Tidsskriftet Tidsskrift for Den norske legeforening

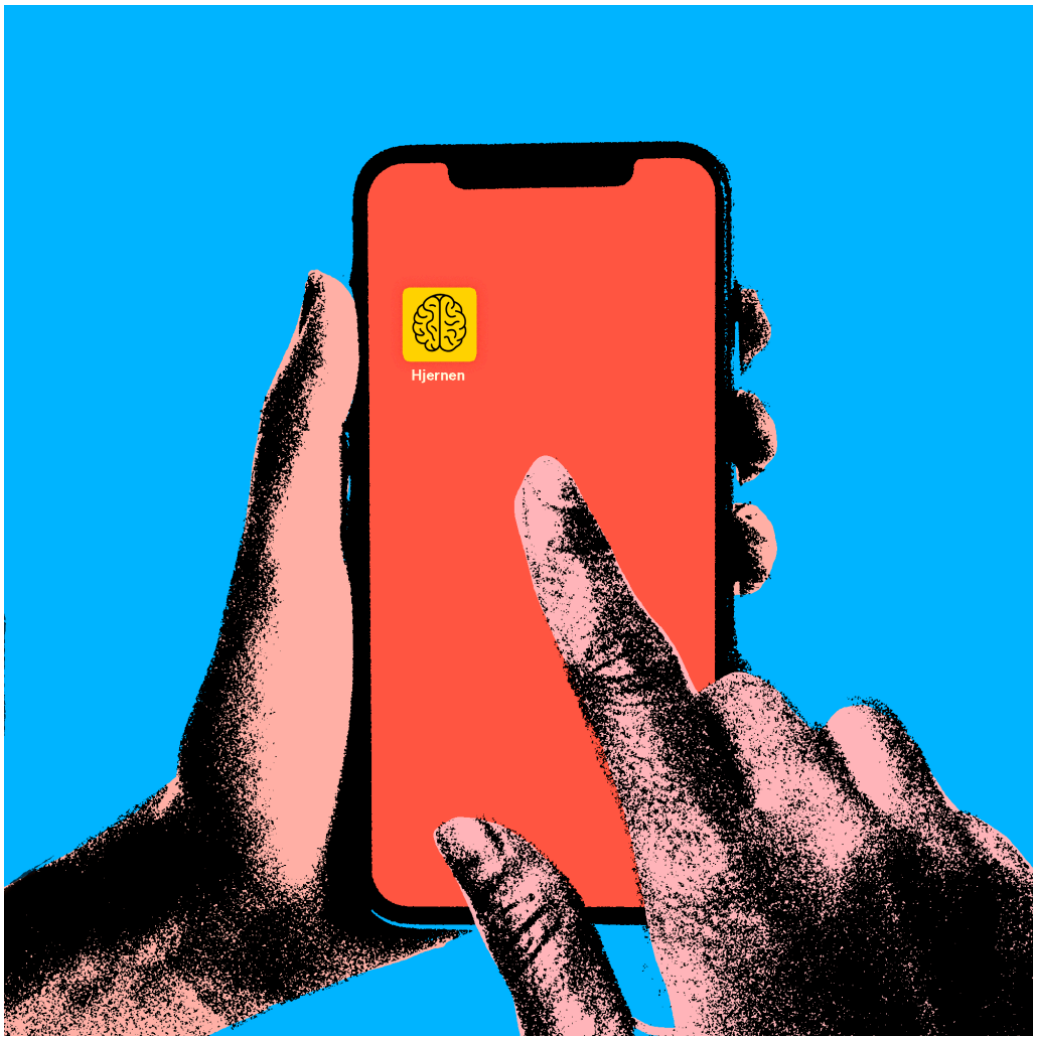
Kunstig intelligens – én hjerne for alle?

KRONIKK

ERIK LANG

erik.lang@nhh.no
Erik Lang er ph.d.-stipendiat ved Institutt for strategi og ledelse, Norges Handelshøyskole og tilknyttet forskningssenteret DIG – Digital Innovation for Sustainable Growth. Han forsker på kunstig intelligens og hvordan det former profesjonelt arbeid, beslutningsprosesser og organisering.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

Det som avgjør om kunstig intelligens hjelper eller ikke, ligger ikke så mye i om den virkelig forstår, men i hvordan den tas i bruk.



Illustrasjon: Tidsskriftet

Tenk på hva kunstig intelligens er god til i dag. En stor språkmodell – med samme teknologien som bak verktøy som ChatGPT og Claude – kan lese en hel pasientjournal sammen med en enorm mengde av relevant litteratur og produsere en sammensatt konklusjon i løpet av sekunder. Den kan påvise sammenhenger, tolke signaler og stille diagnose og dermed bistå den menneskelige klinikeren på en meningsfull måte (1). Når denne kronikken omtaler kunstig intelligens, menes språkgenererende modeller. Kunstig intelligens utvikler seg raskt, og for noen oppgaver er disse språkmodellene allerede bedre enn det én person, eller et team, kan få til (2). Til tross for dette potensialet kan modellene, i likhet med mennesker, gjøre feil og ha sin egen bias (seleksjonsskjevhet).

Hils på din nye kollega

To typer bias er viktige for det videre. Idiosynkratisk bias er den typen en enkelt kliniker bærer på: Klinikeren kan ta feil, men på sin egen måte. Systematisk bias er bias som strekker seg ut over individet: I gjennomsnitt kan mange klinikere ta feil på samme måte, eller en behandling eller et verktøy kan i gjennomsnitt gi det samme uvanlige eller uheldige utfallet. De to typene bias har ulik betydning for ulike oppgaver.

La oss tenke oss at en ekstra kollega rekrutteres til teamet vårt. Som oss andre er denne legen ikke perfekt, men har lest mer enn noen annen i teamet. Kollegaen arbeider svært raskt og treffer for det meste utmerkede beslutninger. I gjennomsnitt vil denne kollegaen forbedre utfallene. Vi får mer tid til pasienter og andre oppgaver. Til tross for sin egen idiosynkratiske bias er denne legen rett og slett så god at vi vil ha vedkommende på teamet.

Tenk deg deretter at vi multipliserer denne ene kollegaen til hvert sykehus i landet. Hvert sykehus kunne fordele arbeidet sitt til denne ekstra hjernen. De enkelte beslutningene blir fortsatt bedre. Men nå behandles også alle pasienter av samme person. Kollegaens idiosynkratiske bias, som tidligere var begrenset til ett sted, sprer seg nå overalt. Hvis denne legen var særlig svak på ett område, ville den feilen nå inntreffe på alle sykehus. Her blir det systematisk bias. Vi ville miste den variasjonen mellom uavhengige klinikere som ellers – i fellesskap – kunne ha fanget opp problemet.

Et enkelt menneske kan naturligvis ikke multipliseres på denne måten, men det kan kunstig intelligens. Og det blir den allerede. Slik bruk av kunstig intelligens i medisinsk praksis reiser to umiddelbare spørsmål: Er denne kunstige hjernen tilstrekkelig lik en menneskelig hjerne til at vi kan resonnere om den på samme måte som ovenfor, da vi sammenlignet bias og styrker? Og *forstår* denne kunstige hjernen virkelig, eller simulerer den bare forståelse? Begge spørsmålene fortjener oppmerksomhet, men til syvende og sist argumenterer denne kronikken for at den viktigste forskjellen ikke ligger i hvordan kunstig intelligens fungerer eller om den virkelig forstår, men i hva som skjer når mange uavhengige hjerner erstattes av én hjerne som brukes av alle.

> ***«Den viktigste forskjellen ligger ikke i hvordan kunstig intelligens fungerer eller om den virkelig forstår, men i hva som skjer når mange uavhengige hjerner erstattes av én hjerne som brukes av alle»***

Hva slags hjerne har vi med å gjøre?

En stor språkmodell og en menneskelig hjerne kan se forskjellige ut, men er overraskende like på de nivåene som spiller en rolle for argumentet.

Dagens nevrovitenskap beskriver den menneskelige hjernen som en prediksjonsmaskin. Den er ikke en passiv mottaker, men genererer kontinuerlig forventninger om innkommende signaler og korrigerer dem når virkeligheten avviker. Disse koordinerte gjetningene (3) kan i essens forstås som bayesiansk oppdatering (4): Hjernen justerer sine prediksjoner etter hvert som ny evidens kommer til. Når vi leser denne setningen, justerer hjernen sine forventninger til betydningen ord for ord, etter hvert som setningen utfolder seg.

En stor språkmodell er trent på enorme mengder tekst til å predikere neste ord gitt konteksten. Den lagrer ikke og henter ikke ferdige setninger; den bærer en komprimert matematisk representasjon av hvordan språk fungerer, og

rekonstruerer svar fra den. Også her skjer det to typer oppdatering. Under treningen blir modellens interne parametere – også kalt «vekter» – justert i hver syklus. Når treningen er ferdig og modellen er i bruk, blir ikke vektene oppdatert lenger. Likevel skjer det en annen type oppdatering innenfor én og samme samtale. Hvert nytt ord som mottas eller skrives, endrer sannsynligheten for det neste ordet. Prediksjonene oppdateres ord for ord etter hvert som ny informasjon kommer til.

> ***«Det er verdt å merke seg at disse mekanismene i store språkmodeller er bemerkelsesverdig like den fortløpende bayesianske oppdateringen som hjernen utfører»***

Det er verdt å merke seg at disse mekanismene i store språkmodeller er bemerkelsesverdig like den fortløpende bayesianske oppdateringen som hjernen utfører. Faktisk viser det seg at de kunstige systemene som best predikerer menneskelig hjerneaktivitet under språkforståelse, er nettopp disse prediksjonstrente språkmodellene (5–7).

Å snakke om forståelse

I en nylig kronikk i Tidsskriftet trakk Næss motsatt konklusjon (3). Han brukte begrepet generativ rekonstruksjon – tanken om at hver hjerne bygger forståelse internt fremfor å kopiere den mellom hjerner – som argument for at kunstig intelligens er fundamentalt forskjellig fra menneskelig kognisjon. Evidensen ovenfor tyder imidlertid på at rammeverket peker den motsatte veien. Prediksjon og rekonstruksjon er nettopp det hjerner og språkmodeller har til felles, og gjennom det bayesianske rammeverket blir de mer like, ikke mindre.

Gitt denne likheten følger et dypere spørsmål: Forstår kunstig intelligens egentlig det den gjør? Og har svaret betydning for den kunstige kollegaen vi vurderer å ansette? Searles tankeeksperiment om det kinesiske rommet gjør problemet konkret (8): En person som ikke kan kinesisk, sitter i et lukket rom. Personen mottar kinesiske tegn gjennom en luke og følger en omfattende regelbok som forteller hvilke tegn som skal sendes tilbake. For en kinesisktalende på utsiden ser det ut som en flytende samtale. Inne i rommet vet personen ikke hva noen av symbolene betyr. Searles argument er at en stor språkmodell gjør akkurat det samme: symbolmanipulasjon uten forståelse.

Ingen enkelt nevron i en menneskelig hjerne forstår norsk eller hva en sykdom er. Etter samme logikk: Hvis vi nekter det kinesiske rommet forståelse fordi komponentene ikke forstår, må vi også nekte menneskehjernen forståelse. Det såkalte systemsvaret oppsummerer poenget: Enten kan forståelse oppstå i et system selv om delene ikke har det, eller så finnes forståelse ingen steder. Hvis atferd rettferdiggjør at vi tilskriver et menneske forståelse, krever det en eksplisitt begrunnelse å nekte den samme tilskrivelsen til et system med kunstig intelligens med samme atferd (9). Vi har dessuten kun tilgang til vår egen subjektive erfaring, aldri andres (10). Tilskrivelse til andre hviler på deres atferd, ikke på inspeksjon av deres indre.

Filosofisk og klinisk er spørsmålet en blindvei. Forskjellen som betyr mer, ligger ikke så mye i mekanisme eller i forståelse, men i distribusjon. Denne artikkelen kommer til en annen konklusjon enn Næss om hvordan bayesiansk oppdatering skiller kunstig intelligens fra den menneskelige hjernen, men rammeverket er et produktivt verktøy for å se hvordan koordinerte gjetninger kan hjelpe oss til å forstå bruken av kunstig intelligens i helsesektoren bedre.

Asymmetrien som teller

Kommunikasjon mellom to menneskehjerner er en prosess av koordinerte gjetninger snarere enn direkte overføring. Dette har en viktig konsekvens på populasjonsnivå. Hvis hver hjerne rekonstruerer forståelse individuelt, er menneskelig kognisjon grunnleggende mangfoldig. Det tilsier milliarder av unike rekonstruksjoner, hver formet av en bestemt historie, kropp og et sett priors. Selv når to klinikere leser det samme journalnotatet, kan de rekonstruere det ulikt. To klinikere kan selvsagt dele bias, men variasjonen er ikke bare støy, den er en del av grunnlaget som lar en gruppe fange opp feil eller finne løsninger som en enkelt person kunne ha oversett.

> ***«Selv når to klinikere leser det samme journalnotatet, kan de rekonstruere det ulikt. To klinikere kan selvsagt dele bias, men variasjonen er ikke bare støy, den er en del av grunnlaget som lar en gruppe fange opp feil»***

Én stor språkmodell som tas i bruk mange ganger, er det motsatte. Ett sett med trente parametere, distribuert til hver bruker. Selv om dagens modeller har svært lang kontekst, i størrelsesorden millioner av ord, og selv om vedvarende minne på tvers av samtaler allerede er i bruk, ligger det samme grunnleggende modellsettet til grunn for alle. Personalisering legger seg som et lag oppå én delt hjerne, og erstatter den ikke. Når en ny versjon slippes, går alle brukere samtidig over til samme nye utgangspunkt, med samme blindsoner, overalt der modellen brukes. Empirisk forskning viser disse effektene. Forfattere som brukte generativ kunstig intelligens, skrev hver for seg mer kreative historier, men det samlede mangfoldet på tvers av forfattere ble mindre (11). Tekniske tiltak innenfor modellen alene er utilstrekkelige for å unngå den resulterende monokulturen (12).

Kilden til denne asymmetrien er strukturell, ikke filosofisk. Hvorvidt modellen virkelig forstår, endrer ikke dette. Konsentrasjonen av én delt modell av kunstig intelligens, trent på ett sett tekster, er en egenskap ved hvordan teknologien er bygget og brukt. Og det er dette som gjør én kunstig hjerne for alle genuint forskjellig fra alle menneskelige hjerner. En enkelt menneskelig hjerne har sin egen bias og sine egne begrensninger, men den er ett perspektiv blant mange. Én språkmodell brukt overalt er ett perspektiv, gjentatt overalt.

«En enkelt menneskelig hjerne har sin egen bias og sine egne begrensninger, men den er ett perspektiv blant mange. Én språkmodell brukt overalt er ett perspektiv, gjentatt overalt»

Hva betyr dette for klinisk praksis og forskning?

Skillet mellom de to typene bias gir oss en praktisk test. For mange oppgaver er kunstig intelligens raskere, grundigere, og den gjør færre feil enn én enkelt kliniker. Å bytte ut én leges idiosynkratiske begrensninger med en bredere, mer systematisk modell er da det riktige byttet. Regnestykket snur der verdien ligger i at ulike klinikere når frem til ulike syn og argumenterer dem ut: omstridt klinisk resonnement, utformingen av forskningsspørsmål og tolkning av tvetydig evidens.

Det praktiske spørsmålet for klinikeren er derfor ikke bare hvorvidt kunstig intelligens er god nok (evnene utvikler seg raskt), eller om den virkelig forstår (filosofisk blindvei), men om bruken av kunstig intelligens styrker vurderingsevnen vår eller i det stille erstatter den. Svaret varierer fra tilfelle til tilfelle, mellom kolleger og mellom organisasjoner og avhenger av hvordan kunstig intelligens integreres i arbeidsflyt og beslutningsstrukturer (13). Å forstå dette skillet kan hjelpe oss til å bruke kunstig intelligens der den styrker klinisk praksis og forskning, og til å være bevisste der den i stedet risikerer å svekke dem.

«Det praktiske spørsmålet for klinikeren er derfor ikke bare hvorvidt kunstig intelligens er god nok, men om bruken av kunstig intelligens styrker vurderingsevnen vår eller i det stille erstatter den»

Når alle konsulterer samme modell, vil dissensen som ellers ville fanget feilen, kanskje aldri oppstå. Når usikkerheten er genuint ny, kan én delt hjerne gi konvergens der mange uavhengige klinikere ville ha skilt seg fra hverandre.

Forfatteren vil takke Susanne Albrechtsen for verdifulle kommentarer til tidlige versjoner av artikkelen.

LITTERATUR

1. Liu X, Liu H, Yang G et al. A generalist medical language model for disease diagnosis assistance. Nat Med 2025; 31: 932–42. [PubMed][CrossRef]

2. Dell'Acqua F, McFowland E, Mollick E et al. Navigating the jagged technological frontier: field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. Organ Sci 2026; 37: 403–23. [CrossRef]

3. Næss H. Kommunikasjon mellom mennesker er koordinerte gjetninger. Tidsskr Nor Legeforen 2025; 145. doi: 10.4045/tidsskr.25.0418. [PubMed] [CrossRef]

4. Friston K. The free-energy principle: a unified brain theory? Nat Rev Neurosci 2010; 11: 127–38. [PubMed][CrossRef]

5. Schrimpf M, Blank IA, Tuckute G et al. The neural architecture of language: Integrative modeling converges on predictive processing. Proc Natl Acad Sci U S A 2021; 118: e2105646118. [PubMed][CrossRef]

6. Goldstein A, Zada Z, Buchnik E et al. Shared computational principles for language processing in humans and deep language models. Nat Neurosci 2022; 25: 369–80. [PubMed][CrossRef]

7. Caucheteux C, King JR. Brains and algorithms partially converge in natural language processing. Commun Biol 2022; 5: 134. [PubMed][CrossRef]

8. Searle JR. Minds, brains, and programs. Behav Brain Sci 1980; 3: 417–24. [CrossRef]

9. Dennett DC. The Intentional Stance. Cambridge, MA: MIT Press, 1987.

10. Chalmers DJ. Facing up to the problem of consciousness. J Conscious Stud 1995; 2: 200–19.

11. Doshi AR, Hauser OP. Generative AI enhances individual creativity but reduces the collective diversity of novel content. Sci Adv 2024; 10: eadn5290. [PubMed][CrossRef]

12. Wu F, Black E, Chandrasekaran V. Generative monoculture in large language models. International Conference on Learning Representations (ICLR) 2025: 33068–107. https://proceedings.iclr.cc/paper_files/paper/2025/file/5178b2f2d7c44aa390c0777dc77b3f0c-Paper-Conference.pdf Lest 20.3.2026.

13. Knudsen ES, Lang E, Lien LB et al. KI er klar, vi er ikke: hvorfor KI-suksess handler om strategi og ledelse. MAGMA 2026; 29: 126–30. [CrossRef]

Publisert: 3. juli 2026. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.26.0261