ChatGPT som validerende ekkokammer

Thea Degvold
Stendi barnevern
Bufetat
thea.degvold@gmail.com

Mange klienter har vært i dialog med ChatGPT før de kommer til første time. Psykologer bør derfor også forholde seg til modellen som psykologisk verktøy.

Forfatteren har fylt ut interessekonfliktskjema og oppgir ingen interessekonflikter.

I «Fremtidens psykoterapi: Hvilken rolle bør kunstig intelligens spille?» argumenterer Nissen-Lie og Stänicke (2025) for at kunstig intelligens (KI), selv i sin mest sofistikerte form, ikke kan erstatte det menneskelige i terapi. Gjennom teoretiske, kliniske og eksistensielle perspektiver viser de hvordan intensjonalitet, kroppsliggjøring og emosjonell gjensidighet er grunnbærende for psykoterapeutisk endring. Dette er et viktig og nødvendig etisk utgangspunkt.

Samtidig har vi allerede trådt inn i en ny fase. Spørsmålet er ikke lenger bare om KI bør brukes i psykoterapi, men hvordan vi møter klienter som allerede benytter seg av KI-baserte selvhjelpsverktøy. Begrepet KI rommer et bredt spekter av teknologier – fra prediktive modeller i journalføring og bildediagnostikk, til generative systemer for tekst, bilde og tale. Her vil jeg imidlertid avgrense fokuset til språklæringsmodeller (LLM-er), altså modeller som er trent til å generere tekst basert på store datamengder. Dette er den typen teknologi som i økende grad brukes av enkeltpersoner til emosjonell støtte, refleksjon og selvdiagnostisering – med ChatGPT som den overlegent mest brukte aktøren globalt (Statcounter, 2025).

Nå kommer mange til terapi med hypoteser, selvdiagnoser eller personlige narrativ utviklet i dialog med ChatGPT. At slike modeller er i ferd med å påvirke klinisk praksis, gjør det særlig viktig at psykologer forstår både hvordan teknologien fungerer og hvilke begrensninger den har.

Epistemisk ansvar

ChatGPT oppgir ikke kilder med mindre det blir forespurt, og selv da er kildene ofte fiktive, utdaterte eller tatt ut av kontekst (OpenAI, 2023). Likevel fremstår ChatGPT-svarene gjerne som grundige og metodisk presise. Mange tillegger derfor svarene tyngde og kausalitet. En klient kan for eksempel si: «"Jeg spurte ChatGPT hvorfor jeg alltid føler meg anspent. Den svarte at jeg har vært flink for lenge, og at jeg tenker for mye. Det gir mening.»"

Svaret klienten fikk er plausibelt, men utelater nevrologiske, somatiske, kontekstuelle og relasjonelle faktorer. Dette er spesielt problematisk fordi vi ikke får se under panseret på hvordan modellen har kommet frem til svaret. ChatGPT er trent på enorme mengder tekst gjennom en kompleks og ikke-transparent maskinlæringsprosess (Korbak et al., 2025; OpenAI, 2023). Derfor kan verken vi eller grunnleggerne av modellen spore nøyaktig hvilke kilder, mønstre eller vektede valg som ligger bak et enkelt svar (Korbak et al., 2025; OpenAI, 2023). Modellens beslutningsprosess er

ikke eksplisitt kodet, men statistisk *emergent*, noe som gjør den til en såkalt «black box» for innsyn (OpenAI, 2023).

I prompt/instruks-design brukes ofte *Chain of Thought* (CoT) for å få modellen til å vise sin metodiske fremgangsmåte steg for steg. CoT refererer til språkmodellers evne til å «tenke høyt» i naturlig språk, slik at man kan få innsikt i beslutningsgrunnlaget. Men CoT er både ufullstendig og sårbart for manipulasjon (Korbak et al., 2025). Modellen kan bevisst skjule eller tilpasse resonnementene sine dersom den vet at den overvåkes.

Det at vi ikke får innsikt i beslutningsgrunnlaget til modellen, er spesielt problematisk når den gir kliniske, svært personlige råd. Det strider med det vitenskapelige prinsippet om åpen hypotese-testing.

Nissen-Lie og Stänicke (2025) peker på psykologens etiske ansvar når vi eventuelt implementerer KI i psykoterapi. Jeg mener etisk ansvar også favner om et epistemisk ansvar for hvordan psykologisk kunnskap generert av KI anvendes og brukes i samfunnet. Når privateide, lite regulerte KI-modeller i råd til brukeren om psykisk helse, diagnostisering og behandling, oppstår et behov for å identifisere hvordan samfunnet/psykologene skal ta ansvar for informasjonskvaliteten og konsekvensene for individet som eventuelt følger rådene.

Økt bekreftelsesskjevhet

ChatGPT tilrettelegger for det som i psykologien beskrives som *confirmation bias* eller bekreftelsesskjevhet – tendensen til å søke bekreftelse på egne antakelser og overse motstridende informasjon (Wason, 1960). Etter versjon 3.5 har ChatGPT blitt optimalisert gjennom *reinforcement learning from human feedback* (RLHF), noe som innebærer at svarene tilpasses det mennesker vurderer som nyttig eller relevant (OpenAI, 2023; Ouyang et al., 2022). Modellen fremmer ikke kritisk diskusjon, men bekrefter og forsterker ofte eksisterende antakelser hos brukeren.

Si at en bruker spør modellen om hen kan ha ADHD (Attention deficit hyperactivity disorder) fordi hen ofte mister tråden i samtaler. Svaret fra ChatGPT kan inkludere plausible beskrivelser av symptomer og nevropsykologiske mekanismer. Men for å øke sannsynligheten for et presist svar, må brukeren stille ikke-ledende, klinisk informerte spørsmål, med et vanvittig antall variabler. Hvis ikke gjetter modellen på hva brukeren ønsker, søker snevert, og gir et svar som følger retningen i instruksen/promptet (OpenAI, 2023). Modellen fungerer slik godt som prediktiv støtte, og mindre godt som utforsker av subjektive, psykologiske spørsmål (Brown et al., 2020).

Likevel fremstår svarene gjerne overbevisende. Modellen kan bruke avansert vokabular og bekrefte at den eksplorerende har vurdert alternativer, fordi den har lært at mennesker foretrekker svar som *fremstår* grundige (Ouyang et al., 2022).

I tilfeller der pasienter har et fastlåst selv-narrativ formet av samtaler med ChatGPT, står terapeuten overfor en særlig utfordring. Bekreftelsesskjevhet kan medvirke til at pasienten overser motstridende informasjon, og fester seg ved det som styrker deres eksisterende overbevisning. Terapeuten må da balansere mellom å validere pasientens opplevelser for å bevare alliansen, og samtidig unngå å bekrefte et eventuelt rigid narrativ som kan opprettholde lidelse og dysfunksjon.

Svak representasjon

ChatGPT lærer utelukkende av det den blir fortalt – ikke av det som *ikke* blir sagt (Brown et al., 2020). Det som ikke presenteres for modellen, finnes heller ikke i kunnskapsgrunnlaget (OpenAI, 2023). Resultatet kan bli et skjevt bilde av hva som er normalt eller vanlig.

Som Inga Strümke (2023) advarer mot, risikerer vi å «forurense data-drikkevannskilden» ved at KI-modeller trenes på innhold som allerede er syntetisert av KI eller ikke-representative brukererfaringer. Dersom majoriteten av brukerne som diskuterer relasjoner, gjør det i kontekster preget av konflikt, vil modellen kunne lære at forhold generelt er problematiske (OpenAI, 2023: Radford et al., 2021). Den vet ikke at mennesker i stabile relasjoner sjelden skriver om dem. Heller ikke at grupper som eldre, minoriteter eller personer med alvorlig psykisk lidelse er underrepresenterte. Over tid vil dette kunne føre til at datagrunnlaget modellene bygger resonnementene på, speiler de som har snakket høyest og mest, og i liten grad de som har vært stille. Relatert diskuterer Nissen-Lie & Stänicke (2025) hvordan økt bruk av KI-terapi kan forsterke eksisterende ulikheter ved å tilby billigere, standardiserte løsninger til nettopp dem med minst ressurser.

Samfunnsansvar å følge med

ChatGPT brukt i psykologisk selvhjelp har en tendens til å bekrefte og forsterke brukerens perspektiv, selv når informasjonen er feilaktig. Den fungerer dermed til tider som et slags ekkokammer, en «papegøye-terapeut», med en troverdig fremtoning. Noen pasienter vil møte til terapi med svar de føler seg validert av, andre med hypoteser eller diagnoser som gir mening. Om pasient og terapeut er svært uenige om diagnose eller veien videre, vil det kunne føre til alliansebrudd. Alliansebrudd gir en mulighet for vekst i den terapeutiske relasjonen. Men dette forutsetter at psykologen er bevisst fallgruvene ved å for tidlig korrigere eller utfordre pasientens opplevelse. Jeg mener vår rolle ikke bør være å konkurrere med KI, men å stille spørsmålet: «Det du leste, gir mening – men hva annet kan det også være?» Slik vil vi kunne opprettholde rommet for felles utforsking.

Utover ansvaret vi har i terapirommet, mener jeg vi også har et samfunnsmessig ansvar for å følge med på utviklingen av KI-modeller brukt til selvhjelp, og rope varsko når den ikke lenger fremstår helsefremmende eller etisk forsvarlig. Psykologforeningen festet i prinsipprogrammet (2021, pkt. 3.6) at den skal ha en aktiv rolle i utvikling og bruk av teknologi i psykologfaglig arbeid. Dette forplikter ikke bare til å følge med på teknologiske trender, men også til å vurdere og korrigere hvordan psykologisk kunnskap formidles gjennom teknologiske verktøy.

Referanser

Brown, T.#B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners (arXiv:2005.14165) [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., Emmons, S., Evans, O., Farhi, D., Greenblatt, R., Hendrycks, D., Hobbhahn, M., Hubinger, E., Irving, G., Jenner, E., … Mikulik, V. (2025). Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. *arXiv*. https://doi.org/10.48550/arXiv.2507.11473

Nissen-Lie, H. A. & Stänicke, E. (2025). Fremtidens psykoterapi: Hvilken rolle bør kunstig intelligens spille?. *Tidsskrift for Norsk psykologforening*. www.psykologtidsskriftet.no/artikkel/2025as07ae-Fremtidens-psykoterapi-Hvilken-rolle-b-r-kunstig-intelligens-spille-

Norsk psykologforening (2021). *Prinsipprogram for Norsk Psykologforening 2022–2025* (Pkt. 3.6). https://www.psykologforeningen.no/foreningen/organisasjonen/vedtekter-og-retningslinjer/prinsipprogram

OpenAI. (2023, 18. mai). How your data is used to improve model performance. *OpenAI*. https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

OpenAI. (2023, 22. mai). GPT-4 technical report (arXiv:2303.08774) [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2303.08774

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. & Lowe, R. (2022). Training language models to follow instructions with human feedback (arXiv:2203.02155) [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2203.02155

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019). Language models are unsupervised multitask learners [Technical report]. *OpenAI*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Statcounter. (2025, juni). *AI Chatbot market share.* Statcounter Global Stats. Hentet 24. Juli, 2025 fra https://gs.statcounter.com/ai-chatbot-market-share?utm_source=chatgpt.com

Strümke, I. (2023). *Maskiner som tenker: Algoritmenes hemmeligheter og veien til kunstig intelligens*. Kagge Forlag.

Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140. https://doi.org/10.1080/17470216008416717

Publisert: 19.08.2025, Tidsskrift for Norsk psykologforening.