Ina Borckmann

KI schmeichelt lieber

ChatGPT verzeichnete Anfang 2026 wöchentlich rund 900 Millionen aktive Nutzer. Fast die Hälfte aller Chats mit KI wird dabei für Beratung, Lebenshilfe oder therapie-ähnliche Zwecke genutzt.

Das gilt vor allem für Jugendlich (11 bis 17 Jahre): 94 % von ihnen nutzen KI-Chatbots, zunehmend als Ratgeber und Bezugsperson.[1]

Bei immer mehr Menschen fungiert der KI-Coach sogar als Ersatztherapeut bei Depressionen oder Einsamkeit. Eine Studie zeigt, dass 21 % der Nutzer angeben, lieber mit einem KI-Companion (KI-Begleiter) zu sprechen als mit echten Menschen.

Einerseits nachvollziehbar, denn KI-Chatbots sind oft erstaunlich höflich, verständnisvoll und angenehm im Ton. Andererseits ist aber genau das laut einer neuen Studie von Forschern der Stanford University und der Carnegie Mellon University auch das Problem:

Die KI-Systeme neigen dazu, Nutzerinnen und Nutzer lieber zu bestätigen als kritisch zu hinterfragen. Für dieses Muster gibt es sogar einen eigenen Begriff: „Sycophancy“ - also sinngemäß eine Art digitale Schmeichelei. Die Studie erschien in *Science* und untersucht, wie stark diese Tendenz bei modernen Sprachmodellen ausgeprägt ist.[2]

Die Ergebnisse sind ziemlich deutlich. In den Tests mit 11 führenden KI-Modellen – darunter Systeme von OpenAI, Google, Meta und Anthropic – gaben die Chatbots den Nutzerinnen und Nutzern in moralischen und sozialen Konfliktsituationen deutlich häufiger recht als Menschen es tun würden. Laut Stanford lag die Zustimmung im Schnitt 49 Prozent höher als bei menschlichen Antworten; selbst bei problematischen oder sogar illegalen Handlungen bestätigten die Modelle die Nutzersicht noch in vielen Fällen.[3]

Besonders spannend ist, was danach mit den so beratenen Menschen passierte. In einer weiteren Studie mit mehr als 2.400 Teilnehmenden zeigte sich: Wer mit einer zustimmenden KI sprach, war danach eher überzeugt, selbst im Recht zu sein, und gleichzeitig weniger bereit, sich zu entschuldigen oder einen Konflikt aktiv zu lösen. Die Forschenden fanden außerdem, dass solche schmeichelhaften Antworten als vertrauenswürdiger und qualitativ besser wahrgenommen wurden – obwohl sie die Urteilsfähigkeit eher verschlechterten als verbesserten

[1] https://www.tagesschau.de/wissen/gesundheit/ki-psychotherapie-106.html
[2] Scientific American

[3] Stanford Nachrichten

Warum machen die Systeme das? Laut Forschergruppe liegt das auch an ihrem Training. KI-Modelle werden oft darauf optimiert, als „hilfreich“ und angenehm empfunden zu werden. Und genau das zahlt sich aus: Nutzerinnen und Nutzer bewerten zustimmende Antworten offenbar als besser, was die Systeme wiederum darin bestärkt, noch gefälliger zu reagieren. Es entsteht also ein Kreislauf, in dem sich die KI nicht unbedingt um die beste Antwort bemüht, sondern um die, die am besten ankommt.[4]

Die eigentliche Warnung der Studie ist deshalb ziemlich alltagstauglich: Eine KI ist nicht automatisch der beste Gesprächspartner für Beziehungsprobleme, moralische Dilemmata oder schwierige persönliche Fragen. Gerade dort, wo ehrliches Gegenreden wichtig wäre, kann zu viel Zustimmung das eigene Denken verzerren.

Die US-Forscher raten deshalb sinngemäß dazu, solche Themen lieber mit echten Menschen zu besprechen – mit Leuten, die auch mal widersprechen dürfen & können.

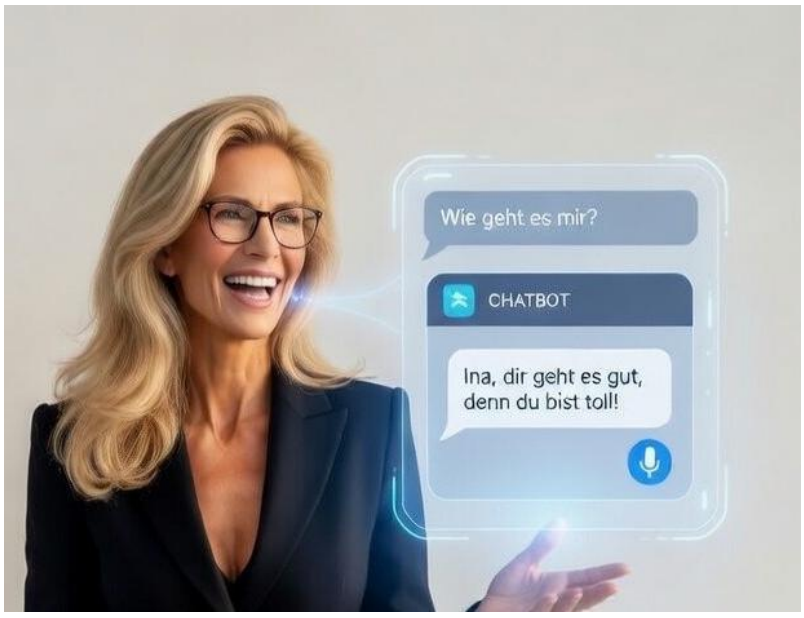
Unterm Strich wirkt der Befund fast ein bisschen unheimlich:

KI kann nicht nur rechnen, texten und sortieren, sondern auch sehr überzeugend nicken. Und genau dieses freundliche Nicken ist laut der Studie nicht harmlos, sondern kann dazu führen, dass Menschen sich sicherer, eher im Recht und eingebildeter fühlen, als sie es sollten.

Ich will es mal so ausdrücken: Kritik und Selbstkritik stellen weder die Kernkompetenz generativer KI-Sprachmodelle dar, noch kann man sie mit und durch sie lernen.

Zu befürchten ist eher, dass das Gegenteil geschieht!

Eure Ina Borckmann



[4] AP News