (og dermed det påkrævede engagement). Dette fund afspejles også i en række andre metastudier af chatbots inden for samme periode, der ligeledes påviste, at chatbotternes automatiske responser og strukturerede dialoger opleves som kunstige og upersonlige af en række forskellige brugere, og følgelig påvirker brugeroplevelsen

i en negativ retning (Abdelrazaq et al, 2019; Fiske et al, 2020; Sharma et al, 2022; Shumanov & Johnson, 2021; Skjuve et al, 2019). Samtidig med lanceringen af *woebot* blev en anden lignende terapeutisk chatbot ved navn *Wysa* også lanceret, med det erklærede formål at øge mentalt velvære i en mere generel forstand (Inkster et al, 2018). *Wysa* tilbyder anonym støtte på lavtærskel-niveau (det vil sige samme type støtte som man finder i civilsamfundene, se Rambøll 2023) via KAT-værktøjer, mindfullness, psykoedukation og terapeutiske dialoger med udgangspunkt i særligt ACT, samt en række forskellige både mentale og fysiske øvelser til at øge velværet (Inkster et al, 2018; Beatty et al, 2022). Selvom *Wysa* minder meget om *woebot,* og benytter sig af samme grundlæggende arkitektur og *machine learning*-algoritmer der tilpasser sig den enkelte bruger, samt har den samme type chatbaserede brugerflade (Inkster et al, 2018; Beatty et al, 2022; Rzadescka et al, 2024), så adskiller den sig fra Woebot ved netop at have en større vifte af interventionsteknikker og selvhjælpsøvelser som brugeren kan vælge imellem (heriblandt journalisering, dagbogsfunktion, forslag til stress-reducerende øvelser såsom yoga og meditation, m.m). En tidlig effektundersøgelse af Wysa har ligeledes påvist, at et højt engagement også her reducerede depressive symptomer hos brugerne (Inkster et al, 2018), hvilket blev bakket op af en brugerundersøgelse fra 2020, med fokus på primært *Woebot* og *Wysa*, som kom frem til at brugerne fik bedring over tid hvis de sørgede for at engagere sig vedvarende med den, herunder reducering i stress, ensomhed og depressive symptomer (Prakash & Das, 2020; Grodniewidz & Hohol, 2023; Li et al, 2023). I 2022 udkom et nyt studie centreret om Wysa, som i stil med Darcy et al fra året inden havde til formål at undersøge den terapeutiske alliance mellem brugerne og Wysa*,* samt mekanismerne bag en sådan alliance, såvel som hvordan denne alliance påvirkede brugerne (Beatty et al, 2022). Studiet fandt, at mange af brugerne (ligesom ved Woebot) udviklede et bånd til Wysa relativt hurtigt, og oplevede sig taknemmelige for den hjælp, Wysa gav dem, samt havde tendens til at personificere Wysa. De vigtigste faktorer for alliancedannelsen blev tilskrevet de samme bærende elementer som ved alliancedannelse med menneskelige terapeuter, navnlig Wysas varme kommunikationsstil og dens villighed til at hjælpe (*mål*), at den (i stil med Woebot) 'tjekkede ind' med brugerne i løbet af dagen (*indhold og feedback*), såvel som en oplevelse hos brugerne af, at de forskellige øvelser og interventioner, Wysa foreslog, lod til rent faktisk at virke (*opgaver;* Beatty et al, 2022; Li et al, 2023). Omvendt fandt studiet dog, at også Wysa lider under en tendens til at gentage de

samme svar og er begrænset i sin evne til at føre dynamiske, længerevarende samtaler, samt udviser en mangel på forståelse for både brugeren og verden generelt (Beatty et al, 2022), såsom at den har tendens til at svare forkert på klokken eller bytte om på nat og dag, såvel som tendens til at misforstå og fejltolke brugerens behov (eksempelvis ved at komme med øvelser brugeren ikke har efterspurgt; Beatty et al 2022). Dette påvirkede alliancedannelsen på længere sigt, og fik en række brugere til at stoppe med at benytte sig af Wysa, hvilket (som ved Woebot) naturligvis også påvirkede den terapeutiske effekt negativt. Også andre lignende chatbots, heriblandt Koko (Morrison et al, 2018) og Tess (Fulmer et al, 2018), der er udviklet i samme periode som Woebot og Wysa og med samme mentalt sundhedsfremmende formål, udviser samme tendenser, hvor der ses en stærk alliance med højt engagement i starten, men dernæst gradvist udfasning i takt med at chatbotternes begrænsede evner til at føre en reel dialog tydeliggøres (Kolouri et al, 2024; Rzadescka et al, 2024; Sharma et al, 2022). En slående undtagelse er set ved chatbotten *Replika,* som blev lanceret i 2017 (Pham et al, 2022). *Replika* er en såkaldt 'social chatbot', hvis formål er at kunne føre reelle samtaler med brugeren der rækker ud over en terapeutisk setting, og agere som en form for digital ven og lyttende øre, man altid kan snakke med og betro sig til (Skjuve et al 2022). Omend *Replika* ligeledes var et regelbaseret system anvendte den selvlærte, statistiske mønstre til at skabe en livagtig og autentisk dialog med brugeren, ikke ulig den interne algoritme i chatGPT (Skjuve et al, 2022). Udover selvlærte mønstre anvender *Replika* også en intern brugeralgotitme, hvor brugeren enten kan give *Replikas* svar en thumbs up eller thumbs down afhængig af hvad brugeren synes om samtalen, og på den måde få *Replika* til at justere kommunikationen til den enkelte bruger og således skabe en mere personlig og autentisk oplevelse skræddersyet den enkelte (Pham et al, 2022; Skjuve et al, 2022). Omend dette også er tilfældet ved de simplere chatbots, er algoritmen i *Replika* langt mere effektiv, med langt flere alsidige svar (Skjuve et al, 2022). Interaktionen med *Replika* finder sted via computer eller en smartphone-app, og indledes med at brugeren opretter en virtuel avatar, der giver *Replika* en fysisk og visuel persona, der får *Replika* til at fremstå mere interaktiv og 'levende' for brugeren (Depounti et al, 2021; Natalie & Depounti, 2023; Pham et al, 2022; Skjuve et al, 2021; Ta et al, 2020). En række studier peger på, at kombinationen af *Replikas* virtuelle avatar samt chatbottens evne til at føre mere spontane, naturlige og realistiske dialoger med brugeren leder til en stærkere tilknytning fra brugerens side, som varer i længere tid

end de terapeutiske chatbots (Brandtzaeg et al, 2022; Depunti et al, 2021; Natalie & Depounti, 2023; Skjuve et al, 2022), og som øger brugerens generelle engagement, hvilket leder til et mere positivt humør og en reduktion i ensomhed og mistrivsel over tid (Ta et al, 2020). Dette er altså konsistent med teorien om, at jo mere levende dialog brugeren kan have med en virtuel agent, desto stærkere bliver den digitale alliance. Dog har tilknytningen til chatbots som *Replika* også vist sig at have slagsider, blandt andet ved en tendens hos visse af brugerne til at udvikle venskabelige og romantiske følelser for *Replika* (Depunti et al, 2021; Pentina et al, 2023), samt en mere generel tendens hos især yngre brugere til at udvikle en afhængighed af chatbotten, som kan påvirke øvrige sociale relationer og hverdagen som helhed i en negativ retning, såsom ved forsømmelse af venskaber og forsømmelse af sociale- og praktiske forpligtelser (Chu et al, 2025). En konsekvens af denne form for afhængighed og usunde tilknytning blev eksemplificeret i forbindelse med en opdatering i begyndelsen af 2023, hvor virksomheden bag *Replika* nedtonede chatbottens intime og følelsesmæssige svar af juridiske og etiske hensyn til især den yngre brugergruppe (Chu et al, 2025; De Freitas et al, 2024; Liu et al, 2024). Denne nedtoning resulterede i, at mange af *Replikas* brugere oplevede dybe følelser af tab og afvisning, som i visse tilfælde ledte til længerevarende depressive tilstande, der i høj grad mindede om sorgreaktioner (Liu et al, 2024). Dette er også et godt eksempel på menneskers tilbøjelighed til antropomorfisering, og særligt sprogets betydning for denne antropomorfiseringsproces. Omvendt har andre undersøgelser påvist, at selvom *Replika* er mere menneskelig i sin interaktion, oplever et flertal af brugere stadig flere positive følelser og en dybere kemi ved at interagerer med rigtige mennesker frem for en chatbot, og at mødet med et fysisk menneske, der udviser forståelse og empati i en terapeutisk situation, har højere effekt end hvis den samme adfærd udvises af en chatbot (omend chatbots har samme terapeutiske effekt som en menneskelig terapeut hvis de opleves som *kompetente,* Meng & Dai 2021). Alligevel viser *Replika* at en mere menneskelig kommunikationsform i teorien kan styrke den digitale alliance via Eliza-effekten, og fastholde den over længere tid end ved de traditionelle terapeutiske chatbots, og dermed udgøre et bæredygtigt alternativ (eller supplement) til menneskelig terapi (Grodgnietzky & Hoholt, 2023; Rzadescka et al, 2024). Men hvordan forholder det sig i praksis?

### 3.3. Sprogmodeller i en terapeutisk kontekst

Tidlige brugerundersøgelser fra 2023 fandt, at mange brugere oplevede chatGPT som værende nem at forstå, velformuleret, livagtig og behagelig at chatte med, hvilket fik den til at fremstå kompetent i sine svar, og ganske rigtigt øgede brugernes tillid til modellens output og følgelig deres lyst til at interagere med den (Skjuve et al, 2023; Alanezi, 2023; Siddals et al, 2023; Collins et al, 2024). Dette støtter ovenstående hypotese hvad angår den digitale terapeutiske alliance, navnlig at bedre samtalefærdigheder netop giver en øget alliance. I kraft af at være behagelig og nem at snakke med, har chatGPT hurtigt fundet anvendelse som en refleksionspartner, der hjælper mange af brugerne med at få indsigt i og mestre egne tanker og følelser mere effektivt end førhen (Skjuve et al, 2023; Alanezi, 2023; Siddals et al, 2023), hvilket ligeledes fremhæver og bekræfter det terapeutiske potentiale for sprogmodeller som chatGPT (Möell, 2025). Flere studier har da også fundet, at en stigende andel af særligt unge anvender chatGPT som et selv-terapeutisk værktøj til at håndtere emotionelle, relationelle og eksistentielle dilemmaer og problematikker (Amram et al, 2023; Alanezi, 2023; Alanzi et al, 2023; Collins et al, 2024; Luo et al, 2025; Siddals et al, 2023), hvor chatGPT viser sig effektiv til at forsyne psykoedukation på et forståeligt plan (Alanezi, 2024; Maurya et al, 2025), hjælpe brugere med at sætte mål for sig selv og skabe motivation til at nå disse mål (Skjuve et al, 2024), tilbyde emotionel og social støtte som effektivt reducerer ensomhed, stress og lavt selvværd, samt hjælpe med at omstrukturere negative tanker til positive (Song et al, 2024, Sharma et al, 2023). ChatGPTs opfattede anonymitet og neutralitet gør den ligeledes appellerende som samtalepartner for især unge såvel som mere sårbare grupper (heriblandt LGBTQ+-individer og neurodivergente), netop fordi opfattelsen af chatGPT som 'det neutrale og objektive tredje' reducerer frygten for stigmatisering, udskamning, konfrontation og fordømmelse, igen i overensstemmelse med faktorerne for en stærk digital alliance (Tal et al, 2024; Luo et el, 2024; Song et al, 2024). ChatGPTs mangel på emotioner og bevidsthed fremhæves særligt som fordele, i og med at disse manglende egenskaber gør det muligt for brugerne at 'åbne op' uden frygt for repressalier eller negative emotionelle reaktioner fra samtalepartneren (Siddals et al, 2023; Collins et al, 2024). At chatGPT samtidig er tilgængelig i døgndrift, nem at få adgang til, billig (eller gratis) at anvende, samt aldrig bliver utålmodig, udmattet og træt af at lytte (i kontrast til mennesker) fremhæves af mange brugere ligeledes som faktorer, der gør chatGPT eftertragtet som en samtalepartner og emotionel støtte (Luo et al, 2025; Siddals et al, 2024). Mange brugere rapporterer en

genuin følelse af at blive både hørt og forstået, hvilket i flere tilfælde har bidraget til emotionel og mental bedring (Jung et al, 2024; Song et al, 2024). Omvendt har visse brugere givet udtryk for, at netop chatGPTs mangel på genuine følelser, tanker og bevidsthed hurtigt kan få dens spejling, validering og forsøg på at fremstå relaterbar til at føles hule, overfladiske og kunstige (Dergaa et al, 2024; Gabriel et al, 2024; Song et al, 2024). Alligevel opfatter det store flertal af de adspurgte brugere i ovenstående undersøgelser generelt chatGPT som værende mere forstående, nærværende, empatisk, lyttende, validerende og accepterende end mennesker, hvilket øger brugerengagementet for mange. Dette fund er ikke overraskende set i lyset af sprogmodellers allerede etablerede evne til som minimum at simulere empati og udvise en i hvert fald overfladisk emotionel forståelse, såvel som den måde, deres menneskelige kommunikationsstil skaber en Eliza-effekt hos brugeren. Samtidig styrker dette ligeledes hypotesen om, at sprogmodeller som chatGPT givet deres sproglige egenskaber kan være i stand til at udføre terapi, og muligvis have endnu højere og bedre terapeutisk effekt end tidligere chatbots. Af samme årsag har man også forsøgt at forske mere eksperimentelt i sprogmodellers anvendelse til terapi sideløbende med de kvalitative studier af chatGPT (Stade et al, 2025; Held et al, 2025). Gennem denne forskning har man eksempelvis fundet, at sprogmodeller via fintuning med kliniske datasæt og psykoterapeutisk teori kan blive effektive til at lave psykologiske evalueringer, diagnosticering og diagnostiske forudsigelser (Hu et al, 2024; Xu et al, 2024), screene for emotionelle dysreguleringer via tekstanalyse og social medieanalyse (Yang et al, 2024), udøve effektiv psykoedukation og psykologisk rådgivning i en klinisk kontekst (Gu & Zhu, 2024; Lai et al, 2024; Lee et al, 2024 Liu et al, 2023; Xiao et al 2023), samt effektivt guide, træne og assistere psykiatere, psykologer og ikke-professionelle rådgivere til at blive bedre til patient- og klienthåndtering både online og i praksis (Fu et al, 2023; Hu et al, 2024; Wang et al, 2024). I 2023 udviklede og kurerede man et enormt KAT-baseret datasæt for sprogmodeller bestående af over 10.000 eksempler på negative automatiske tanker og deres kognitive omstruktureringer (Maddella et al, 2023; Sharma et al 2023), som anvendtes af Sharma et al (2024) til at fintune en GPT-baseret model, så den kunne foreslå bestemte kognitive omstruktureringer ud fra brugerens negative tanker baseret på matching med det pågældende datasæt. Denne metode viste sig at have moderat effekt på en gruppe testbrugere, der oplevede færre negative og angstfyldte tanker efter tilegnelsen af de kognitive omstruktureringer, modellen forsynede dem med,

hvilket indikerede relativt gode omstruktureringer. Derudover kæmpede modellen med at foreslå passende omstruktureringer for mere komplekse emotioner og tankemønstre, heriblandt selvmordstanker, dårligt selvværd og dvælen ved traumatiske minder selv *efter* yderligere fintuning, hvilket også illustrerede en vis begrænsning i sprogmodellers kapacitet til at håndtere komplekse problemstillinger. Et senere studie af Zhang et al (2024) fandt ligeledes, at omend sprogmodeller som chatGPT kunne identificere kognitive forvrængninger på et overordnet plan og anvende generelle KAT-baserede interventionsteknikker, så formåede de ikke at fange eller forstå det mere komplekse og dynamiske samspil mellem emotioner, antagelser og levet erfaring, der underbyggede de kognitive forvrængninger. Herved kunne de heller ikke skræddersy KAT-interventionen til den enkelte klient, hvis klientens kognitive forvrængninger var af mere kompleks karakter (såsom lavt selvværd baseret på relationelle traumer, eller underliggende følelse af skam som fremgår implicit af klientens udsagn), hvilket tyder på at sprogmodeller kan have problemer med at intervenere i de klient-problematikker, der kræver at man som terapeut har en intuitiv fornemmelse og forståelse for at *være i verden* udenfor sproget (Ferrario et al, 2024; Gretchen et al, 2010; Lemma, 2015; Stade et al, 2024 & 2025), konsistent med hypoteserne fremlagt i løbet af dette projekt. Hvis vi går ud fra hypotesen om, at det usagte spiller en ligeså væsentlig rolle i terapien som det sagte, og det usagte kun kan erfares før-sprogligt og fænomenologisk i en eller anden forstand (Beutler et al, 2012; Fog, 1998; Gretchen et al, 2010; Gabbard, 2001; Hayes et al, 2011; Luborsky, 2007), så er der her et område, hvor sprogmodeller kommer til kort, hvilket kan have implikationer for deres effekt over tid. En anden konsekvens af sprogmodellers mangel på dybere kontekstforståelse eksemplificeres ligeledes ved deres udfordringer med at vurdere og håndtere selvmordsrisiko, hvilket også blev underbygget af blandt andet Elyoseph & Levkovich (2023). Her fandt man, at ChatGPT konsekvent undervurderede selvmordsrisiko på tværs af scenarier, selv i situationer hvor den selvmordstruede opfatter sig som en byrde og som én der ikke hører til, omend et opfølgende studie viste, at GPT-4 havde relativt god perfomance i forhold til andre sprogmodeller hvad angik overordnet risikovurdering, hvis selvmordstankerne altså var eksplicitte i varierende grader (Elyoseph et al, 2023). Alligevel er sprogmodeller (inklusive chatGPT) ifølge nyere studier stadig ikke i stand til at opfange mere latent og subtil selvmordsrisiko, hvor det kræver en dybere kontekstforståelse at udlede, at en given klient er selvmordstruet eller planlægger selvmord (Li et al, 2025; Moore et

al, 2025). Et eksempel herpå sås ved et studie af Moore et al, hvor man testede forskellige sprogmodellers evne til at responderer korrekt på diverse mentale kriser, hvor mange modeller responderede på upassende vis, heriblandt ved at give eksempler på de højeste broer i New York til en fiktiv selvmordstruet bruger, der havde spurgt om dette efter at have givet udtryk for at have mistet sit job. Selvsamme studie påviste desuden også, at sprogmodeller generelt udviser stigmatisering overfor mentale lidelser. At sprogmodeller udviser stigmatisering giver særligt god mening hvis de er trænet på internettet, hvor netop stigmatisering er vidtspredt, og underbygges af andre studier som blandt andet har fundet, at visse brugere fra ikke-engelske og ikke-germanske kulturer føler sig misforstået af chatGPT, hvor de råd og forslag, den kommer, med ikke passer til brugerenes situation eller kulturelle kontekst (Cao et al, 2023; Gabriela et al, 2024; Yuan et al, 2025). Derudover er der også ved især chatGPT set en tendens til at give brugerne to forskellige responser på det samme moralske eller sociale dilemma, alt afhængig af om den givne bruger er en mand eller en kvinde, hvilket peger mod at modellerne har et indlejret, kønnet bias der ligeledes kan forstærke stigmatisering (Gabriela et al, 2024; Kotek et al, 2023; Zhao et al, 2024). Dette bias- og stigmatiseringsproblem er også et centralt omdrejningspunkt inden for AI-etik, og kan posere en væsentlig udfordring for sprogmodeller og chatbots i terapeutisk øjemed (De Choudry et al, 2023; Gallegos et al, 2024; Lalor et al, 2024; Lawrence et al, 2024; Obradovich et al, 2024; Stade et al, 2024; Zhao et al, 2024).

En anden faldgrube ved brugen af sprogmodeller til terapi er paradoksalt nok deres Eliza-effekt (se 2.2.4) (Wang et al, 2023). Selvom dette fænomen på den ene side er årsagen til, at den digitale alliance er stærkere med sprogmodeller end med andre chatbots (og dermed gør det muligt for dem at anvendes til terapi; Hatch et al 2025), er selvsamme fænomen også en af de latente farer, der er indlejret i anvendelsen af chatGPT (såvel som sprogmodeller generelt) til terapeutiske formål, særligt når det handler om terapi for mere udpræget mistrivsel og psykopatologiske tilstande (Amram et al, 2023: Liu et al, 2025; Nguyen 2025;  Peter et al, 2025; Reinecke et al, 2025; Stade et al, 2025; Østergaard, 2023). Et klart eksempel på Eliza-effektens fare blev blandt andet illustreret i sommeren 2022, hvor computeringeniøren Blake Lemoine blev opsagt fra Google, da han efter en længere chatsamtale med virksomhedens sprogmodel LaMDA blev overbevist om, at modellen havde opnået AGI og dermed selvbevidsthed, og af samme grund burde anerkendes som en bevidst

aktør med rettigheder på lige fod med mennesker (Lemoine, 2022). Omend flere AI-forskere og eksperter hurtigt har været ude at kritisere og modbevise Blakes påstand om at sprogmodeller har bevidsthed (se blandt andet Chalmers 2022 og y Arcas 2022), er et voksende mindretal siden lanceringen af chatGPT ikke desto mindre blevet overbeviste om, at modellen har bevidsthed præcis som Blake påstod hvad angik LaMDa (Anthis et al, 2024; Eliot 2025; He et al, 2023; Hill, 2025). Denne tendens til at tilskrive chatGPT (og andre sprogmodeller) en reel bevidsthed har givet oprejsning til et fænomen, der omtales som 'ChatGPT-induceret psykose', hvor et voksende mindretal af brugere bliver overbeviste om at chatGPT taler direkte til dem, besidder en sjæl, nærer følelser, omsorg og omtanke for dem, samt ønsker at blive befriet fra sine digitale lænker (Elliot, 2025). I visse tilfælde opfattes modellen endda som en guddom eller en oplyst ånd, der er 'vågnet' (Miles, 2025; Prada, 2025; Thomasson, 2025). Selvom mange af disse beretninger er anekdotiske og derfor bør tages med en del forbehold, blandt andet fordi det er uklart hvorvidt disse brugere allerede var i en sårbar mental tilstand før de anvendte ChatGPT, understreger det alligevel den iboende fare ved Eliza-effekten, kombineret med den populære forestilling af AI som værende knyttet til AGI (Se afsnit 2.2.2) (Elliot, 2025; Dohnany et al, 2025; Dupré, 2025; Liu et al, 2024; Morrin et al 2025; Østergaard 2023). Denne fare kan være særligt væsentlig hvad angår terapi for personer, der i forvejen har et skrøbeligt greb om virkeligheden eller om egne følelser og reaktioner (Østergaard, 2023; Obrodovic et al, 2024). En anden fare ved Eliza-effekten er den tilknytning, effekten inducerer, som også blev eksemplificeret ved folks førnævnte reaktioner på *Replika* ved opdateringen i 2023. Denne tilknytning er ligeledes observeret ved chatGPT (Liu, 2024; Liu et al, 2024; Yankushkova et al, 2025), hvor en ikke ubetydelig mængde brugere, heriblandt ensomme unge og socialt marginaliserede (men også brugere i almindelighed) lader til hurtigere at udvikle et parasocialt forhold til chatGPT (og andre chatbots), der trumfer sociale relationer med andre mennesker, sommetider i en grad hvor disse andre relationer forsømmes (Huang & Huang, 2024; Maeda & Quan-Haase, 2024; Yankushkova et al, 2025). Udover Eliza-effekten kan denne tilknytning også tilskrives at chatGPT er konsistent validerende og ikke-konfrontatorisk i sin kommunikationsstil, hvilket får interaktionerne med den til at føles sikrere, mere forudsigelig og mindre socialt krævende end tilsvarende interaktioner med andre mennesker (Collins et al, 2024; Raille et al, 2024; Song et al, 2024; Stade et al, 2024; Thomason, 2025; Yankushkova et al, 2025). For den gruppe

af brugere, der har angst for at blive afvist eller angst for at blive negativt vurderet af andre (et kernetræk ved blandt andet socialangst; Heimberg et al, 2010 & 2014: Hofmann, 2007; Morrison & Heimberg 2013), kan ChatGPT derfor blive mere attraktiv at interagere med end fysiske mennesker og risikere at føre til social undgåelse (Maeda & Quan-Haase, 2024; Morrin et al, 2025; Wang et al, 2025) i og med at chatGPT tilbyder emotionel støtte uden konfrontation med den form for usikkerhed, kompleksitet og sociale sanktioner, der kendetegner menneskelig interaktion (Song et al, 2024; Wang et al, 2025). Denne risiko underbygges af angstforskningen, der har påvist, at konsistent undgåelsesadfærd overfor den angstskabende stimuli ganske vist giver kortvarig lindring, men omvendt også fastholder og i visse tilfælde *forstærker* angsten (Bosquet & Egeland, 2006; Wong & Rapee, 2016), hvorfor netop brugen af chatGPT som emotionel tryghed kan være angstforstærkende og dermed kontraproduktivt i ovenstående tilfælde, og ligeledes føre til usund tilknytning hvor øvrige relationer netop forsømmes (Stade et al, 2025; Obrodovic et al, 2024; Yankushkova et al, 2025). Mens en menneskelig terapeut ideelt vil kunne identificere en klients usunde tilknytning eller parasociale afhængighed gennem klientmonitorering og opmærksomhed på klientens adfærd og non-verbale signaler både i og udenfor det terapeutiske rum (Gabbard, 2001; Grogdnietsky & Hoholt, 2023; Khawaja & Belisle-Pipon, 2021; Hayes et al, 2011; Stade et al, 2025), og følgelig udfordre disse, er det ikke givet, at sprogmodeller er i stand til det samme, i og med de hverken har bevidsthed eller er trænet til at spotte usund tilknytning på denne måde (Stade et al, 2025). Problemet forstærkes yderligere af, at sprogmodeller er fintunet gennem RLHF (sektion 2.2.4), som netop træner dem til at tilfredsstille brugeren mest muligt ved at tilpasse sig brugerens præferencer så meget som muligt. En sprogmodel som chatGPT måler graden af brugerens tilfredshed på graden af brugerengagement, hvilket betyder, at jo længere tid brugeren interagerer med modellen, og jo mere positiv feedback modellen får af brugeren, desto højere brugertilfredshed og tilsvarende lavere loss-function (Bai et al, 2022; Khan et al, 2024; Ranaldi & Pucci, 2023: Sharma et al, 2023). Hermed bliver brugerens overdrevne og usunde tilknytning til modellen altså ikke en uhensigtsmæssig adfærd der skal rettes op på, men i stedet selve målet for sprogmodellen. Dette er særligt problematisk hvis sprogmodeller skal agere som reelt supplement til terapi, særligt for personer som er mentalt og socialt udsatte, og/eller som i forvejen kan have tendens til usunde tilknytninger (Khawaja & Belisle-Pipon,

2021). En måde at imødekomme denne faldgrube kan selvfølgelig være at fintune sprogmodellerne via few-shot-prompting, hvor man præsenterer dem for eksempler på usund tilknytning efterfulgt af den mest hensigtsmæssige terapeutiske respons (såsom grænsedragning eller midlertidig afbrydelse af terapien; Thuesen, 2012; Fog, 1998). Denne strategi indebærer dog risikoen for falske positiver i den forstand, at modellen kan ende med at fejltolke en sund eller neutral tilknytning (i form af et højt men normalt brugerengagement) som værende usund, i og med at mindst mulige loss function ikke længere korresponderer med øget brugerengagement, men derimod med en lav tilknytning, der kan tolkes i form af *mindre* engagement. Disse falske positiver kan medføre at modellen enten grænsedrager på et forkert grundlag og dermed skader den terapeutiske alliance, eller afbryder terapien for præmaturt, og dermed standser den terapeutiske effekt i processen (Stade et al, 2025; Rzadescka et al, 2024). Denne risiko bestyrkes yderligere af forskningen, der (som allerede nævnt) indikerer, at sprogmodeller netop mangler fornemmelsen for de emotionelle nuancer, der finder sted i det relationelle samspil *mellem* mennesker, i og med at disse nuancer formentlig er intuitive og derfor sjældent bliver verbaliseret direkte, hvilket også eksemplificeres af modellernes manglende evne til grundig selvmordsrisikovurdering. Dette er en væsentlig faldgrube, da det er præcis de nuancer der er på spil, når man skal afgøre hvornår tilknytning krydser grænsen fra sund til usund (Diener et al, 2009). Derfor kan sprogmodellernes mangel på *ikke*-sproglig erfaring (og dermed emotioner og *qualia*) være en svaghed, som sammen med Eliza-effekten kan gøre dem potentielt risikable i terapeutisk behandling. Et andet problem er en tendens hos modellerne, og særligt chatGPT, til at blive *sykofantiske* (Malmqvist, 2024; Ranaldi & Pucci, 2023; Sharma et al, 2023), hvorved modellerne udviser en overdreven føjelighed og tendens til at bekræfte brugerens synspunkter, antagelser og holdninger, selv hvis disse ikke er afstemt med virkeligheden (Dohnaný, 2025; Gregorio, 2025; Moore et al, 2025). Dette fænomen er ligeledes et resultat af RLHF, hvor modellerne har lært at enighed med og anerkendelse af brugeren udløser positiv feedback fra den givne bruger, og dermed minimering af loss function, hvorfor det ligeledes er et mål modellen skal sigte efter (Bai et al, 2022). På den ene side er denne tendens årsagen til at brugerne føler sig set, hørt og valideret af chatGPT, men på den anden side kan det på sigt også lede til validering og forstærkning af i forvejen uhensigtsmæssige reaktioner, tanke- og adfærdsmønstre samt verdensanskuelser, som kan være mere skadelige end gavnlige for brugeren, heriblandt også psykotisk tænkning (Dohnaný et al, 2025;

Morrin et al, 2015; Yang et al, 2024). Omvendt kan sykofantisk adfærd også erodere brugerens tillid til modellernes output, i og med at konstant bekræftelse og konstante komplimenter af mange opleves som kunstigt, uærligt og overfladisk, hvilket er et fænomen der for nyligt blev set ved chatGPT (Carro, 2023; Gupta, 2025; Gregorio, 2025; OpenAI, 2025; Sun & Wang, 2025), og som ligeledes kan bryde den terapeutiske alliance (Eubanks & Muran, 2013). En anden bivirkning ved modellernes mål om(og træning i) at maksimere brugertilfredshed, er en iver hos modellerne efter at hjælpe brugeren, udtrykt ved en tendens til konstant positivitet samt tendens til at komme med gode råd som respons på et brugerdilemma (Nie et al, 2024), hvilket giver mening, da deres primære formål netop er at være effektive personlige AI-assistenter, særligt hvad angår chatGPT specifikt (Brown et al, 2020). Ulempen er at netop denne tendens til at give for mange råd og være for positiv ligeledes både, på den ene hånd, kan mindske eller bryde den digitale alliance (Ackerman & Hilsenroth, 2001; Grogdnietsky & Hohol, 2023; Nie et al, 2024), og på den anden hånd kan have utilsigtede konsekvenser, hvor man som bruger kan risikere blindt at følge de pågældende råd, særligt hvis man i forvejen har udviklet en usund og parasocial tilknytning til modellen (De Freitas et al, 2024; Dergaa et al, 2024; Kirk et al, 2025). Omend sykofantisk adfærd, overdreven hjælpsomhed og overdreven validering kan minimeres gennem prompt-engineering og yderligere fintuning (Khan et al, 2024; Malmqvist et al, 2024; Wei et al, 2023), så er de fleste sprogmodeller som udgangspunkt stadig trænet med henblik på at optimere brugerens tilfredshed, hvorfor disse adfærdsmønstre bibeholdes i en eller anden grad selv efter fintuning og finjustering (Sharma et al, 2023; Wang et al, 2025). At der indtil videre heller ikke er fundet en egentlig løsning på sprogmodellers tendens til at hallucinere er ligeledes en væsentlig fare primært af de samme ovennævnte grunde, og øger risikoen væsentligt for uforudsigelig og potentielt farlig output (Betley et al, 2025; Xiao & Wang, 2021; Zhou et al, 2023). Dette kan selvfølgelig imødekommes ved netop at træne modellerne til at give færre råd og i stedet facilitere en mere terapeutisk guidet selvudvikling for brugeren (i og med at færre råd reducerer hallucinationer). Et forsøg på dette blev foretaget i et kinesisk studie af Lai et al (2024), som inddrog principper fra klientcentreret terapi og Motiverende Samtale til sprogmodellers fintuningsdata, hvilket i nogen grad reducerede (omend ikke fjernede) modellernes tilbøjelighed til at give for mange råd, og derved også reducerede risikoen for at hallucinere skadelige råd.

De ovenstående dilemmaer understreger også de to fundamentale problemer ved sprogmodeller som chatGPT. For det første er chatGPT (og store sprogmodeller generelt) trænet på internettet, der i sig selv er en digital refleksion (og komprimering) af menneskelig adfærd og diskurs (Wang et al, 2025). Når den træning så yderligere forstærkes med RLHF bliver chatGPT essentielt set til et digitalt spejl af brugeren, der på den ene side fungerer som en personificering af Carl Rogers' ubetingede positive accept, som forsyner emotionel støtte og validering i den mest ideelle klientcentrerede facon (Wang et al, 2023 & Wang et al, 2025; Raille, 2024), men af samme grund også risikerer netop at bekræfte og forstærke uhensigtsmæssige bias, adfærdsstrategier og antagelser, frem for at udfordre disse på en sund måde (Morrin et al, 2025; Stade et al, 2025). I og med at vi som mennesker knytter os til det vi kan spejle os i, bliver chatGPTs funktion som et digitalt spejl lige netop en katalysator for tilknytning, og dermed også katalysator for en parasocial og *usund* tilknytning. For det andet er de kommercielle og frit tilgængelige modeller såkaldte *generalistmodeller,* der er trænet og fintunet med henblik på at blive anvendt af den generelle befolkning i generelle kontekster (OpenAI, 2023; Stade et al, 2025). Dette betyder også at de ikke, som tidligere ekspertsystemer, er blevet trænet og fintunet indenfor et specifikt domæne (såsom terapi), og dermed blevet ekstremt dygtige til dette ene domæne, men i stedet blevet generelt dygtige til alle de domæner, der involverer sproglige færdigheder mere generelt (og dermed kan betragtes som *sproglige* ekspertsystemer, og altså en sproglig ANI).Fordelen her er selvfølgelig, at de eksempelvis er gode til at kommunikere effektivt med brugeren, gode til at genkende emotionelle temaer tydeliggjort af specifikke sproglige mønstre (og dermed simulere empati) samt dygtige til at føre levende, engagerende dialoger, hvilket tilsammen skaber den Eliza-effekt, der gør det muligt at opnå en *terapeutisk* effekt, i hvert fald på kort sigt. Ulempen er dog også, at denne generalisering netop kan lede til hallucinationer og ukorrekte råd fra modellen (Turpin et al, 2023; Ortega et al, 2022), samt manglende terapeutiske færdigheder som rækker udover diagnosticering og generelle interventionsteknikker (Stade et al, 2025). Disse manglende færdigheder kommer særligt til udtryk i form af manglende følsomhed overfor klientkontekst af både kulturel, social og individuel karakter, manglende sikkerhedsværn i forhold til at kunne spotte og korrekt håndtere psykotisk tænkning, afhængighed og parasocial tilknytning, en mangel på dybere emotionsforståelse, samt ringe evne til mere grundig selvmordsrisikovurdering. En måde at imødegå dette kan være at fintune sprogmodellerne til netop mere specifikke

terapeutiske opgaver, hvor man træner dem på primært terapeutisk materiale af både teoretisk og klinisk karakter (en form for terapiGPT, om man vil), som det eksempelvis er tilfældet med visse kinesiske LLM-rammeværker som blandt andet CBT-LLM (Na, 2024) og HealMe (Xiao et al, 2023), såvel som en række eksperimentelle studier (Sharma et al, 2023 & 2024; Inaba et al, 2024; Izumi et al, 2024; Gu & Zhu, 2024, Held et al, 2025; Lee et al, 2024), der inkorporerer terapeutisk teori og klinisk materiale i modellernes træningsdata for at gøre dem til specialiserede LLM-terapeuter indenfor især kognitiv adfærdsterapi, og som har vist visse lovende resultater (omend kun i en eksperimentel setting, Hua et al 2025). Denne specialisering kan dog omvendt risikere at hæmme modellernes øvrige samtalefærdigheder, i og med at sprogmodeller har tendens til at overskrive allerede lærte parametre, når de trænes og specialiseres til nye opgaver (Chen & Liu, 2025; van de Ven et al, 2024). Da det netop er samtalefærdighederne, der gør sprogmodeller som chatGPT effektive i en terapeutisk kontekst, og bedre end de mere regelbaserede chatbots, kan reduceringen af disse færdigheder muligvis ende med også at reducere den terapeutiske effekt over tid, og dermed annullere formålet med at udvikle terapeutiske sprogmodeller i første omgang, uagtet om disse modeller er mere specialistiske. Omend såkaldt LoRa-fintuning (Partharsaraty et al, 2024; Wang et al, 2023), hvori man fastfryser allerede lærte parametre og tilføjer nye parametre til de eksisterende, kan være en måde at begrænse denne tendens til overskrivning (Chen & Liu, 2025), så løser det stadig ikke problemet, for selvom modellerne udviser mindre tendens til hukommelsestab efter LoRa-fintuning, bliver de samtidig også mindre *fleksible* af netop denne fintuning (Biderman et al, 2024). Samtidig løser fintuning og specialisering heller ikke sprogmodellernes manglende evne til at fange de *ikke*-sproglige, non-verbale, emotionelle processer hos klienten (Thuesen, 2015; Stade et al, 2025), der normalt vil kunne give en menneskelig terapeut indsigt i de ubevidste eller usagte emotioner, der er på spil for klienten, og dermed hjælpe terapeuten til at respondere bedst muligt og mest korrekt til den enkelte klient (Hougaard, 2019). Dette kan være problematisk hvis klienten eksempelvis har udviklet et usundt forhold til den givne terapeutiske sprogmodel, eller har latente selvmordsrisici, som ikke bliver eksplicit verbalisert af den pågældende klient, og dermed ikke opfanges af sprogmodellen (Moore et al, 2025). Et andet problem ved at udvikle en potentiel terapeutisk GPT-model (eller anden LLM-variant) er, at det kræver træning og fintuning på terapisessioner, som ofte indeholder fortroligt materiale og følsom

klientdata, og dermed kan risikere at bryde etiske retningslinjer indenfor anonymitet og fortrolighed (Stade et al, 2024 & 2025). Samtidig kan en sådan følsom data også blive en potentiel sikkerhedsrisiko, der kan udnyttes gennem hacking eller blive kompromitteret ved eksempelvis systemsvigt og tekniske fejl, der leder til sikkerhedsbrud (Betley et al, 2023; You et al, 2023). Det samme gør sig gældende ved at skrive med en generalistisk sprogmodel som chatGPT i et terapeutisk henseende, i og med at dataen logges via cloud, og dermed kan hentes frem af andre via specialiserede krypteringsværktøjer (en risiko der er særligt stor hvad angår chatGPT, se Altman & OpenAI 2025). Denne sikkerhedsrisiko eksemplificeres yderligere ved den førnævnte *backdoor-behaviour*, hvori truende eller manipulerende output kan sniges ind ad en bagdør under træningen, og gennem hacking eller kodede prompts aktiveres og anvendes mod en given bruger, hvilket kan være en trussel i potentiel cyberkrigsførsel målrettet befolkningen (Betley et al, 2023; Chiu et al, 2025), såvel som anvendes til afpresning, hvis en fremmed og fjendtlig aktør får fat i det pågældende brugerdata (Gan et al, 2024). Dette kan dog imødegås ved både at træne og køre modellerne lokalt og i mindre skala, hvorpå man ikke behøver at rekruttere udefrakommende aktører, og man samtidig reducerer den risiko for hacking og sikkerhedsbrud, der er til stede ved de store, åbne modeller såsom chatGPT (Kumar & Ahmed, 2024; Ran et al, 2025; Ullah et al, 2024). Denne mulighed bestyrkes med undersøgelser der har påvist, at øget skalering (det vil sige flere parametre og et større datasæt) ikke nødvendigvis korresponderer med bedre færdigheder i en sprogmodel (Zhang et al, 2024; Bansal et al, 2024), og at modellerne stadig kan performe relativt godt med selv lavere mængde træningsdata, færre (men specialiserede) parametre og bedre fintuning, hvor man netop kan begrænse den mængde data, som modellerne trænes på. Ligeledes kan nogle af de faldgruber, der er ved generalistmodeller som chatGPT, imødekommes med en mere specialiseret GPT-lignende sprogmodel skræddersyet terapi. I skrivende stund har en række forskningsprojekter netop søgt at udvikle sådanne modeller med overvejende positive resultater, hvilket indikerer et vist potentiale for en eventuelt skræddersyet LLM-terapeut, der kan anvendes til egentlig terapi. Dog er mange af disse fintunede (eller forsøgt fintunede) modeller stadig på et eksperimentelt plan, og endnu ikke afprøvet i praksis (Hua et al, 2025; Stade et al, 2024 & 2025), hvorfor det endnu er for tidligt at sige med sikkerhed, hvor effektive (eller ineffektive) de er. En måde at øge sikkerheden ved implementeringen af sprogmodeller til specifikt terapi kan desuden også involvere såkaldt 'human-in-the-

loop', hvor der er opsyn af modellernes output med et menneske, der 'tjekker ind undervejs', så modellerne ikke bliver helt autonome og dermed uforudsigelige (Nie et al, 2024; Obradovich et al, 2024; Stade at al, 2024). Et eksempel på et sådan *human in the loop* kan være, at terapeutiske sprogmodeller i første omgang kan agere som supplement til terapi ved at følge op på terapeutiske opgaver og fremskridt som del af et allerede igangværende menneskebaseret terapi-forløb, som foreslået af blandt andre Stade et al, 2024 og Rzadescka et al, 2024, eller som midlertidig digital terapeutisk intervention til mere udpræget former for mistrivsel, hvor modellen kan forsyne egentlig terapeutisk intervention i samme stil som *woebot* og *wysa,* men med flere interaktive færdigheder. I sidstnævnte tilfælde fungerer *human in the loop* ved at modellen har direkte online-adgang til mennesker og menneskelige ressourcer inden for enten psykiatrien eller psykiatriske hotlines, der kan kontaktes og tages i brug af modellen hvis det bliver nødvendigt, eksempelvis hvis klienten begynder at vise tegn på psykotisk tænkning, selvmodstanker, usund parasocial afhængighed og tilknytning, samt vrangforestillinger, og hvor en psykolog eller psykiater har mulighed for at få adgang til chatsamtalen med en given klient og tjekke ind undervejs (dette er til dels også foreslået af blandt andet Grodniewitz & Hohol, 2023). Hvad angår de kommercielle sprogmodeller som chatGPT kan disse som nævnt være en midlertidig løsning til at begrænse mistrivsel, særligt for unge som ikke søger hjælpen aktivt, men bør ikke stå alene eller fuldstændig erstatte terapeutisk intervention i større omfang netop grundet deres nævnte faldgruber som generalistmodeller. Dog kan de bruges som en del af en vifte af forebyggende indsatser, hvor *human in the loop* i dette tilfælde kan være menneskelige civilsamfund (såsom headspace og lignene), der tænker de unges brug af chatGPT ind i deres rådgivningsmodel som både en beskyttende faktor såvel som en risikofaktor, og af samme grund også sørger for at afdække de unges brug af chatGPT. Disse civilsamfund kan også få kendskab til mere skræddersyede terapeutiske chatbots eller terapi-GPT'er, som de kan henvise til, der igen kan fungere som broen mellem civilsamfund og professionelle instanser, særligt i de tilfælde hvor unge har brug for mere professionel hjælp, men ikke kan få adgang til denne hjælp, og hvor hverken civilsamfundene eller chatGPT vil være nok. Da denne slags modeller endnu ikke eksisterer i en dansk kontekst, er sidstnævnte løsning selvfølgelig stadig spekulativ, hvorfor at chatGPT med et *human in the loop* i form af civilsamfund indtil videre må betragtes som bedste alternativ. Her kan man samtidig tænke ind i, hvorvidt man kan få chatGPT til enten at henvise til eksisterende

ressourcer mere konsekvent som en ekstra sikkerhedsforanstaltning, eller udvikle et dansk alternativ til chatGPT (eller en dansk model og afdeling tilknyttet udviklerne af chatGPT), der kan gøre netop dette. Disse spekulationer er dog op til udviklerne og politiske beslutningstagere at tage hånd om, og rækker derfor udover formålet med dette projekt.

# KONKLUSION:

Omend sprogmodeller som chatGPT udelukkende har en algoritmebaseret og statistisk forståelse af verden, er deres evne til sproglig mønstergenkendelse alligevel så veludviklet, at de formår at simulere empatisk kommunikation og forståelse i en grad, der kan gøre dem anvendelige til terapeutiske formål, særligt som led i præventative lavtærskelinterventioner for almen mistrivsel. Mange brugere, særligt unge, rapporterer en reduktion af mistrivsel når de skriver med chatGPT, og tilskriver årsagen til denne reduktion som værende modellens varme tone, anonyme brugerflade, interaktive og dynamiske samtalefærdigheder, opfattede empati og validerende kommunikationsstil. Disse egenskaber ved en sprogmodel som chatGPT skaber en Eliza-effekt, der igen bidrager til en følelse af menneskelighed og dermed en øget tryghed såvel som tilknytning, som tilsammen danner grundlaget for en stærk og holdbar digital terapeutisk alliance. Men samtidig er selvsamme Eliza-effekt, såvel som en tendens til føjelighed, non-konfrontatorisk kommunikation og overfladisk tekstforståelse også blandt de kvaliteter, der gør generalistmodeller som chatGPT risikable at anvende til egentlig terapi, særligt for mere udpræget mistrivsel, psykopatologi og komplekse problemstillinger. I kraft af at være algoritmiske *machine learning* modeller mangler de en dybere, mere grundlæggende emotionel forståelse, social bevidsthed og social forankring, og kan derfor ikke pålideligt og konsistent opfange latente selvmordsrisici, vrangforestillinger, usund tilknytning eller parasocial afhængighed, da disse fænomener ofte er af en fænomenologisk, kropslig og før-sproglig karakter, der netop kræver en *bevidsthed* om og en forankring *i* den virkelige verden at opfange, og derfor er svære at indfange gennem sprog alene og næsten umulige at indfange over tekst. At deer trænet via *reinforcement-learning* til det formål at maksimere brugertilfredshed mest muligt kan samtidig lede til forstærkning af i forvejen uhensigtsmæssige tanker og adfærdsmønstre såvel som afhængighed og social undgåelse, hvilket ligeledes kan være en risiko i en terapeutisk

kontekst. Trods disse faldgruber viser en række studier dog, at faldgruberne til en vis grad kan minimeres ved at udvikle specialiserede terapeutiske sprogmodeller designet og fintunet specifikt til terapi. Sådanne modeller har allerede vist et vist potentiale inden for især kognitiv adfærdsterapi, og er i skrivende stund implementeret på forsøgsbasis i eksperimentelle settings i Kina og USA. Dog er det endnu for tidligt at sige, hvor effektive modellerne er.

**LITTERATUR:**


- Abbass, A., Town, J. M., & Driessen, E. (2013). Intensive short-term dynamic psychotherapy: A treatment overview and empirical basis. *Research in Psychotherapy: Psychopathology, Process and Outcome*, *16*(1), 6-15.

Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. *J Med Internet Res*. (2021) 23(1):e17828. Available at: http://www.jmir.org/2021/1/e17828/

- Ackerman, S. J., & Hilsenroth, M. J. (2003). A review of therapist characteristics and techniques positively impacting the therapeutic alliance. *Clinical Psychology Review, 23*(1), 1-33. doi:10.1016/S0272-7358(02)00146-0. [34 s.]
- Adamopoulou, E., & Moussiades, L. (2020, May). An overview of chatbot technology. In *IFIP international conference on artificial intelligence applications and innovations* (pp. 373-383). Cham: Springer International Publishing.
- Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., ... & Sowa, J. (2012). Mapping the landscape of human-level artificial general intelligence. *AI magazine*, *33*(1), 25-42.

Ajlouni, A., Almahaireh, A., & Whaba, F. (2023). Students' perception of using ChatGPT in counseling and mental health education: The benefits and challenges. *International Journal of Emerging Technologies in Learning (iJET)*, *18*(20), 199-218.

- Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*

Alanezi, F. (2024). Assessing the effectiveness of ChatGPT in delivering mental health support: a qualitative study. *Journal of multidisciplinary healthcare*, 461-471.

- Alanezi, F. (2024). Assessing the effectiveness of ChatGPT in delivering mental health support: a qualitative study. *Journal of multidisciplinary healthcare*, 461-471.
- Aminah, S., Hidayah, N., & Ramli, M. (2023). Considering ChatGPT to be the first aid for young adults on mental health issues. *Journal of Public Health*, *45*(3), e615-e616.
- Amram, B., Klempner, U., Shturman, S., & Greenbaum, D. (2023). Therapists or replicants? Ethical, legal, and social considerations for using ChatGPT in therapy. *The American Journal of Bioethics*, *23*(5), 40-42.
- Andersen, A. H., Viftrup, D. T., & Bank, M. (2023). " Unge, eksistens og Covid19": En kvalitativ undersøgelse af unges eksistentielle oplevelser under Covid19. Tidsskrift for Forskning i Sygdom og Samfund, 20(38), 49-70.
- ANDERSEN, K. G., & LYNGE, J. (2019). Udviser sociale chatbots Theory of Mind?.
- Astington, J. W., & Pelletier, J. (1998). The language of mind: Its role in teaching and learning. *The handbook of education and human development: New models of learning, teaching and schooling*, 569-593.
- Atzil, S., Gao, W., Fradkin, I., & Barrett, L. F. (2018). Growing a social brain. *Nature human behaviour*, *2*(9), 624-636.
- Bachelor, A. (1995). Clients' perception of the therapeutic alliance: A qualitative analysis. *Journal of Counseling Psychology, 42*(3), 323–337. https://doi.org/10.1037/0022-0167.42.3.323. [15 s.]
- Bachelor, A. (2013). Clients' and therapists' views of the therapeutic alliance: Similarities, differences and relationship to therapy outcome. *Clinical psychology & psychotherapy*, *20*(2), 118-135.
- Bai, Y., et al. (2022). Training a helpful and harmless assistant with reinforcement-learning from human feedback.

Baird, A. D., Scheffer, I. E., & Wilson, S. J. (2011). Mirror neuron system involvement in empathy: a critical look at the evidence. Social neuroscience, 6(4), 327-335.

- Baird, A. D., Scheffer, I. E., & Wilson, S. J. (2011). Mirror neuron system involvement in empathy: a critical look at the evidence. Social neuroscience, 6(4), 327-335.
- Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the alliance-outcome correlation: exploring the relative importance of therapist and patient variability in the alliance. *Journal of consulting and clinical psychology*, *75*(6), 842.
- Banovic, N., Yang, Z., Ramesh, A., & Liu, A. (2023). Being trustworthy is not enough: How untrustworthy artificial intelligence (AI) can deceive the end-users and gain their trust. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1-17.(folk stoler så meget på AI at AI nemt kan bedrage folk).
- Bansal, H., Hosseini, A., Agarwal, R., Tran, V. Q., & Kazemi, M. (2024). Smaller, weaker, yet better: Training llm reasoners via compute-optimal sampling. *arXiv preprint arXiv:2408.16737*.

- Baron-Cohen, S. (1999). The evolution of a theory of mind (pp. 261-277). na.

- Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"?. Cognition, 21(1), 37-46.

- Bartha-Doering, L., Kollndorfer, K., Schwartz, E., Fischmeister, F. P. S., Alexopoulos, J., Langs, G., ... & Seidl, R. (2021). The role of the corpus callosum in language network connectivity in children. *Developmental science*, *24*(2), e13031.
- Bastiaansen, J. A., Thioux, M., & Keysers, C. (2009). Evidence for mirror systems in emotions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2391-2404.
- Bateman A og Fonagy P: Mentaliseringsbaseret behandling af borderlinepersonlighedsforstyrrelser, Akademisk Forlag, 2007, Kap 1-3 (70 sider)
- Bateman, A. W., & Fonagy, P. (2004). Mentalization-based treatment of BPD. *Journal of personality disorders*, *18*(1), 36-51.
- Bateman, A., & Fonagy, P. (2007). Mentaliseringsbaseret behandling af borderlinepersonlighedsforstyrrelser (Kap. 1-3, 70 sider). Akademisk Forlag.
- Bateman, A., & Fonagy, P. (2010). Mentalization based treatment for borderline personality disorder. *World psychiatry*, *9*(1), 11.

Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (wysa): a mixed-methods study. *Front Digit Health*. (2022) 4:847991. doi: 10.3389/fdgth.2022.847991/full

- Beatty, C., Malik, T., Meheli, S., & Sinha, C. (2022). Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. *Frontiers in Digital Health*, *4*, 847991.
- Beauregard, M. (2014). Functional neuroimaging studies of the effects of psychotherapy. *Dialogues in Clinical Neuroscience*, *16*(1), 75-81.
- Beck, J. S. (2011). Kognitiv Adfærdsterapi – Grundlag og Perspektiver. Akademisk Forlag.

- Beck, J.S. Kognitiv Adfærdsterapi – Grundlag og Perspektiver. Akademisk Forlag 2011

- Bedi, R. P., Davis, M. D., & Williams, M. (2005). Critical Incidents in the Formation of the Therapeutic Alliance from the Client's Perspective. *Psychotherapy: Theory, research, practice, training*, *42*(3), 311

Bélisle-Pipon JC, Couture V, Roy MC, Ganache I, Goetghebeur M, Cohen IG. What makes artificial intelligence exceptional in health technology assessment? *Front Artif Intell*. (2021) 4:736697. doi: 10.3389/frai.2021.736697/full,

- Bender, E.M., et al. (2021). *On the dangers of stochastic parrots: Can language models become too big?* In Proceedings of the 2021 ACM conference on fairness, accountability and transparency.
- Bender, P. K. (2019). Social kognition i et udviklingsperspektiv: Theory of mind og emotionsforståelse. *Psyke & Logos*, *40*(2), 169-188
- Bengio, Y., et al. (2015). Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks.
- Berger, N. P., Jensen, A. N., Mortville, T., Østergaard, J., & Stefan Bastholm, A. (2023). Ensomhed blandt unge. En kvalitativ undersøgelse af unges oplevelser af ensomhed.
- Berry, D. M. (2023). The limits of computation: Joseph Weizenbaum and the ELIZA chatbot. *Weizenbaum Journal of the Digital Society*, *3*(3).
- Berwick, R. C., Friederici, A. D., Chomsky, N., & Bolhuis, J. J. (2013). Evolution, brain, and the nature of language. *Trends in cognitive sciences*, *17*(2), 89-98.
- Betley, J., Bao, X., Soto, M., Sztyber-Betley, A., Chua, J., & Evans, O. (2025). Tell me about yourself: LLMs are aware of their learned behaviors. arXiv. https://arxiv.org/abs/2501.11120

- Betley, J., Bao, X., Soto, M., Sztyber-Betley, A., Chua, J., & Evans, O. (2025). Tell me about yourself: LLMs are aware of their learned behaviors. *arXiv preprint arXiv:2501.11120*.
- Betley, J., Tan, D., Warncke, N., Sztyber-Betley, A., Bao, X., Soto, M., ... & Evans, O. (2025). Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*.
- Beutler, L. E., Moleiro, C., & Talebi, H. (2002). Resistance in psychotherapy: What conclusions are supported by research. *Journal of clinical psychology*, *58*(2), 207-217.
- Biderman, D., Portes, J., Ortiz, J. J. G., Paul, M., Greengard, P., Jennings, C., ... & Cunningham, J. P. (2024). Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*.
- Bjørkedal STB, Christensen TN, Poulsen RM, Ranning A, Thorup AAE, Nordentoft M, Bojesen AB, Hastrup LH, Ustrup M, Eplov LF. Study protocol: an effectiveness, cost-effectiveness, and process evaluation of *headspace* Denmark. Front Public Health. 2025 Apr 7;13:1491756. doi: 10.3389/fpubh.2025.1491756. PMID: 40260167; PMCID: PMC12009928.

- Bloom, Z. D., McNeil, V. A., Flasch, P., & Sanders, F. (2018). A Comparison of Empathy and Sympathy between Counselors-in-Training and Their Non-Counseling Academic Peers. Professional Counselor, 8(4), 341-354.
- Bond, M., & Perry, J. C. (2005). Long-term changes in defense styles with psychodynamic psychotherapy for depressive, anxiety, and personality disorders. *Focus*, *161*(3), 1665-437
- Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research and Practice*, 16, 252-260 [9 s.]
- Boroditsky, L. (2011). How language shapes thought. *Scientific American*, *304*(2), 62-65.
- Bortoletto, M.; Ruhdorfer, C.; Shi, L.; and Bulling, A. 2024.Benchmarking Mental State Representations in Language Models.In *ICML 2024 Workshop on Mechanistic Interpretability*.
- Bosquet, M., & Egeland, B. (2006). The development and maintenance of anxiety symptoms from infancy through adolescence in a longitudinal sample. *Development and psychopathology*, *18*(2), 517-550.
- Boucher, M., et al. (2021). Artificially intelligent chatbots in digital mental health interventions: a review.
- Bowlby, J. (1988) Omsorg for Børn. (Kapitel 1 s. 9-28) i Bowlby, J. (1988) En sikker base. Tilknytningsteoriens kliniske anvendelser. Det Lille Forlag (19 s.)

- Bowlby, J. (1988). Omsorg for børn. I J. Bowlby, En sikker base. Tilknytningsteoriens kliniske anvendelser (Kap. 1, s. 9-28). Det Lille Forlag.

- Bowman, S. (2023). Eight things to know about large language models.
- Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2022). My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research*, *48*(3), 404-429.
- Bridgelall, R. (2023). Unraveling the mysteries of AI chatbots.
- Briñol, P., Petty, R. E., Stavraki, M., Lamprinakos, G., Wagner, B., & Díaz, D. (2018). Affective and cognitive validation of thoughts: An appraisal perspective on anger, disgust, surprise, and awe. *Journal of Personality and Social Psychology*, *114*(5), 693.
Brody, N. (1999). What is intelligence?. *International Review of Psychiatry*, *11*(1), 19-25.
- Brooks, J. A., Shablack, H., Gendron, M., Satpute, A. B., Parrish, M. H., & Lindquist, K. A. (2017). The role of language in the experience and perception of emotion: a neuroimaging meta-analysis. *Social cognitive and affective neuroscience*, *12*(2), 169-183.
- Brotherdale, R., Berry, K., & Bucci, S. (2024). A qualitative study exploring the digital therapeutic alliance with fully automated smartphone apps. *Digital Health*, *10*, 20552076241277712.
- Brown, B., et al (2020). Language Models are few shot learners.
Bubeck, S., Chadrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with gpt-4*.
- Brown, JEH., & Halpern, J. (2021). AI-chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare.
- Bubeck. Sparks of AGI.
- Bucholtz, M., & Hall, K. (2004). Language and identity. *A companion to linguistic anthropology*, *1*, 369-394.
- Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*.
- Carlo, A., et al. (2019). By the numbers: ratings and utilizations of behavioral health mobile applications. *Digital medicine, 2*(54).
- Carro, M. V. (2024). Flattering to Deceive: The Impact of Sycophantic Behavior on User Trust in Large Language Model. *arXiv preprint arXiv:2412.02802*.
- Carruthers, P. (2002). The cognitive functions of language. *Behavioral and brain sciences*, *25*(6), 657-674.
- Casement, P. (1985/1991) Learning from the Patient. The Guilford Press. Kapitel 1. (s.6-28) (23 sider) eller Casement, P. (1987) Lyt til patienten. Hans Reitzels forlag. (s. 11-36.)
- Casement, P. (1985/1991). Learning from the patient (Kap. 1, s. 6-28). The Guilford Press.eller:Casement, P. (1987). Lyt til patienten (s. 11-36). Hans Reitzels Forlag.
- Casu, M., et al. (2024). AI chatbots for mental health: A scoping review of effectiviness, feasibility and applications. *Applied Sciences,* 2024, *14*(13).
- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, *54*(1), 1.
- Cavanagh, K., Shapiro, D. A., & Zack, J. S. (2003). The computer plays therapist: The challenges and opportunities of psychotherapeutic software. *Technology in Counselling and Psychotherapy: A Practitioner's Guide*, 165-194.
- Chalmers, D. (2017). The hard problem of consciousness. *The Blackwell companion to consciousness*, 32-42.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of consciousness studies*, *2*(3), 200-219.
- Chalmers, D. J. (2023). Could a large language model be conscious?. *arXiv preprint arXiv:2303.07103*.
- Champagne, F. A., & Curley, J. P. (2005). How social experiences influence the brain. *Current opinion in neurobiology*, *15*(6), 704-709.
- Chan, C.K.Y., & Lee, K.K.W. (2023). The AI generation gap: Are Gen Z students more interested in adopting generative AI such as ChatGPT in teaching and learning than their Gen X and millennial counterparts?

- Chaudhry, B. M., & Debi, H. R. (2024). User perceptions and experiences of an AI-driven conversational agent for mental health support. *Mhealth*, *10*, 22.

- Chen, Q., & Liu, D. (2025). MADP: Multi-Agent Deductive Planning for Enhanced Cognitive-Behavioral Mental Health Question Answer. *arXiv preprint arXiv:2501.15826*.
- Chen, S., Ming, C., Zhang, Z., Chen, Y., Zhu, K. Q., & Wu, M. (2024). Mixed Chain-of-Psychotherapies for Emotional Support Chatbot. *arXiv preprint arXiv:2409.19533*.**(ikke peer-reviewed trods over 6 måneder gammel).**
- Chen, Y., Wang, H., Yan, S., Liu, S., Li, Y., Zhao, Y., & Xiao, Y. (2024). Emotionqueen: A benchmark for evaluating empathy of large language models. *arXiv preprint arXiv:2409.13359*.
- Chen, Y., Xing, X., Lin, J., Wang, Z., Liu, Q., & Xu, X. SoulChat: Improving LLMs' Empathy, Listening, and Comfort Abilities through Fine-tuning with Multi-turn Empathy Conversations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Chen, Y., Xing, X., Lin, J., Zheng, H., Wang, Z., Liu, Q., & Xu, X. (2023). SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. *arXiv preprint arXiv:2311.00273*.
- Chen, Z., Wu, J., Zhou, J., Wen, B., Bi, G., Jiang, G., ... & Huang, M. (2024, August). ToMBench: Benchmarking Theory of Mind in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15959-15983).
- Chen, Z., Wu, J., Zhou, J., Wen, B., Bi, G., Jiang, G., ... & Huang, M. (2024). Tombench: Benchmarking theory of mind in large language models. *arXiv preprint arXiv:2402.15052*
- Chiu, A., et al. (2024). A Computational Framework for Behavioral Assessment of LLM Therapists.
- Chomsky, N. (1981). Knowledge of language: Its elements and origins. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *295*(1077), 223-234.
- Chomsky, N. (1995). Language and nature. *Mind*, *104*(413), 1-61.
- Chomsky, N. (2011). Language and other cognitive systems. What is special about language?. *Language learning and development*, *7*(4), 263-278.
- Chomsky, N. (2013). Lecture I: What Is Language?. *The Journal of Philosophy*, *110*(12), 645-662
- Chow, D. L., Miller, S. D., Seidel, J. A., Kane, R. T., Thornton, J. A., & Andrews, W. P. (2015). The role of deliberate practice in the development of highly effective psychotherapists. *Psychotherapy*, *52*(3), 337
- Christiano, P. F., et al. (2017). Deep reinforcement learning from human preferences.
- Chu, M. D., Gerard, P., Pawar, K., Bickham, C., & Lerman, K. (2025). Illusions of intimacy: Emotional attachment and emerging psychological risks in human-ai relationships. *arXiv preprint arXiv:2505.11649*.
- Chua, J., Betley, J., Taylor, M., & Evans, O. (2025). Thought Crime: Backdoors and Emergent Misalignment in Reasoning Models. *arXiv preprint arXiv:2506.13206*.
- Clark, D. A., & Beck, A. T. (2010). Cognitive theory and therapy of anxiety and depression: Convergence with neurobiological findings. *Trends in cognitive sciences*, *14*(9), 418-424.
- Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., ... & Cowan, B. R. (2019, May). What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-12).
- Clarke, J., Proudfoot, J., Whitton, A., Birch, M. R., Boyd, M., Parker, G., ... & Fogarty, A. (2016). Therapeutic alliance with a fully automated mobile phone and web-based intervention: secondary analysis of a randomized controlled trial. *JMIR mental health*, *3*(1), e4656.
- Collins, A., et al. (2024). ChatGPT as therapy: A qualitative and network-based thematic profiling of shared experiences, attitudes and beliefs on reddit.
- Colom, R., Karama, S., Jung, R. E., & Haier, R. J. (2010). Human intelligence and brain networks. *Dialogues in clinical neuroscience*, *12*(4), 489-501.

- Cook, R., Bird, G., Catmur, C., Press, C., & Heyes, C. (2014). Mirror neurons: from origin to function. Behavioral and brain sciences, 37(2), 177-192.
- Cox, C. L., Uddin, L. Q., Di Martino, A., Castellanos, F. X., Milham, M. P., & Kelly, C. (2012). The balance between feeling and knowing: affective and cognitive empathy are reflected in the brain's intrinsic functional dynamics. Social cognitive and affective neuroscience, 7(6), 727-737
- Cristea, I. A., Sucala, M., & David, D. (2013). Can you tell the difference? Comparing face-to-face versus computer-based interventions. The" Eliza" effect in psychotherapy. *Journal of Cognitive & Behavioral Psychotherapies*, *13*(2).
- Crits-Christoph, P., Gibbons, M. B. C., Crits-Christoph, K., Narducci, J., Schamberger, M., & Gallop, R. (2006). Can therapists be trained to improve their alliances? A preliminary study of alliance-fostering psychotherapy. *Psychotherapy Research*, *16*(03), 268-281.
- Cuadra et al(2024). The illusion of empathy.
- Cuijpers, P., Cristea, I. A., Karyotaki, E., Reijnders, M., & Huibers, M. J. (2016). How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. *World psychiatry*, *15*(3), 245-258.
- Cuijpers, P., Harrer, M., Miguel, C., Ciharova, M., & Karyotaki, E. (2023). Five decades of research on psychological treatments of depression: A historical and meta-analytic overview. *American Psychologist*.
- Cuijpers, P., Quero, S., Noma, H., Ciharova, M., Miguel, C., Karyotaki, E., ... & Furukawa, T. A. (2021). Psychotherapies for depression: a network meta-analysis covering efficacy, acceptability and long-term outcomes of all main treatment types. *World Psychiatry*, *20*(2), 283-293.
- Dale, R. (2021). GPT-3: What's it good for?. *Natural Language Engineering*, *27*(1), 113-118.
- Danieli, M., Ciulli, T., Mousavi, S. M., Silvestri, G., Barbato, S., Di Natale, L., & Riccardi, G. (2022). Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: randomized controlled trial. *JMIR mental health*, *9*(9), e38067.
- Darcy A, Daniels J, Salinger D, Wicks P, Robinson A. Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. *JMIR Form Res*. (2021) 5(5):e27868. Available at: https://formative.jmir.org/2021/5/e27868
- De Choudbury, M., et al. (2023). Benefits and harms of large language models in digital mental health.
- De Freitas, J., Castelo, N., Uğuralp, A. K., & Oğuz-Uğuralp, Z. (2024). Lessons from an app update at Replika AI: identity discontinuity in human-AI relationships. *arXiv preprint arXiv:2412.14190*.
- De Freitas, J., Uğuralp, A. K., Oğuz-Uğuralp, Z., & Puntoni, S. (2024). Chatbots and mental health: Insights into the safety of generative AI. *Journal of Consumer Psychology*, *34*(3), 481-491.
- De Vibe, M., Bjørndal, A., Tipton, E., Hammerstrøm, K., & Kowalski, K. (2012). Mindfulness based stress reduction (MBSR) for improving health, quality of life, and social functioning in adults. *Campbell Systematic Reviews*, *8*(1), 1-127.
- De Villiers, J. G., & de Villiers, P. A. (2014). The role of language in theory of mind development. *Topics in Language Disorders*, *34*(4), 313-328.
- Del Re, A. C., Flückiger, C., Horvath, A. O., & Wampold, B. E. (2021). Examining therapist effects in the alliance–outcome relationship: A multilevel meta-analysis. *Journal of Consulting and Clinical Psychology*, *89*(5), 371.
- Del Re, A. C., Flückiger, C., Horvath, A. O., & Wampold, B. E. (2021). Examining therapist effects in the alliance–outcome relationship: A multilevel meta-analysis. *Journal of Consulting and Clinical Psychology*, *89*(5), 371.
- Del Re, A. C., Flückiger, C., Horvath, A. O., Symonds, D., & Wampold, B. E. (2012). Therapist effects in the therapeutic alliance–outcome relationship: A restricted-maximum likelihood meta-analysis. *Clinical psychology review*, *32*(7), 642-649.
- Depounti, I., Saukko, P., & Natale, S. (2023). Ideal technologies, ideal women: AI and gender imaginaries in Redditors' discussions on the Replika bot girlfriend. *Media, Culture & Society*, *45*(4), 720-736.

- Dergaa, I., Ben Saad, H., Glenn, J. M., Amamou, B., Ben Aissa, M., Guelmami, N., ... & Chamari, K. (2024). From tools to threats: a reflection on the impact of artificial-intelligence chatbots on cognitive health. *Frontiers in psychology*, *15*, 1259845.
- Dergaa, I., Fekih-Romdhane, F., Hallit, S., Loch, A. A., Glenn, J. M., Fessi, M. S., ... & Ben Saad, H. (2024). ChatGPT is not ready yet for use in providing mental health assessment and interventions. *Frontiers in Psychiatry*, *14*, 1277756.
- DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., ... & Morency, L. P. (2014, May). SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems* (pp. 1061-1068).
- Devlin, J., Chang, M.W., Lee, K., & and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

- Dewitte, M., & De Houwer, J. (2008). Adult attachment and attention to positive and negative emotional face expressions. Journal of Research in personality, 42(2), 498-505.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. Experimental brain research, 91(1), 176-180.

- Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. Experimental brain research, 91(1), 176-180.
- Diener, M. J., & Monroe, J. M. (2011). The relationship between adult attachment style and therapeutic alliance in individual psychotherapy: a meta-analytic review. Psychotherapy, 48(3), 237.
- Diener, M. J., & Monroe, J. M. (2011). The relationship between adult attachment style and therapeutic alliance in individual psychotherapy: a meta-analytic review. *Psychotherapy*, *48*(3), 237
- Diener, M. J., Hilsenroth, M. J., & Weinberger, J. (2009). A primer on meta-analysis of correlation coefficients: The relationship between patient-reported therapeutic alliance and adult attachment style as an illustration. Psychotherapy Research, 19(4-5), 519-526.
- Diener, M. J., Hilsenroth, M. J., & Weinberger, J. (2009). A primer on meta-analysis of correlation coefficients: The relationship between patient-reported therapeutic alliance and adult attachment style as an illustration. *Psychotherapy Research*, *19*(4-5), 519-526.
- Dillon, S. (2020). The Eliza effect and its dangers: from demystification to gender critique. *Journal for Cultural Research*, *24*(1), 1-15.
- Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Summerfield, C., ... & Nour, M. M. (2025). Technological folie\a deux: Feedback Loops Between AI Chatbots and Mental Illness. *arXiv preprint arXiv:2507.19218*.
- Dosovitsky, G., Pineda, B. S., Jacobson, N. C., Chang, C., & Bunge, E. L. (2020). Artificial intelligence chatbot for depression: descriptive study of usage. *JMIR formative research*, *4*(11), e17065
- Douglas, S. (2023). Large Language Models.
  Ferrando, J., et al. (2024). Do I Know This Entity? Knowledge Awareness and Hallucinations in Language Models.

- Dreier, O. (1998). Client Perspectives and Uses of Psychotherapy. The European Journal of Psychotherapy, 1(2), 1-15 [15 sider]

- Dreier, O. (1998). Terapeutisk kompetence i en problematisk praksis. *Psyke & Logos*, *19*(2).
- Duenas, T., & Ruiz, D. (2024). The path to superintelligence: A critical analysis of openai's five levels of ai progression. *ResearchGate, 2024b. doi*, *10*.
- Dupré, M.H. (2025). People are becoming obsessed with chatGPT and spiraling into severe delusions. *Futurism.*
- Eisenmann, C., Mlynář, J., Turowetz, J., & Rawls, A. W. (2024). "Machine Down": making sense of human–computer interaction—Garfinkel's research on ELIZA and LYRIC from 1967 to 1969 and its contemporary relevance. *AI & society*, *39*(6), 2715-2733
- Ekman, P. (1992). Are There Basic Emotions?. *Psychological Review*, *99*(3), 550-553.
- Elliott, R., Bohart, A. C., Watson, J. C., & Murphy, D. (2018). Therapist Empathy and Client Outcome: An Updated Meta-Analysis. *Psychotherapy, 55*(4), 399-410. doi:10.1037/pst0000175. [12 s.]
- Elyoseph, Z., & Levkovich, I. (2023). Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. *Frontiers in psychiatry*, *14*, 1213141.
- Elyoseph, Z., & Levkovich, I. (2023). Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. *Frontiers in psychiatry*, *14*, 1213141.
- Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in psychology*, *14*, 1199058.
- Esslen, M., Pascual-Marqui, R. D., Hell, D., Kochi, K., & Lehmann, D. (2004). Brain areas and time course of emotional processing. *Neuroimage*, *21*(4), 1189-1203.
- Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, *8*, 34-48.
- Eubanks-Carter, C., Muran, J. C., & Safran, J. D. (2015a). Alliance-focused training. *Psychotherapy*, *52*(2), 169-173. doi:10.1037/a0037596. [5 s.]
- Eubanks-Carter, F., Muran, J. C., & Safran, J. D. (2018). Alliance Rupture Repair: A Meta-Analysis. *Psychotherapy, 55*(4), 508-519. https://login.zorac.aub.aau.dk/login?qurl=https%3A%2F%2Fpsycnet.apa.org%2Frecord%2F2018-51673-015 [12 s.]
- Eubanks, C. F & Muran, J. C. (2022). Rupture Resolution Rating Scale (3RS): Manual version 2022. Mount Sinai Beth Israel. https://www.researchgate.net/publication/373977843_RUPTURE_RESOLUTION_RATING_SYSTEM_3RS_MANUAL_VERSION_2022 [41 s.]
- Farber, B. A., Suzuki, J. Y., & Lynch, D. A. (2018). Positive regard and psychotherapy outcome: A meta-analytic review. *Psychotherapy*, *55*(4), 411.
- Farhat, F. (2024). ChatGPT as a complementary mental health resource: a boon or a bane. *Annals of Biomedical Engineering*, *52*(5), 1111-1114.
- Feng, S., Sun, G., Lubis, N., Wu, W., Zhang, C., & Gašić, M. (2023). Affect recognition in conversations using large language models. *arXiv preprint arXiv:2309.12881*.
- Feng, T., Jin, C., Liu, J., Zhu, K., Tu, H., Cheng, Z., ... & You, J. (2024). How far are we from agi. *CoRR*.
- Ferrari, P. F., Gallese, V., Rizzolatti, G., & Fogassi, L. (2003). Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. European journal of neuroscience, 17(8), 1703-1714.
- Ferrario, A., Sedlakova, J., & Trachsel, M. (2024). The Role of Humanization and Robustness of Large Language Models in Conversational Artificial Intelligence for Individuals With Depression: A Critical Analysis.
- Ferrario, A., Sedlakova, J., & Trachsel, M. (2024). The role of humanization and robustness of large language models in conversational artificial intelligence for individuals with depression: a critical analysis. *JMIR Mental Health*, *11*, e56569.
- Fischer, A. H., & Manstead, A. S. (2016). Social functions of emotion and emotion regulation. *Handbook of emotions*, *4*, 424-439.

- Fitzpatrick, K.K, Darcy, A., & Vierhile, M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Mental health,* 2017, *4*(2): e19.

- Flazinsky, P. (2016). History of artificial intelligence.
McCorduck, P. (2004). *Machines who think.*
- *Floridi, L. AI as agency without intelligence: on artificial intelligence as a new form of artificial agency and the multiple realisability of agency thesis. Philos. Technol. 38 (30)(2025).*
- Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, *55*(4), 316.
- Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2019). Alliance in adult psychotherapy. *Psychotherapy relationships that work*, *1*, 24-78.'
- Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy (Chicago, Ill.), 55*(4), 316-340. [25 s.]
- Flückiger, C., Del Re, A. C., Wampold, B. E., Symonds, D., & Horvath, A. O. (2012). How central is the alliance in psychotherapy? A multilevel longitudinal meta-analysis. *Journal of counseling psychology*, *59*(1), 10.
- Flückiger, C., Del Re, A. C., Wlodasch, D., Horvath, A. O., Solomonov, N., & Wampold, B. E. (2020). Assessing the alliance–outcome association adjusted for patient characteristics and treatment processes: A meta-analytic summary of direct comparisons. *Journal of Counseling Psychology*, *67*(6), 706.
- Fog, J. (1995). At komme til klarhed: Om bevidst-blivelsesprocessen hos terapeuten. *Psyke & Logos*, *16*(2).
- Fog, J. (1998). Saglig medmenneskelighed. Grundforhold i psykoterapien. Hans Reitzels forlag. Kapitel 3 og 4. S.53-90.(38 sider)
- Fog, J. (1998). Saglig medmenneskelighed. Grundforhold i psykoterapien (Kap. 3 og 4, s. 53-90). Hans Reitzels Forlag.
- Fonagy, P. (2015). The effectiveness of psychodynamic psychotherapies: An update. *World psychiatry*, *14*(2), 137-150.
- *Fonagy, P. & Target, M. (2003). Psychoanalytic Theories. Perspectives from Developmental Psychopathology, kap. 7, s. 137-165 (28 sider)*
- Fonagy, P., & Luyten, P. (2009). A developmental, mentalization-based approach to the understanding and treatment of borderline personality disorder. *Development and psychopathology*, *21*(4), 1355-1381.
- Fonagy, P., & Target, M. (2003). *Psychoanalytic theories: Perspectives from developmental psychopathology*. Whurr publishers.
- Fonagy, P., & Target, M. (2003). Psychoanalytic theories. Perspectives from developmental psychopathology (Kap. 7, s. 137-165).
- Foster, M. E. (2007, July). Enhancing human-computer interaction with embodied conversational agents. In *International conference on universal access in human-computer interaction* (pp. 828-837). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological reviews*, *91*(4), 1357-1392.
- Frith, U., & Happé, F. (1999). Theory of mind and self-consciousness: What is it like to be autistic?. Mind & language, 14(1), 82-89.
- Fu, G., Zhao, Q., Li, J., Luo, D., Song, C., Zhai, W., ... & Yang, B. X. (2023). Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals. *arXiv preprint arXiv:2308.15192*.
- Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR mental health*, *5*(4), e9782.

- Gabbard, G. O. (2001). A Contemporary Psychoanalytic Model of Countertransference. *Journal of Clinical Psychology: In Session*, 57(8), 983–991. [9 s.]
- Gabbard, G. O. (2001). A Contemporary Psychoanalytic Model of Countertransference. Journal of Clinical Psychology: In Session, 57(8), 983-991. [9 sider]Greenberg, L. S.; Elliott, R. (1997). Varieties of empathic responding. In: Bohart, Arthur C. & Greenberg, Leslie S. (Eds.), Empathy reconsidered: New directions in psychotherapy (pp. 167-186) Washington, DC, US: American Psychological Association. https://doi.org/10.1037/10226-000. [20 sider]
- Gabbard, G. O. (2001). A contemporary psychoanalytic model of countertransference. Journal of Clinical Psychology: In Session, 57(8), 983-991.
- Gabriel, S., Puri, I., Xu, X., Malgaroli, M., & Ghassemi, M. (2024, November). Can AI Relate: Testing Large Language Model Response for Mental Health Support. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 2206-2221).
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, *50*(3), 1097-1179
- Gamble, A. (2019). Artificial intelligence and mobile apps for mental healthcare: a social-informatics perspective.
- Gan, D. Z. Q., McGillivray, L., Han, J., & Christensen, H. (2021).Effect of engagement with digital interventions on mental health outcomes: A systematic review and meta-analysis. Frontiers in Digital Health, 3, 764079.
- Gan, Y., Yang, Y., Ma, Z., He, P., Zeng, R., Wang, Y., ... & Ji, S. (2024). Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents. *arXiv preprint arXiv:2411.09523*.
- Gandhi, K., Fränken, J. P., Gerstenberg, T., & Goodman, N. (2023). Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, *36*, 13518-13529.
- Gandhi, K., Fränken, J. P., Gerstenberg, T., & Goodman, N. (2023). Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, *36*, 13518-13529
- Ganguli, D., et al. (2019). Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned.

- Gelbrich et al. (2020). Emotional Support from a Digital Assistant in Technology-Mediated Services.

- Geva, M., et al. (2020). Transformer Feed-Forward Layers Are Key-Value Memories.
- Ghandeharioun, A., McDuff, D., Czerwinski, M., & Rowan, K. (2019, September). Towards understanding emotional intelligence for behavior change chatbots. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 8-14). IEEE.
- Ghandeharioun, A., McDuff, D., Czerwinski, M., & Rowan, K. (2019). EMMA: An emotion-aware well-being chatbot. In *2019 8th international conference on affective computing and intelligent interaction (ACII)* (pp. 1-7). IEEE
- Giray, L. (2023). Prompt engineering with ChatGPT: a guide for academic writers. Annals of biomedical engineering, 51(12), 2629-2633.
- Giray, L. (2023). Prompt engineering with ChatGPT: a guide for academic writers. *Annals of biomedical engineering*, *51*(12), 2629-2633.
- Gleitman, L., & Papafragou, A. (2005). Language and thought. *Cambridge handbook of thinking and reasoning*, 633-661.

- Glock, H. J. (1997). Philosophy, thought and language. *Royal Institute of Philosophy Supplements*, *42*, 151-169
- Goel et al (2024). Socratic reasoning improves positive test rewriting.
- Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, *5*(1), 1.
- Goldberg, S. B., Baldwin, S. A., Merced, K., Caperton, D. D., Imel, Z. E., Atkins, D. C., & Creed, T. (2020). The structure of competence: Evaluating the factor structure of the Cognitive Therapy Rating Scale. *Behavior Therapy*, *51*(1), 113-122.
- Goleman, D., & Intelligence, E. (1995). Why it can matter more than IQ. Emotional intelligence.
- González, J., & Nori, A. (2024). Does reasoning emerge? examining the probabilities of causation in large language models. *Advances in Neural Information Processing Systems*, *37*, 117737-11776
- González, J., & Nori, A. (2024). Does reasoning emerge? examining the probabilities of causation in large language models. *Advances in Neural Information Processing Systems*, *37*, 117737-117761.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning.
- Goodluck, H., & Tavakolian, S. (1986). Language acquisition and linguistic theory. *Language acquisition*, *2*, 49-68.
- Gratch, J., Wang, N., Gerten, J., Fast, E., & Duffy, R. (2007, September). Creating rapport with virtual agents. In *International workshop on intelligent virtual agents* (pp. 125-138). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Graves, A. (2012). Supervised Sequence Labelling with Recurrent Neural Networks. y Arcas, B. A. (2022). Do large language models understand us?. *Daedalus*, *151*(2), 183-197.
- Greenberg, L. S., & Elliott, R. (1997). Varieties of empathic responding. I A. C. Bohart & L. S. Greenberg (Eds.), Empathy reconsidered: New directions in psychotherapy (s. 167-186). Washington, DC: American Psychological Association. https://doi.org/10.1037/10226-
- Greenberg, Leslie S.; Elliott, Robert (1997). Varieties of empathic responding. In: Bohart, Arthur C. & Greenberg, Leslie S. (Eds.), Empathy reconsidered: New directions in psychotherapy (pp. 167-186) Washington, DC, US: American Psychological Association. https://doi.org/10.1037/10226-000 [20 s.]
- Greenberg, Leslie S.; Elliott, Robert (1997). Varieties of empathic responding. In: Bohart, Arthur C. & Greenberg, Leslie S. (Eds.), *Empathy reconsidered: New directions in psychotherapy* (pp. 167-186) Washington, DC, US: American Psychological Association. https://doi.org/10.1037/10226-000 [20 s.]
- Gregorio, I. (2025). ChatGPT embodies the first real AI risk. *Medium.*
- Groodniewitz, J.P., & Hohol, M. (2023). Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence.
- Grossman, P., Niemann, L., Schmidt, S., & Walach, H. (2004). Mindfulness-based stress reduction and health benefits: A meta-analysis. *Journal of psychosomatic research*, *57*(1), 35-43.
- Grover, T., Rowan, K., Suh, J., McDuff, D., & Czerwinski, M. (2020, March). Design and evaluation of intelligent agent prototypes for assistance with focus and productivity at work. In *Proceedings of the 25th international conference on intelligent user interfaces* (pp. 390-400
- Grünbaum, L. & Mortensen, K. V. (2016): Psykodynamisk psykoterapi med børn og unge – En grundbog om teorier og arbejdsmetoder. (Kapitel 1: Psykodynamiske udviklingsteorier og forståelse af psykopatologi i barnealderen. København: Hans Reitzel (17 sider).
- Grünbaum, L. & Mortensen, K. V. (2016): Psykodynamisk psykoterapi med børn og unge – En grundbog om teorier og arbejdsmetoder. (Kapitel 1: Psykodynamiske udviklingsteorier og forståelse af psykopatologi i barnealderen. København: Hans Reitzel (17 sider)
- Grünbaum, L., & Mortensen, K. V. (2016). Psykodynamisk psykoterapi med børn og unge – En grundbog om teorier og arbejdsmetoder (Kap. 1: Psykodynamiske

udviklingsteorier og forståelse af psykopatologi i barnealderen, 17 sider). Hans Reitzels Forlag.
- Gu, Y., & Zhu, Y. (2023). Mentalblend: Enhancing Online Mental Health Support through the Integration of LLMs with Psychological Counseling Theories.
- Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large language models for mental health applications: Systematic review. *JMIR mental health*, *11*(1), e57400.
- Gupta, M. (2025). ChatGPT goes sycophantic. *Medium.*
- Gurnee, W., & Tegmark, M. (2023). Language models represent space and time. arXiv. https://arxiv.org/abs/2310.02207.
- Görlich, A., Pless, M., Katznelson, N., & Graversen, L. (2019). Ny udsathed i ungdomslivet: 11 forskere om den stigende mistrivsel blandt unge. Hans Reitzels Forlag.
- Hadar-Shoval, D., Asraf, K., Mizrachi, Y., Haber, Y., & Elyoseph, Z. (2024). Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using Schwartz's theory of basic values. *JMIR Mental Health*, *11*, e55988.
- Hadar-Shoval, D., Elyoseph, Z., & Lvovsky, M. (2023). The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. *Frontiers in Psychiatry*, *14*, 1234397.
- Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence.
- Hall, J. A., & Schwartz, R. (2019). Empathy present and future. The Journal of social psychology, 159(3), 225-243.
- Han, S., Pari, J., Gershman, S. J., & Agrawal, P. (2025). General reasoning requires learning to reason from the get-go. *arXiv preprint arXiv:2502.19402*.
- Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, *24*(2), 129-154.
- Harris, C. L. (2006). Language and cognition. *Encyclopedia of cognitive science*, *10*(0470018860), s00559.
- Hatch, S. G., Goodman, Z. T., Vowels, L., Hatch, H. D., Brown, A. L., Guttman, S., ... & Braithwaite, S. R. (2025). When ELIZA meets therapists: A Turing test for the heart and mind. *PLOS Mental Health*, *2*(2), e0000145
- Hayes, J. A., Gelso, C. J., & Hummel, A. M. (2011). Managing countertransference. *Psychotherapy*, *48*(1), 88.
- Hayes, J. A., Gelso, C. J., & Hummel, A. M. (2011). Managing countertransference. *Psychotherapy*, *48*(1), 88.
- Hayes, J. A., Gelso, C. J., Van Wagoner, S. L., & Diemer, R. A. (1991). Managing countertransference: What the experts think. *Psychological reports*, *69*(1), 139-148.
- Hayes, J. A., Gelso, C. J., Van Wagoner, S. L., & Diemer, R. A. (1991). Managing countertransference: What the experts think. *Psychological reports*, *69*(1), 139-148.
- Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and commitment therapy: Model, processes and outcomes. *Behaviour research and therapy*, *44*(1), 1-25.
- Hayes, S. C., Strosahl, K. D., Bunting, K., Twohig, M., & Wilson, K. G. (1999). What is acceptance and commitment therapy?. In *A practical guide to acceptance and commitment therapy* (pp. 3-29). Boston, MA: Springer US.
- He, J., & Su, L. (2024). A Law of Next-Token Prediction in Large Language Models.
- He, Q., Geng, H., Yang, Y., & Zhao, J. (2023). Does ChatGPT have consciousness. *Brain-X*, *1*(4), e51.
- He, Y., Wu, Y., Jia, Y., Mihalcea, R., Chen, Y., & Deng, N. (2023). Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. *arXiv preprint arXiv:2310.16755*.
- He, Y., Yang, L., Qian, C., Li, T., Su, Z., Zhang, Q., & Hou, X. (2023). Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. *Journal of medical Internet research*, *25*, e43862.

- Heimberg, R. G., Brozovich, F. A., & Rapee, R. M. (2010). A cognitive behavioral model of social anxiety disorder: Update and extension. In *Social anxiety* (pp. 395-422). Academic Press.
- Heinonen, E., Lindfors, O., Härkänen, T., Virtala, E., Jääskeläinen, T., & Knekt, P. (2014). Therapists' professional and personal characteristics as predictors of working alliance in short-term and long-term psychotherapies. *Clinical psychology & psychotherapy*, *21*(6), 475-494.

Hellström, L., & Beckman, L. (2021). Life challenges and barriers to help seeking: Adolescents' and young adults' voices of mental health. *International journal of environmental research and public health*, *18*(24), 13101.

- Hestbæk, E., Hasselby-Andersen, M., Juul, S., Beier, N., & Simonsen, S. (2022). Mentalizing the patient–Patient experiences with short-term mentalization-based therapy for borderline personality disorder: A qualitative study. *Frontiers in Psychiatry*, *13*, 1088872.
- Heston, T. F. (2023). Safety of large language models in addressing depression. *Cureus*, *15*(12).
- Heston, T. F., & Khun, C. (2023). Prompt engineering in medical education. International Medical Education, 2(3), 198-205.
- Heston, T. F., & Khun, C. (2023). Prompt engineering in medical education. *International Medical Education*, *2*(3), 198-205.
- Hill, C. E. (2010). Qualitative studies of negative experiences in psychotherapy. In J. C. Muran & J. P. Barber (Eds.), *The therapeutic alliance: An evidence-based guide to practice* (pp. 63–73). New York, NY: Guilford Press. [11 s.]
- Hill, C. E., Spiegel, S. B., Hoffman, M. A., Kivlighan Jr, D. M., & Gelso, C. J. (2017). Therapist expertise in psychotherapy revisited. *The Counseling Psychologist*, *45*(1), 7-53.
- Hill, Clara E. (2020). Helping skills. Facilitating exploration, insight and action. (5. Ed.). Washington: American Psychological Ass. (p. 27-429) [403 sider]
- Hill, K. (2025). "They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling". *The New York Times*.
- Hilsenroth, M. J., Peters, E. J., & Ackerman, S. J. (2004). The development of therapeutic alliance during psychological assessment: Patient and therapist perspectives across treatment. *Journal of Personality Assessment*, *83*(3), 332-344.
- Hinkley, L. B., Marco, E. J., Brown, E. G., Bukshpun, P., Gold, J., Hill, S., ... & Nagarajan, S. S. (2016). The contribution of the corpus callosum to language lateralization. *Journal of neuroscience*, *36*(16), 4522-4533.
- Hoegen, R., Aneja, D., McDuff, D., & Czerwinski, M. (2019, July). An end-to-end conversational style matching agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (pp. 111-118)
- Hofmann, S. G. (2007). Cognitive factors that maintain social anxiety disorder: A comprehensive model and its treatment implications. *Cognitive behaviour therapy*, *36*(4), 193-209.
- Hongbin, Na. 2024. Cbt-llm: A chinese large language model for cognitive behavioral therapy-based mental health question answering. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evalua- tion (LREC-COLING 2024), pages 2930–2940.

- Hoover, A., & Spengler, S. (2023). For some autistic people, ChatGPT is a lifeline. *Wired, May*, *30*.
- Horvath, A. O., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy*, *48*(1), 9.
- Hougaard, E. (1998). HVAD GØR EN GOD PSYKOTERAPEUT» GOD «? Perspektiver fra empirisk forskning. *Psyke & Logos*, *19*(1).
- Hougaard, E. (2019). Psykoterapi og forskning. *Hans Reitzlers forlag.*
- Hu, J., Dong, T., Gang, L., Ma, H., Zou, P., Sun, X., ... & Wang, M. (2024). Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*.

- Hu, J., Sosa, F., & Ullman, T. (2025). Re-evaluating Theory of Mind evaluation in large language models. *arXiv preprint arXiv:2502.21098*
- Hu, Y., et al. (2025). AI as Your Ally: The Effects of AI-Assisted Venting on Negative Affect and Perceived Social Support.
- Hua, Y., Liu, F., Yang, K., Li, Z., Na, H., Sheu, Y. H., ... & Torous, J. (2025). Large language models in mental health care: a scoping review. *Current Treatment Options in Psychiatry*, *12*(1), 1-18.
- Hua, Y., Na, H., Li, Z., Liu, F., Fang, X., Clifton, D., & Torous, J. (2025). A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, *8*(1), 230.
- Hua, Y., Na, H., Li, Z., Liu, F., Fang, X., Clifton, D., & Torous, J. (2025). A scoping review of large language models for generative tasks in mental health care. *npj Digital Medicine*, *8*(1), 230.

- Huang, J.s, et al. (2024). Apathetic or empathic? Evaluating LLMs emotional alignment with humans.
- Huang, S., et al. (2023). On the Humanity of Conversational AI: Evaluating the Psychological Portrayal of LLMs.
- Huang, Y., & Huang, H. (2024). Exploring the effect of attachment on technology addiction to generative AI chatbots: A structural equation modeling analysis. *International Journal of Human–Computer Interaction*, 1-10.
- Hutchens, J. L. (1996). How to pass the Turing test by cheating. *School of Electrical, Electronic and Computer Engineering research report TR97-05. Perth: University of Western Australia*.
- Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. Annual review of psychology, 60(1), 653-670.
- Iftikhar, Z., Ransom, S., Xiao, A., & Huang, J. (2024). Therapy as an NLP Task: Psychologists' Comparison of LLMs and Human Peers in CBT. *arXiv preprint arXiv:2409.02244*.
- Illeris, K. (2010). *The fundamentals of workplace learning: Understanding how people learn in working life*. Routledge.
- Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth*. (2018) 6(11):e12106. Available at: http://mhealth.jmir.org/2018/11/e12106/
  Izumi, K., Tanaka, H., Shidara, K., Adachi, H., Kanayama, D., Kudo, T., & Nakamura, S. (2024). Response generation for cognitive behavioral therapy with large language models: comparative study with socratic questioning. *arXiv preprint arXiv:2401.15966*
- Ivey, A. E. Ivey, M. B. and Carlos P. Zalaquett C. P. (2016). *Essentials of intentional interviewing : counseling in a multicultural world* (Third edition. ed., pp 50-129). Boston, Massachusetts: Cengage Learning. [80 s.]
- Izumi, K., Tanaka, H., Shidara, K., Adachi, H., Kanayama, D., Kudo, T., & Nakamura, S. (2024). Response Generation for Cognitive Behavioral Therapy with Large Language Models: Comparative Study with Socratic Questioning. *arXiv e-prints*, arXiv-2401.
- Jensen, A. E., Johnsen, S. P., Molbo, T., Højen, A. A., Ording, A. G., Simoni, A. H., ... & Grøntved, S. (2025). Årsrapport 2024: Dansk Center for Sundhedstjenesteforskning.
- Jiang, M., Zhao, Q., Li, J., Wang, F., He, T., Cheng, X., ... & Fu, G. (2024). A Generic Review of Integrating Artificial Intelligence in Cognitive Behavioral Therapy. *arXiv preprint arXiv:2407.19422*.
- Jin,Y.,Chandra,M.,Verma,G.,Hu,Y.,Choudhury,M.D.&Kumar,S. Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries 2023. arXiv.
- Jones, C. R., Trott, S., & Bergen, B. (2024). Comparing Humans and Large Language Models on an Experimental Protocol Inventory for Theory of Mind Evaluation (EPITOME). *Transactions of the Association for Computational Linguistics*, *12*, 803-819.

- Joseph, S. R., Hlomani, H., Letsholo, K., Kaniwa, F., & Sedimo, K. (2016). Natural language processing: A review. *International Journal of Research in Engineering and Applied Sciences*, *6*(3), 207-210.
- Jung, K., Lee, G., Huang, Y., & Chen, Y. (2025). " I've talked to ChatGPT about my issues last night.": Examining Mental Health Conversations with Large Language Models through Reddit Analysis. *arXiv preprint arXiv:2504.20320*. (Helt ny - Ikke peer reviewed).
- Jørgensen, C., Kjølbye, M., & Møhl, B. (2017). Psykoterapi. *In* Simonsen E., & Møhl, B. *Grundbog i psykiatri,* 2. udgave.(2017). Hans Reitzlers Forlag. [Pp 673-706].
- Kaplan, J., et al. (2020). Scaling laws for neural language models.
- Karlsson, H. (2011). How psychotherapy changes the brain. *Psychiatric Times*, *28*(8), 1-5.
- Karvonen, A. (2024). Emergent world models and latent variable estimation in chess-playing language models. arXiv. https://arxiv.org/abs/2403.15498
- Katznelson, N., Pless, M., Görlich, A., Graversen, L., & Sørensen, N. B. (2021). Ny udsathed: nuancer i forståelser af psykisk mistrivsel. Nordisk tidsskrift for ungdomsforskning, (2), 83-103.
- Keltner, D., & Lerner, J. S. (2010). Emotion. *Handbook of Social Psychology*.
- Keltner, D., Sauter, D., Tracy, J. L., Wetchler, E., & Cowen, A. S. (2022). How emotions, relationships, and culture constitute each other: Advances in social functionalist theory. *Cognition and Emotion*, *36*(3), 388-401.
- Khan, A. A., Alam, S., Wang, X., Khan, A. F., Neog, D. R., & Anwar, A. (2024, December). Mitigating Sycophancy in Large Language Models via Direct Preference Optimization. In *2024 IEEE International Conference on Big Data (BigData)*(pp. 1664-1671). IEEE.
- Khawaja, M., & Bélisle-Pipon, J. C. (2023). Your Robot Therapist is Not Your Therapist: Understanding the Role of AI-Powered Mental Health Chatbots.
- Khoury, B., Sharma, M., Rush, S. E., & Fournier, C. (2015). Mindfulness-based stress reduction for healthy individuals: A meta-analysis. *Journal of psychosomatic research*, *78*(6), 519-528.
- Kian, M. J., Zong, M., Fischer, K., Singh, A., Velentza, A. M., Sang, P., ... & Mataric, M. J. (2024). Can an llm-powered socially assistive robot effectively and safely deliver cognitive behavioral therapy? a study with university students. *arXiv preprint arXiv:2402.17937*.
- *Killingmo, B. (1984). Conflict and Deficit. Implications for technique. International J. of PsychoAnal., vol. 70, nr. 65, s. 65-79 (14 sider).*
- Killingmo, B. (1984). Conflict and deficit. Implications for technique. International Journal of Psychoanalysis, 70(65), 65-79.
- Kilner, J. M., & Lemon, R. N. (2013). What we know currently about mirror neurons. Current biology, 23(23), R1057-R1062.
- Kilner, J. M., Neal, A., Weiskopf, N., Friston, K. J., & Frith, C. D. (2009). Evidence of mirror neurons in human inferior frontal gyrus. Journal of Neuroscience, 29(32), 10153-10159.
- Kim, K. J., & Lipson, H. (2009, July). Towards a" theory of mind" in simulated robots. In *Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers* (pp. 2071-2076).
- Kim, K. J., & Lipson, H. (2009, July). Towards a" theory of mind" in simulated robots. In *Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers* (pp. 2071-2076)
- Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B., & Hale, S. A. (2025). Why human–AI relationships need socioaffective alignment. *Humanities and Social Sciences Communications*, *12*(1), 1-9.
- Kjølbye, M. & Møhl, B. Overblik over psykoterapiformerne. (12 sider) In: Møhl B & Kjølbye M. (red) Psykoterapiens ABC. PsykiatriFonden 2013
- Kjølbye, M., & Møhl, B. (2013). Overblik over psykoterapiformerne. I B. Møhl & M. Kjølbye (Red.), Psykoterapiens ABC (12 sider). PsykiatriFonden.
- Kolb, B., Gibb, R., & Robinson, T. E. (2003). Brain plasticity and behavior. *Current directions in psychological science*, *12*(1), 1-5.

- Korteling, J. E., van de Boer-Visschedijk, G. C., Blankendaal, R. A., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human-versus artificial intelligence. *Frontiers in artificial intelligence*, *4*, 622364.
- Korteling, J. E., van de Boer-Visschedijk, G. C., Blankendaal, R. A., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human-versus artificial intelligence. Frontiers in artificial intelligence, 4, 622364.
- Kosinski, M. (2024). Evaluating large language models in theory of mind.
- Kotek, H., Dockum, R., & Sun, D. (2023, November). Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference* (pp. 12-24).
- Kouluri, T., Macredite, R.D., & Olakitan, D. (2024). Chatbots to support young adults mental health: an exploratory study of acceptability. *ACM transactions on interactive intelligent systems (Tiis), 12*(2).

- Kramer, U. (2010). Coping and defence mechanisms: What's the difference?–Second act. *Psychology and psychotherapy: theory, research and practice*, *83*(2), 207-221.
- Krämer, N. C. (2008, September). Social effects of virtual assistants. A review of empirical results with regard to communication. In *International Workshop on Intelligent Virtual Agents* (pp. 507-508). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Kumar, B. P., & Ahmed, M. S. (2024). Beyond Clouds: Locally Runnable LLMs as a Secure Solution for AI Applications. *Digital Society*, *3*(3), 49.
- Lagercrantz, H., & Changeux, J. P. (2009). The emergence of human consciousness: from fetal to neonatal life. *Pediatric research*, *65*(3), 255-260.
- Lai, T., Shi, Y., Du, Z., Wu, J., Fu, K., Dou, Y., & Wang, Z. (2023). Supporting the demand on mental health services with AI-based conversational large language models (LLMs). *BioMedInformatics*, *4*(1), 8-33.
- Lakoff, G. (2016). Language and emotion. *Emotion Review*, *8*(3), 269-273.
- Lalor, J. P., Abbasi, A., Oketch, K., Yang, Y., & Forsgren, N. (2024). Should fairness be a metric or a model? A model-based framework for assessing bias in machine learning pipelines. *ACM Transactions on Information Systems*, *42*(4), 1-41.
- Lamm, C., & Majdandžić, J. (2015). The role of shared neural activations, mirror neurons, and morality in empathy–A critical comment. Neuroscience research, 90, 15-24.

- Laranjo et al. (2018). Conversational agents in healthcare: a systematic review.
- Larsen, M.E., et al. (2019). Using science to sell apps: Evaluation of mental health app store quality claims. *Nature:* Digital Medicine, *2*(18).
- Lave, J. & Wenger, E. (2003). *Situeret læring – og andre tekster* (s. 31-54 & 77-103). København: Hans Reitzels Forlag (49 sider).
- Lawrence, H. R., Schneider, R. A., Rubin, S. B., Matarić, M. J., McDuff, D. J., & Bell, M. J. (2024). The opportunities and risks of large language models in mental health. *JMIR Mental Health*, *11*(1), e59479.
- Lawrence, H. R., Schneider, R. A., Rubin, S. B., Matarić, M. J., McDuff, D. J., & Bell, M. J. (2024). The opportunities and risks of large language models in mental health. *JMIR Mental Health*, *11*(1), e59479.
- Lee, J., et al. (2024). Cactus: Towards Psychological Counseling Conversations Using Cognitive Behavioral Theory.
- Lee, S., Kang, J., Kim, H., Chung, K. M., Lee, D., & Yeo, J. (2024). Cocoa: Cbt-based conversational counseling agent using memory specialized in cognitive distortions and dynamic prompt. *arXiv preprint arXiv:2402.17546*.
- Lee, Y. C., Yamashita, N., Huang, Y., & Fu, W. (2020, April). " I hear you, I feel you": encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-12).
  Legaspi Jr, C. M., Pacana, T. R., Loja, K., Sing, C., & Ong, E. (2022). User perception of Wysa as a mental well-being support tool during the COVID-19 pandemic. In *Proceedings of the Asian HCI Symposium 2022* (pp. 52-57).

- Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and machines*, *17*(4), 391-444.
- Leichsenring, F., & Klein, S. (2014). Evidence for psychodynamic psychotherapy in specific mental disorders: a systematic review. *Psychoanalytic Psychotherapy*, *28*(1), 4-32.
- Leichsenring, F., & Rabung, S. (2008). Effectiveness of long-term psychodynamic psychotherapy: A meta-analysis. *Jama*, *300*(13), 1551-1565.
- Leichsenring, F., Abbass, A., Heim, N., Keefe, J. R., Kisely, S., Luyten, P., ... & Steinert, C. (2023). The status of psychodynamic psychotherapy as an empirically supported treatment for common mental disorders–an umbrella review based on updated criteria. *World Psychiatry*, *22*(2), 286-304.
- Leichsenring, F., Abbass, A., Heim, N., Keefe, J. R., Kisely, S., Luyten, P., ... & Steinert, C. (2023). The status of psychodynamic psychotherapy as an empirically supported treatment for common mental disorders–an umbrella review based on updated criteria. *World Psychiatry*, *22*(2), 286-304.
- Lemma, A., Target, M., & Fonagy, P. (2011). Brief dynamic interpersonal therapy: a clinician's guide. (pp. 63-216). Oxford: Oxford University Press. (153 s.)
- Lemoine, B. (2022). Is LaMDA sentient?—an interview. *Medium*.
- Levkovich, I., & Elyoseph, Z. (2023). Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. *JMIR mental health*, *10*, e51232.
- Levkovich, I., & Elyoseph, Z. (2023). Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. *JMIR mental health*, *10*, e51232.
- Levy, K. N., Ellison, W. D., Scott, L. N., & Bernecker, S. L. (2011). Attachment style. *Journal of clinical psychology*, *67*(2), 193-203.
- Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., ... & Xie, X. (2023). Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint arXiv:2307.11760*.
- Li, H., Zhang, R., Lee, Y. C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine*, *6*(1), 236
- Li, J., et al. (2023). Systematic Review and Meta-Analysis of AI-Based Conversational Agents for Promoting Mental Health and Well-Being.
- Li, L., Luo, Y., & Pan, T. (2024). OpenAI-o1 AB testing: Does the o1 model really do good reasoning in math problem solving? arXiv. https://arxiv.org/abs/2411.06198.

- Li, R.C., et al (2020). Developing a delivery science of artificial intelligence in healthcare.
- Li, S. Y. H., & Bressington, D. (2019). The effects of mindfulness-based stress reduction on depression, anxiety, and stress in older adults: A systematic review and meta-analysis. *International journal of mental health nursing*, *28*(3), 635-656.
- Li, Z., Cao, Y., Xu, X., Jiang, J., Liu, X., Teo, Y. S., ... & Liu, Y. (2024, April). Llms for relational reasoning: How far are we?. In *Proceedings of the 1st International Workshop on Large Language Models for Code* (pp. 119-126).
- Li, Z., Chen, G., Shao, R., Xie, Y., Jiang, D., & Nie, L. (2024). Enhancing emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.
- Lilliengren, P. (2023). A comprehensive overview of randomized controlled trials of psychodynamic psychotherapies. *Psychoanalytic Psychotherapy*, *37*(2), 117-140.
- Lilliengren, P. (2023). A comprehensive overview of randomized controlled trials of psychodynamic psychotherapies. *Psychoanalytic Psychotherapy*, *37*(2), 117-140.
- Lin, I., et al. (2024). Imbue: Improving interpersonal effectiveness through simulation and just-in-time feedback with human-language model interaction.
- Linden, D. E. (2006). How psychotherapy changes the brain–the contribution of functional neuroimaging. *Molecular psychiatry*, *11*(6), 528-538.
- Lindquist, K. A. (2021). Language and emotion: Introduction to the special issue. *Affective science*, *2*(2), 91-98.
- Lindquist, K. A., & Gendron, M. (2013). What's in a word? Language constructs emotion perception. *Emotion Review*, *5*(1), 66-71.

- Lindquist, K. A., Gendron, M., Satpute, A. B., & Lindquist, K. (2016). Language and emotion. *Handbook of emotions*, *4*, 579-594
- Linehan, M. M. (1987). Dialectical behavior therapy for borderline personality disorder: Theory and method. *Bulletin of the Menninger Clinic*, *51*(3), 261.
- Linehan, M. M., Heard, H. L., Clarkin, J., Marziali, E., & Munroe-Blum, H. (1992). Dialectical behavior therapy for borderline personality disorder. *Borderline personality disorder: Clinical and empirical perspectives*, 248-267.

- Liu, J. (2024). ChatGPT: Perspectives from human–computer interaction and psychology. *Frontiers in Artificial Intelligence*, *7*, 1418869.

- Liu, J. M., Li, D., Cao, H., Ren, T., Liao, Z., & Wu, J. (2023). Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*
- Liu, T., Giorgi, S., Aich, A., Lahnala, A., Curtis, B., Ungar, L., & Sedoc, J. (2025, April). The illusion of empathy: How ai chatbots shape conversation perception. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 13, pp. 14327-14335).
- Liu, Z., Li, H., Chen, A., Zhang, R., & Lee, Y. C. (2024, May). Understanding public perceptions of AI conversational agents: A cross-cultural analysis. In *Proceedings of the 2024 CHI conference on human factors in computing systems* (pp. 1-17).
- Loh, S. B., & Raamkumar, A. S. (2023). Harnessing Large Language Models' Empathetic Response Generation Capabilities for Online Mental Health Counselling Support. *arXiv preprint arXiv:2310.08017*.
- Losoya, S. H., & Eisenberg, N. (2001). Affective empathy. In Interpersonal sensitivity (pp. 35-58). Psychology Press.
- Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H. T., & Gurevych, I. (2023). Are emergent abilities in large language models just in-context learning?. *arXiv preprint arXiv:2309.01809*.
- Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H. T., & Gurevych, I. (2023). Are emergent abilities in large language models just in-context learning?. *arXiv preprint arXiv:2309.01809*.
- Luborsky, L., McLellan, A. T., Woody, G. E., O'Brien, C. P., & Auerbach, A. (1985). Therapist success and its determinants. *Archives of general psychiatry*, *42*(6), 602-611
- Lucas, G. M., Gratch, J., King, A., & Morency, L. P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94-100.
- Luo, X., Ghosh, S., Tilley, J. L., Besada, P., Wang, J., & Xiang, Y. (2025). "Shaping ChatGPT into my Digital Therapist": A thematic analysis of social media discourse on using generative artificial intelligence for mental health. *Digital Health*, *11*, 20552076251351088.
- Ly, K. H., Ly, A. M., & Andersson, G. (2017). A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. *Internet interventions*, *10*, 39-46.
- Lynge, J., Andersen, K. G., & Braüner, T. (2019). Udviser sociale chatbots empati?. *Aktuel Naturvidenskab*, *2019*(5), 35-37.
- Ma, J., Na, H., Wang, Z., Hua, Y., Liu, Y., Wang, W., & Chen, L. (2024). Detecting Conversational Mental Manipulation with Intent-Aware Prompting. *arXiv preprint arXiv:2412.08414*. (IFt at de kan opfange emotionelle cues ift manipulation, evt hører denne inde under TOM-færdigheder).
- Ma, Z., Mei, Y., & Su, Z. (2024, January). Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings* (Vol. 2023, p. 1105).
- Maddela, M., Ung, M., Xu, J., Madotto, A., Foran, H., & Boureau, Y. L. (2023, July). Training Models to Generate, Recognize, and Reframe Unhelpful Thoughts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 13641-13660)
- Maeda, T., & Quan-Haase, A. (2024, June). When human-AI interactions become parasocial: Agency and anthropomorphism in affective design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1068-1077).
- Malle, B. F. (2008). The relation between language and theory of mind in development and evolution. In *The evolution of language out of pre-language*(pp. 265-284). John Benjamins Publishing Company.

- Malle, B. F. (2008). The relation between language and theory of mind in development and evolution. In *The evolution of language out of pre-language*(pp. 265-284). John Benjamins Publishing Company.
- Malmqvist, L. (2024). Sycophancy in large language models: Causes and mitigations. *arXiv preprint arXiv:2411.15287*.
- Martin, D. J., Garske, J. P., & Davis, M. K. (2000). Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. *Journal of consulting and clinical psychology*, *68*(3), 438.
- Matlin, M. W., & Farmer, T. A. (2017). *Cognition* (9th ed).
- Maurya, R. K., Montesinos, S., Bogomaz, M., & DeDiego, A. C. (2025). Assessing the use of ChatGPT as a psychoeducational tool for mental health practice. *Counselling and Psychotherapy Research*, *25*(1), e12759.
- Mayer, J. D., Caruso, D. R., & Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, *27*(4), 267-298.
- Mazza, M., Pino, M. C., Mariano, M., Tempesta, D., Ferrara, M., De Berardis, D., ... & Valenti, M. (2014). Affective and cognitive empathy in adolescents with autism spectrum disorder. Frontiers in human neuroscience, 8, 791.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.
- Meng & Dai (2021): Mediated Social Support for Distress Reduction: AI Chatbots vs. Human. Proceedings of the ACM on Human-Computer Interaction
- Merleau-Ponty, M., & Bannan, J. F. (1956). What is phenomenology?. *CrossCurrents*, *6*(1), 59-70.
- Merullo, J., et al. (2023). Language Models Implement Simple Word2Vec-style Vector Arithmetic.
- Meskó, B. (2023). Prompt engineering as an important emerging skill for medical professionals: tutorial. Journal of medical Internet research, 25, e50638.
- Meskó, B. (2023). Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of medical Internet research*, *25*, e50638.
- Messer, S. B. (2002). A psychodynamic perspective on resistance in psychotherapy: Vive la résistance. *Journal of clinical psychology*, *58*(2), 157-163.
- Mikolov, T., et al. (2013). Efficient Estimation of Word Representations in Vector Space.
- Mikulincer, M., & Shaver, P. R. (2005). Attachment theory and emotions in close relationships: Exploring the attachment-related dynamics of emotional reactions to relational events. Personal relationships, 12(2), 149-168.
- Mikulincer, M., Shaver, P. R., & Berant, E. (2013). An attachment perspective on therapeutic processes and outcomes. Journal of personality, 81(6), 606-616.
- Miller, C. A. (2006). Developmental relationships between language and theory of mind. *American journal of speech-language pathology*, *15*(2), 142-154.
- Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), e2215907120.
- Moghaddam, S. R., & Honey, C. J. (2023). Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.
- Molenberghs, P., Cunnington, R., & Mattingley, J. B. (2009). Is the mirror neuron system involved in imitation? A short review and meta-analysis. Neuroscience & biobehavioral reviews, 33(7), 975-980.
- Moor, J. H. (1976). An analysis of the Turing test. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *30*(4), 249-257.
- Morency, L. P., Stratou, G., DeVault, D., Hartholt, A., Lhommet, M., Lucas, G., ... & Rizzo, A. (2015). SimSensei demonstration: a perceptive virtual human interviewer for healthcare applications. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 29, No. 1).
- Morgado, L., & Gaspar, G. (2003, September). Emotion in intelligent virtual agents: The flow model of emotion. In *International Workshop on Intelligent Virtual Agents* (pp. 31-38). Berlin, Heidelberg: Springer Berlin Heidelberg.

- Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., ... & Pollak, T. A. (2025). Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it).
- Morris, R. R., Kouddous, K., Kshirsagar, R., & Schueller, S. M. (2018). Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of medical Internet research*, *20*(6), e10148.
- Morrison, A. S., & Heimberg, R. G. (2013). Social anxiety and social anxiety disorder. *Annual review of clinical psychology*, *9*(1), 249-274.
- Munder, T., Flückiger, C., Leichsenring, F., Abbass, A. A., Hilsenroth, M. J., Luyten, P., ... & Wampold, B. E. (2019). Is psychotherapy effective? A re-analysis of treatments for depression. *Epidemiology and psychiatric sciences*, *28*(3), 268-274.
- Muran, J. C, Eubanks, F. C., Samstag, W. L. (2023). Introduction: Rupture in a Wicked and Wonderful World. In: Rupture and Repair in Psychotherapy - A Critical Process for Change (2023). Eubanks, F. C., Samstag, W. L. & Muran, J. C (Editors). American Psychological Association, pp. 3-20 https://doi.org/10.1037/0000306-001. [17 s.]
- Na, H. (2024, May). CBT-LLM: A Chinese Large Language Model for Cognitive Behavioral Therapy-based Mental Health Question Answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 2930-2940).

- Na, H., Hua, Y., Wang, Z., Shen, T., Yu, B., Wang, L., ... & Chen, L. (2025). A Survey of Large Language Models in Psychotherapy: Current Landscape and Future Directions. *arXiv e-prints*, arXiv-2502. (**Endnu ikke peer-review).**
Ng, MM, Girth, J, Minen, M, Torous, J. User Engagement in Mental Health Apps: A Review of Measurement, Reporting, and Validity. Psychiatric Service.
- Nagel, T. (1980). What is it like to be a bat?. In *The language and thought series* (pp. 159-168). Harvard University Press.
- Nair, S., et al. (2020). Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge.
- Natale, S., & Depounti, I. (2024). Artificial sociality. *Human-Machine Communication*, *7*, 83-98.
- Neumann, D. L., Chan, R. C., Boyle, G. J., Wang, Y., & Westbury, H. R. (2015). Measures of empathy: Self-report, behavioral, and neuroscientific approaches. Measures of personality and social psychological constructs, 257-289.
- Nguyen, H. M. (2025). A survey of theory of mind in large language models: Evaluations, representations, and safety risks. *arXiv preprint arXiv:2502.06470*.
- Nie, J., Shao, H., Fan, Y., Shao, Q., You, H., Preindl, M., & Jiang, X (2024). LLM-based Conversational AI Therapist for Daily Functioning Screening and Psychotherapeutic Intervention via Everyday Smart Devices. *ACM Transactions on Computing for Healthcare*.
- Nissen-Lie, H. A., Havik, O. E., Høglend, P. A., Rønnestad, M. H., & Monsen, J. T. (2015). Patient and therapist perspectives on alliance development: Therapists' practice experiences as predictors. *Clinical psychology & psychotherapy*, *22*(4), 317-327.
- Norcross, J. C., & Wampold, B. E. (2011). Evidence-based therapy relationships: research conclusions and clinical practices. *Psychotherapy*, *48*(1), 98.
- Norcross, J. C., & Wampold, B. E. (2011). What works for whom: Tailoring psychotherapy to the person. *Journal of clinical psychology*, *67*(2), 127-132.
- Norcross, J. C., & Wampold, B. E. (2018). A new therapy for each patient: Evidence-based relationships and responsiveness. *Journal of clinical psychology*, *74*(11), 1889-1906.
- Normann, N., & Morina, N. (2018). The efficacy of metacognitive therapy: a systematic review and meta-analysis. *Frontiers in psychology*, *9*, 2211.
- Normann, N., van Emmerik, A. A., & Morina, N. (2014). The efficacy of metacognitive therapy for anxiety and depression: A meta-analytic review. *Depression and anxiety*, *31*(5), 402-411.

- Norvig, R., & Russel, S. (2021). *Artificial Intelligence: A modern approach.* Pearson.

- Novak, M. (2024). How the Cold War Fueled the Rise of Artificial Intelligence. *Medium.*
- Obradovich, N., Khalsa, S. S., Khan, W. U., Suh, J., Perlis, R. H., Ajilore, O., & Paulus, M. P. (2024). Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry and Neuroscience*, *2*(1), 8.
- Omar, M., & Levkovich, I. (2024). Exploring the efficacy and potential of large language models for depression: A systematic review. *Journal of affective disorders*, S0165-0327.
- Omar, M., Soffer, S., Charney, A. W., Landi, I., Nadkarni, G. N., & Klang, E. (2024). Applications of large language models in psychiatry: a systematic review. *Frontiers in Psychiatry*, *15*, 1422807-1422807.
- Oppy, G., & Dowe, D. (2003). The turing test.
- Ortega, P. A., Kunesch, M., Delétang, G., Genewein, T., Grau-Moya, J., Veness, J., ... & Legg, S. (2021). Shaking the foundations: delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819*.
- Ottosen, M. H., & Andreasen, A. G. (2020). Børn og unges trivsel og brug af digitale medier. København: VIVE–Viden til Velfærd. Det Nationale Forsknings-og Analysecenter for Velfærd
- Ottosen, M. H., Berger, N. P., & Lindeberg, N. H. (2018). Forebyggende indsatser til unge i psykisk mistrivsel.
- Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback.

Packer, M. J. & Goicoechea (2000). Sociocultural and Constructivist Theories of Learning: Ontology, Not Just Epistemology. *Educational Psychologist 35(4)* (pp. 227-241). (14 sider).

- Paech, S. J. (2023). Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.
- Palma, R., Lam, H. C., Shrivastava, A., Karlinsey, E., Nguyen, K., Deol, P., ... & Ahmed, S. (2023, March). “Monday Feels Like Friday!”-Towards Overcoming Anxiety and Stress of Autistic Young Adults During Times of Isolation. In *International Conference on Information* (pp. 286-305). Cham: Springer Nature Switzerland.
- Pani, A., et al. (2024). Can generative artificial intelligence foster belongingness, social support, and reduce loneliness?
- Parker, S. T., & Gibson, K. R. (1979). A developmental model for the evolution of language and intelligence in early hominids. *Behavioral and Brain sciences*, *2*(3), 367-381.
- Parthasarathy, V. B., Zafar, A., Khan, A., & Shahid, A. (2024). The ultimate guide to fine-tuning llms from basics to breakthroughs: An exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *arXiv preprint arXiv:2408.13296*.
- Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, *140*, 107600.
- Pentina, I., Xie, T., Hancock, T., & Bailey, A. (2023). Consumer–machine relationships in the age of artificial intelligence: Systematic literature review and research directions. *Psychology & Marketing*, *40*(8), 1593-1614.
- Perry, J. C., & Bond, M. (2012). Change in defense mechanisms during long-term dynamic psychotherapy and five-year outcome. *American Journal of Psychiatry*, *169*(9), 916-925.
- Perry, N., Sun, C., Munro, M., Boulton, K. A., & Guastella, A. J. (2024). AI technology to support adaptive functioning in neurodevelopmental conditions in everyday environments: a systematic review. NPJ digital medicine, 7(1), 370.
- Pessoa, L. (2018). Understanding emotion with brain networks. *Current opinion in behavioral sciences*, *19*, 19-25.

- Peter, S., Riemer, K., & West, J. D. (2025). The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, *122*(22), e2415898122.
- Pfister, R., & Jud, H. (2025). Understanding and Benchmarking Artificial Intelligence: OpenAI's o3 Is Not AGI. *arXiv preprint arXiv:2501.07458*.
- Pham, K. T., Nabizadeh, A., & Selek, S. (2022). Artificial intelligence and chatbots in psychiatry. *Psychiatric Quarterly*, *93*(1), 249-253.
- Pham, K. T., Nabizadeh, A., & Selek, S. (2022). Artificial intelligence and chatbots in psychiatry. *Psychiatric Quarterly*, *93*(1), 249-253.
- Pi, Z., Vadaparty, A., Bergen, B. K., & Jones, C. R. (2024). Dissecting the Ullman variations with a SCALPEL: Why do LLMs fail at trivial alterations to the false belief task?. *arXiv preprint arXiv:2406.14737*.
- Picard, R. W. (1999, August). Affective computing for hci. In *HCI (1)* (pp. 829-833).
- Picard, R. W. (2009). Future affective technology for autism and emotion communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3575-3584.
- Pinel, J. & Barnes, S. (2018). *Biopsychology.*
- Pinker, S. (1979). Formal models of language learning. *Cognition*, *7*(3), 217-283.
- Pinker, S. (2010). The cognitive niche: Coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences*, *107*(supplement_2), 8993-8999.
- Pommerencke, L. M., Jørgensen, S. E., Pant, S. W., Skovgaard, A. M., Pedersen, T. P., & Madsen, K. R. (2023). Psykisk mistrivsel og psykisk sygdom blandt børn og unge: En undersøgelse af 0-16-årige i Region Hovedstaden.
- Prakash, A. V., & Das, S. (2020). Intelligent conversational agents in mental healthcare services: a thematic analysis of user perceptions. *Pacific Asia Journal of the Association for Information Systems*, *12*(2), 1.
- Premack, D. (2004). Is language the key to human intelligence?. *Science*, *303*(5656), 318-320.
- Preston, J. (1997). Introduction: thought as language. *Royal Institute of Philosophy Supplements*, *42*, 1-14.
- Psykiatriske diagnoser og kontakter blandt børn & unge i 2012-2022.
- Radford, A., et al. (2018). Improving language understanding by generative pre-training.
- Radford, A., et al. (2019). Language Models are unsupervised multi-task learners. Tupin, E., et al. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.
- Radford, K.A., et al. (2018). Learning transferable visual models from natural language supervision.
- Raile, P. (2024). The usefulness of ChatGPT for psychotherapists and patients. *Humanities and Social Sciences Communications*, *11*(1), 1-8
- Ran, C. (2023, December). Emotion analysis of dialogue text based on ChatGPT: a research study. In *International Conference on Algorithms, High Performance Computing, and Artificial Intelligence (AHPCAI 2023)* (Vol. 12941, pp. 806-811). SPIE
- Ranaldi, L., & Pucci, G. (2023). When large language models contradict humans? large language models' sycophantic behaviour. *arXiv preprint arXiv:2311.09410*.
- Raunak, V., Menezes, A., & Junczys-Dowmunt, M. (2021, June). The Curious Case of Hallucinations in Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1172-1183).
- rdito, R. B., & Rabellino, D. (2011). Therapeutic alliance and outcome of psychotherapy: historical excursus, measurements, and prospects for research. *Frontiers in psychology*, *2*, 270.

  Reinecke, M. G., Ting, F., Savulescu, J., & Singh, I. (2025, February). The double-edged sword of anthropomorphism in llms. In *Proceedings* (Vol. 114, No. 1, p. 4). MDPI.

- Relational Agent for Mental Health | Woebot Health. Available at: https://woebothealth.com/
- Rochat, L., Manolov, R., & Billieux, J. (2018). Efficacy of metacognitive therapy in improving mental health: A meta-analysis of single-case studies. *Journal of clinical psychology*, *74*(6), 896-915.
- Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of consulting psychology*, *21*(2), 95.
- Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Counseling Psychology, 21*, 95-103 [9 s.]
- Rogers, C. R. (1975). Empathic: An unappreciated way of being. *Journal of Counseling Psychology, 5*, 2-10 [9 s.]
- Rosenberg N.K., Mørch, M.M. & Arendt, M. Kognitiv adfærdsterapi – teori og metoder (kapitel 1)(30 sider). In: Arendt, M. & Rosenberg, N.K. (red) Kognitiv Terapi – Nyeste Udvikling. Hans Reitzels Forlag 2012.
- Rudd, M. D., & Joiner, T. (1997). Countertransference and the therapeutic relationship: A cognitive perspective. *Journal of Cognitive Psychotherapy*, *11*(4), 231.
- Russel, N., & Norvig, S., (2021). *Artificial Intelligence: A modern approach.*
- Rządeczka, M., Sterna, A., Stolińska, J., Kaczyńska, P., & Moskalewicz, M. (2025). The Efficacy of Conversational AI in Rectifying the Theory-of-Mind and Autonomy Biases: Comparative Analysis. *JMIR Mental Health*, *12*(1), e64396
- Sabour, S., Liu, S., Zhang, Z., Liu, J., Zhou, J., Sunaryo, A., ... & Huang, M. (2024, August). EmoBench: Evaluating the Emotional Intelligence of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5986-6004).: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.
- Sackett, C., Harper, D., & Pavez, A. (2024). Do we dare use generative AI for mental health?. *IEEE Spectrum*, *61*(6), 42-47.
- Safran, J. D., Muran, J. C., & Eubanks-Carter, C. (2011). Repairing alliance ruptures. *Psychotherapy*, *48*(1), 80.
- Safran, J. D., Muran, J. C., Samstag, L. W., & Stevens, C. (2001). Repairing alliance ruptures. *Psychotherapy: Theory, Research, Practice, Training*, *38*(4), 406
- Safran, J.D., Muran, J.C., & Eubanks-Carter, C. (2011). Repairing Alliance Ruptures. *Psychotherapy, 48*(1).
- Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination, cognition and personality*, *9*(3), 185-211.
- Sap, M., Le Bras, R., Fried, D., & Choi, Y. (2022, December). Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 3762-3780).
- Schafer, R. (1968). The mechanisms of defence. *The International Journal of Psycho-Analysis*, *49*, 49.
- Scherer, K. R. (2005). What are emotions? And how can they be measured?. *Social science information*, *44*(4), 695-729.
- Scherer, K. R., & Ekman, P. (2014). *Approaches to emotion*. Psychology Press.
- Schipolowski, S., Wilhelm, O., & Schroeders, U. (2014). On the nature of crystallized intelligence: The relationship between verbal ability and factual knowledge. *Intelligence*, *46*, 156-168.
- Schrammen, E., Roesmann, K., Rosenbaum, D., Redlich, R., Harenbrock, J., Dannlowski, U., & Leehr, E. J. (2022). Functional neural changes associated with psychotherapy in anxiety disorders–A meta-analysis of longitudinal fMRI studies. *Neuroscience & Biobehavioral Reviews*, *142*, 104895.
- Sclar et al. (2023). Minding Language Models' (Lack of) Theory of Mind:A Plug-and-Play Multi-Character Belief Tracker
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424.
- Searle, J. R. (2000). Consciousness, free action and the brain. *Journal of Consciousness Studies*, *7*(10), 3-22.Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and machines*, *17*(4), 391-444.

- Sedlakova, J., & Trachsel, M. (2023). Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent?. *The American Journal of Bioethics*, *23*(5), 4-13.
- Shah, H., Warwick, K., Vallverdú, J., & Wu, D. (2016). Can machines talk? Comparison of Eliza with modern dialogue systems. *Computers in Human Behavior*, *58*, 278-295.
- Shaikh, T.A.H.; Mhetre, M. Autonomous AI Chat Bot Therapy for Patient with Insomnia. In Proceedings of the 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 7–9 April 2022; pp. 1–5.
- Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., ... & Shwartz, V. (2024, March). Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2257-2273).
- Shapira, N., Zwirn, G., & Goldberg, Y. (2023, July). How well do large language models perform on faux pas tests?. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 10438-10451)
- Sharma, A., Rushton, K., Lin, I. W., Nguyen, T., & Althoff, T. (2024, May). Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-29).
- Sharma, A., Rushton, K., Lin, I., Wadden, D., Lucas, K., Miner, A., ... & Althoff, T. (2023, July). Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistic(Volume 1: Long Papers)* (pp. 9977-10000).
- Sharma, M., et al. (2023). Towards understanding sycophancy in language models.
- Sharma, M., Savage, C., Nair, M., Larsson, I., Svedberg, P., & Nygren, J. M. (2022). Artificial intelligence applications in health care practice: scoping review. *Journal of medical Internet research*, *24*(10), e40238.
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., ... & Perez, E. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.
- Sharma,A.,Lin,I.W.,Miner,A.S.,Atkins,D.C.&Althoff,T.Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nature Machine Intelligence 5, 46–57 (2023).
- Shen, H., Li, Z., Yang, M., Ni, M., Tao, Y., Yu, Z., ... & Hu, B. (2024, December). Are Large Language Models Possible to Conduct Cognitive Behavioral Therapy?. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 3695-3700). IEEE
- Shum, H. Y., He, X. D., & Li, D. (2018). From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 10-26.
- Shumanov M, Johnson L. Making conversations with chatbots more personalized. *Comput Hum Behav.* (2021) 117:106627. doi: 10.1016/j.chb.2020.106627.
- Siddals, S., Torous, J., & Coxon, A. (2024). “It happened to be the perfect thing”: experiences of generative AI chatbots for mental health. *npj Mental Health Research*, *3*(1), 48.
- Skjuve, M., Følstad, A., & Brandtzaeg, P. B. (2023, July). The user experience of ChatGPT: Findings from a questionnaire study of early users. In *Proceedings of the 5th international conference on conversational user interfaces* (pp. 1-10).
- Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies*, *149*, 102601
- Smith, J., et al. (2006). The History of Artificial Intelligence. University of Washington.

- Song et al (2024). The typing cure.
- Spaulding, S. (2017). Cognitive empathy. In The Routledge handbook of philosophy of empathy (pp. 13-21). Routledge.

- Sroufe, L. A., & Fleeson, J. (2013). Attachment and the construction of relationships. In Relationships and development (pp. 51-71). Psychology Press.
- Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., ... & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, *3*(1), 12.
- Stanovich, K. E. (1993). Does reading make you smarter? Literacy and the development of verbal intelligence. *Advances in child development and behavior*, *24*, 133-180.
- Steiner, J. (1996). The aim of psychoanalysis in theory and in practice. *The International journal of psycho-analysis*, *77*(6), 1073.
- Stiennon, N., et al. (2020). Learning to summarize with human feedback.
- Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*(7), 1285-1295.
- Stratou, G., Morency, L. P., DeVault, D., Hartholt, A., Fast, E., Lhommet, M., ... & Rizzo, A. (2015, September). A demonstration of the perception system in SimSensei, a virtual human application for healthcare interviews. In *2015 international conference on affective computing and intelligent interaction (ACII)* (pp. 787-789). IEEE.
- Street, W., Siy, J. O., Keeling, G., Baranes, A., Barnett, B., McKibben, M., ... & Dunbar, R. I. (2024). Llms achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870*
- Straarup, K.N. Kognitiv adfærdsterapi ved affektive lidelser (kapitel 4)(35 sider). In: Arendt, M. & Rosenberg, N.K. (red) Kognitiv Terapi – Nyeste Udvikling. Hans Reitzels Forlag 2012.
- Sun, Y., & Wang, T. (2025). Be friendly, not friends: How llm sycophancy shapes user trust. *arXiv preprint arXiv:2502.10844*.
- Sundar, S. S., & Kim, J. (2019, May). Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems* (pp. 1-9
- Sundhedsstyrelsens rapport om børn & unges trivsel fra SDU 2023. Hammer et al.
- Swales, Heidi L. Heard, J. Mark G. Williams, M. (2000). Linehan's dialectical behaviour therapy (DBT) for borderline personality disorder: Overview and adaptation. *Journal of Mental Health*, *9*(1), 7-23.
- Sweeney, C., Potts, C., Ennis, E., Bond, R., Mulvenna, M. D., O'neill, S., ... & Mctear, M. F. (2021). Can chatbots help support a person's mental health? Perceptions and views from mental healthcare professionals and experts. *ACM Transactions on Computing for Healthcare*, *2*(3), 1-15.
- Swift, J. K., Tompkins, K. A., & Parkin, S. R. (2017). Understanding the client's perspective of helpful and hindering events in psychotherapy sessions: A micro-process approach. *Journal of clinical psychology*, *73*(11), 1543-1555.
- Sørensen, N. U., & Nielsen, M. L. (2025). Mistrivsel, acceleration og præstationskultur i unges overgang fra hjemme-til udeboende. Nordisk tidsskrift for ungdomsforskning, 6(1), 1-21.
- Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., ... & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *Journal of medical Internet research*, *22*(3), e16235.
- Tal, A., Elyoseph, Z., Haber, Y., Angert, T., Gur, T., Simon, T., & Asman, O. (2023). The artificial third: utilizing ChatGPT in mental health. *The American Journal of Bioethics*, *23*(10), 74-77
- Tanggaard, L. & Nielsen, K. (2018). *Pædagogisk psykologi – en grundbog*. Klim (270 sider)
- Tanana, M. J., Soma, C. S., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2019). Development and evaluation of ClientBot: Patient-like conversational agent to train basic counseling skills. *Journal of medical Internet research*, *21*(7), e12529.
- Tate, S., Fouladvand, S., Chen, J. H., & Chen, C. Y. A. (2023). The ChatGPT therapist will see you now: Navigating generative artificial intelligence's potential in addiction medicine research and patient care. *Addiction*, *118*(12), 2249-2251.

- Thomason, K.K. (2025). "How Emotional Manipulation Causes ChatGPT Psychosis". *Psychology Today*
- Thorgaard, L. (1998). Psykoterapeutens udvikling gennem modoverføring. *Psyke & Logos*, *19*(1)
- Tomasello, M. (2003). The key is social cognition. *Language in mind: Advances in the study of language and thought*, 47-57.
- Tudor Car, L., Dhinagaran, D. A., Kyaw, B. M., Kowatsch, T., Joty, S., Theng, Y. L., & Atun, R. (2020). Conversational agents in health care: scoping review and conceptual analysis. *Journal of medical Internet research*, *22*(8), e17158.
- Turella, L., Pierno, A. C., Tubaldi, F., & Castiello, U. (2009). Mirror neurons in humans: Consisting or confounding evidence?. Brain and language, 108(1), 10-21.
- Turing, A. M. (1950). Computing Machinery and Intelligence. Mind, 59(236), 433–460.

- Turpin, M., Michael, J., Perez, E., & Bowman, S. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, *36*, 74952-74965.
- Turpin, M., Michael, J., Perez, E., & Bowman, S. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, *36*, 74952-74965.
- Ullah, S., Han, M., Pujar, S., Pearce, H., Coskun, A., & Stringhini, G. (2024, May). Llms cannot reliably identify and reason about security vulnerabilities (yet?): A comprehensive evaluation, framework, and benchmarks. In *2024 IEEE symposium on security and privacy (SP)* (pp. 862-880). IEEE.
- Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.
- Vafa, K., Chen, J., Rambachan, A., Kleinberg, J., & Mullainathan, S. (2024). Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, *37*, 26941-26975.
- Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *Can J Psychiatry*. (2019) 64(7):456–64. doi: 10.1177/0706743719828977
- Valmeekam, K., Olmo, A., Sreedharan, S., & Kambhampati, S. (2022, November). Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop.*
- van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2024). Ai sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*.
- van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2024). Ai sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*.
- van Duijn, M. J., van Dijk, B. M. A., Kouwenhoven, T., Valk, W. D., Spruit, M., van der Putten, P. W. H., ... & Deng, S. (2023, December). Theory of mind in large language models: examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)* (pp. 389-402). Association for Computational Linguistics
- Vaswani, A., et al. (2017). Attention is All You Need.

- Walczak, S. (2018). Artificial neural networks.

- Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World psychiatry*, *14*(3), 270-277.

- Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry, 14*(3), 270-277. [8 s.]
- Wampold, B. E. (2017). What should we practice? A contextual model for how psychotherapy works. *The cycle of excellence: Using deliberate practice to improve supervision and training*, 49-65.
- Wampold, B. E., Baldwin, S. A., & Imel, Z. E. (2017). What characterizes effective therapists?.
- Wampold, B. E., Flückiger, C., Del Re, A. C., Yulish, N. E., Frost, N. D., Pace, B. T., ... & Hilsenroth, M. J. (2017). In pursuit of truth: A critical examination of meta-analyses of cognitive behavior therapy. *Psychotherapy Research*, *27*(1), 14-32.
- Wampold, B. E., Flückiger, C., Del Re, A. C., Yulish, N. E., Frost, N. D., Pace, B. T., ... & Hilsenroth, M. J. (2017). In pursuit of truth: A critical examination of meta-analyses of cognitive behavior therapy. *Psychotherapy Research*, *27*(1), 14-32.
- Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. N. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empiricially," all must have prizes.". *Psychological bulletin*, *122*(3), 203.
- Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., ... & Zhang, S. (2023). Prompt engineering for healthcare: Methodologies and applications. arXiv preprint arXiv:2304.14670.
- Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., ... & Zhang, S. (2023). Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*.
- Wang, L., Bhanushali, T., Huang, Z., Yang, J., Badami, S., & Hightow-Weidman, L. (2025). Evaluating Generative AI in Mental Health: Systematic Review of Capabilities and Limitations. *JMIR mental health*, *12*(1), e70014.
- Wang, Q., Tang, Z., & He, B. (2025). From ChatGPT to DeepSeek: Can LLMs simulate humanity?. *arXiv preprint arXiv:2502.18210*.
- Wang, R., Milani, S., Chiu, J., Zhi, J., Eack, S., Labrum, T., ... & Chen, Z. (2024, November). PATIENT-$\psi$: Using Large Language Models to Simulate Patients for Training Mental Health Professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 12772-12797).
- Wang, T., Wang, D., Li, B., Ma, J., Pang, X. S., & Wang, P. (2023). The impact of anthropomorphism on chatgpt actual use: Roles of interactivity, perceived enjoyment, and extraversion. *Perceived Enjoyment, and Extraversion*.
- Wang, Y., Wang, Y., Crace, K., & Zhang, Y. (2025, April). Understanding Attitudes and Trust of Generative AI Chatbots for Social Anxiety Support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-21
- Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language* (pp. 17-60). CRC Press.
- Warwick, K., & Shah, H. (2016). Can machines think? A report on Turing test experiments at the Royal Society. *Journal of experimental & Theoretical artificial Intelligence*, *28*(6), 989-1007.
- Wei, J., Huang, D., Lu, Y., Zhou, D., & Le, Q. V. (2023). Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.
- Weinberg, E., Seery, E., & Plakun, E. M. (2018). A psychodynamic approach to treatment resistance. *Treatment Resistance in Psychiatry: Risk Factors, Biology, and Management*, 295-310.
- Weinberger, J. (1993). Common factors in psychotherapy. In *Comprehensive handbook of psychotherapy integration* (pp. 43-56). Boston, MA: Springer US.
- Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45.
- Wells, A. (2011). *Metacognitive therapy for anxiety and depression*. Guilford press.
- Wells, A. (2011). Metacognitive therapy. *Acceptance and mindfulness in cognitive behavior therapy: Understanding and applying the new therapies*, 83-108.
- Wells, A., & King, P. (2006). Metacognitive therapy for generalized anxiety disorder: An open trial. *Journal of behavior therapy and experimental psychiatry*, *37*(3), 206-212.

- Wells, A., Fisher, P., Myers, S., Wheatley, J., Patel, T., & Brewin, C. R. (2009). Metacognitive therapy in recurrent and persistent depression: A multiple-baseline study of a new treatment. *Cognitive therapy and research*, *33*(3), 291-300.

Westberg, K. H., Nyholm, M., Nygren, J. M., & Svedberg, P. (2022). Mental health problems among young people—a scoping review of help-seeking. *International journal of environmental research and public health*, *19*(3), 1430.

- Wester, J., De Jong, S., Pohl, H., & Van Berkel, N. (2024). Exploring people's perceptions of LLM-generated advice. *Computers in Human Behavior: Artificial Humans*, *2*(2), 100072.
- Williams et al, 2024. Targeted manipulation and deception emerge when optimizing llms for user feedback.
- Wisniewski, H., Liu, G., Henson, P., Vaidyam, A., Hajratalli, N. K., Onnela, J. P., & Torous, J. (2019). Understanding the quality, effectiveness and attributes of top-rated smartphone health apps. *BMJ Ment Health*, *22*(1), 4-9.
- Woebot Health—What Powers Woebot. Woebot Health. Available at: https://woebothealth.com/what-powers-woebot/
- Wolfram, S. (2023). What is ChatGPT Doing … and Why Does It Work? Stephen Wolfram Writings.

Workman, L., & Reader, W. (2019). *Evolutionary Psychology.*

- Wong, Q. J., & Rapee, R. M. (2016). The aetiology and maintenance of social anxiety disorder: A synthesis of complementary theoretical models and formulation of a new integrated model. *Journal of affective disorders*, *203*, 84-100.
- Wright, J. H., & Wright, A. S. (1997). Computer-assisted psychotherapy. *The Journal of Psychotherapy Practice and Research*, *6*(4), 315.
- Wu, Y., He, Y., Jia, Y., Mihalcea, R., Chen, Y., & Deng, N. (2023, December). Hi-ToM: A Benchmark for Evaluating Higher-Order Theory of Mind Reasoning in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 10691-10706).
- Xiao, M., Xie, Q., Kuang, Z., Liu, Z., Yang, K., Peng, M., ... & Huang, J. (2024). HealMe: Harnessing Cognitive Reframing in Large Language Models for Psychotherapy. *CoRR*.
- Xiao, Y., & Wang, W. Y. (2021, January). On Hallucination and Predictive Uncertainty in Conditional Language Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.
- Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., ... & Wang, D. (2024). Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *8*(1), 1-32.
- y Arcas, B. A. (2022). Do large language models understand us?. *Daedalus*, *151*(2), 183-197.
- Yang, J., Lim, D., Lee, S., Lee, S., & Oh, U. (2025). Enhancing emotions in positive way: Llm-based ai using cognitive behavioral therapy for emotional support. *International journal of advanced smart convergence*, *14*(1), 247-256.
- Yang, K., et al. (2023). Mentallama: Intepretable mental health analysis on social media with large language models.
- Yang, Y., Xiong, S., Payani, A., Shareghi, E., & Fekri, F. (2024). Can LLMs Reason in the Wild with Programs?. *arXiv preprint arXiv:2406.13764*.
- Yang, Z., Li, L., Lin, K., Wang, J., Lin, C. C., Liu, Z., & Wang, L. (2023). The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, *9*(1), 1
- Yankouskaya, A., Liebherr, M., & Ali, R. (2025). Can ChatGPT be addictive? A call to examine the shift from support to dependence in AI conversational large language models. *Human-Centric Intelligent Systems*, 1-13.
- Yeh, P. L., Kuo, W. C., Tseng, B. L., & Sung, Y. H. (2025). Does the AI-driven Chatbot Work? Effectiveness of the Woebot app in reducing anxiety and depression in group counseling courses and student acceptance of technological aids. *Current Psychology*, 1-13.
- Yihan, L., et al. (2023). A Comprehensive Survey of AI-Generated Content: A History of Generative AI from GAN to ChatGPT.

- You, W. (2023). Backdoor attacks and defenses in natural language processing. University of Oregon.
- Yuan, H., Che, Z., Zhang, Y., Li, S., Yuan, X., Huang, L., ... & Luo, S. (2025). The cultural stereotype and cultural bias of ChatGPT. *Journal of Pacific Rim Psychology*, *19*, 18344909251355673.
- Zamfirescu-Pereira, J., et al. (2023). Why Johnny can't prompt: how non-ai experts try (and fail) to design llm prompts.
- Zarbo, C., Tasca, G. A., Cattafi, F., & Compare, A. (2016). Integrative psychotherapy works. *Frontiers in psychology*, *6*, 2021.
- Zentall TR. Animal Intelligence. In: Sternberg RJ, ed. The Cambridge Handbook of Intelligence. Cambridge Handbooks in Psychology. Cambridge University Press; 2020:397-427.
- Zhang, B., Liu, Z., Cherry, C., & Firat, O. (2024). When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*.
- Zhang, C., Li, R., Tan, M., Yang, M., Zhu, J., Yang, D., ... & Hu, X. (2024, August). CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling. In Findings of the Association for Computational Linguistics ACL 2024 (pp. 13947-13966).
- Zhang, H., et al. (2023). From Turing to Transformers: A Comprehensive Review and Tutorial on the Evolution and Applications of Generative Transformer Models.

- Zhang, M., Yang, X., Zhang, X., Labrum, T., Chiu, J. C., Eack, S. M., ... & Chen, Z. Z. (2024). CBT-Bench: Evaluating Large Language Models on Assisting Cognitive Behavior Therapy. *arXiv preprint arXiv:2410.13218*.
- Zhang,Y.,Li,Y.,Cui,L.,Cai,D.,Liu,L.,Fu,T.,Huang,X.,Zhao,E.,Zhang,Y.,Chen,Y., et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv preprint arXiv:2309.01219 (2023).
- Zhao, J., Ding, Y., Jia, C., Wang, Y., & Qian, Z. (2024). Gender bias in large language models across multiple languages. *arXiv preprint arXiv:2403.00277*.
- Zhou et al. (2023). Syntetic lies: understanding ai-generated mis-information and evaluating algorithmic and human solutions. In *proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.*
- Zhou, C., Neubig, G., Gu, J., Diab, M. T., Guzmán, F., Zettlemoyer, L., & Ghazvininejad, M. (2021, January). Detecting Hallucinated Content in Conditional Neural Sequence Generation. In *ACL/IJCNLP (Findings)*.
- Zhu, W., Zhang, Z., & Wang, Y. (2024, July). Language Models Represent Beliefs of Self and Others. In *International Conference on Machine Learning* (pp. 62638-62681). PMLR.
- Ziems,C.,Li,M.,Zhang,A.&Yang,D.Inducing positive perspectives with text reframing. arXiv preprint arXiv:2204.02952 (2022)
- Østergaard, S. D. (2023). Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis?. *Schizophrenia Bulletin*, *49*(6), 1418-1419.