idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Pressemitteilung

Technische Universität Berlin
Stefanie Terp

23.04.2026
http://idw-online.de/de/news869715

Forschungsergebnisse
Informationstechnik, Psychologie
überregional

Erstmals im EEG sichtbar: Wie Vertrauen in KI entsteht

Forscher*innen der TU Berlin machen Vertrauensdynamiken in Mensch-KI-Teams messbar. Die Ergebnisse zeigen, wann Menschen kognitive Prozesse auslagern und wann sie die Kontrolle behalten

Vertrauen in Künstliche Intelligenz entsteht in Sekundenbruchteilen – und entscheidet darüber, ob Menschen sich auf ein System verlassen oder dessen Ergebnisse misstrauisch überprüfen. Forscher*innen der Technischen Universität Berlin zeigen nun erstmals, dass sich diese Dynamik direkt im Gehirn messen lässt: mithilfe von EEG-Signalen. Im DFG-geförderten Forschungsprojekt „Neuronale Korrelate von Vertrauen in Mensch-KI-Interaktion“ untersuchen Prof. Dr. Eva Wiese und Dr. Tobias Feldmann-Wüstefeld vom Fachgebiet Kognitionspsychologie und Kognitive Ergonomie, wie Vertrauen in KI-Systeme mit sogenanntem kognitivem Offloading zusammenhängt. Gemeint ist damit die Frage, ob und wann Menschen geistige Funktionen wie Aufmerksamkeit oder Gedächtnis an ein technisches System auslagern. Das Besondere: Statt Vertrauen vor allem über Fragebögen oder indirekte Verhaltensdaten zu erfassen, setzt das Team auf neuronale Signale, die objektiv und mit hoher zeitlicher Auflösung gemessen werden können.

Warum Vertrauen schwer zu erfassen ist
KI kann Menschen bei anspruchsvollen Aufgaben spürbar entlasten – etwa indem sie Informationen filtert, Optionen vorschlägt oder Teilaufgaben übernimmt. Ob diese Unterstützung tatsächlich nützt, hängt jedoch entscheidend davon ab, wie stark Menschen dem System vertrauen. Zu viel Vertrauen kann dazu führen, dass Fehler übersehen werden. Zu wenig Vertrauen wiederum kostet Zeit und kann den Nutzen der Unterstützung erheblich mindern. Bislang wird Vertrauen meist über Selbstauskünfte oder Verhaltensmaße wie Reaktionszeiten untersucht. Beide Ansätze haben Grenzen: Fragebögen sind subjektiv und unterbrechen den Arbeitsprozess, Verhaltensdaten sind oft nicht eindeutig interpretierbar.

Offloading: Vertrauen zeigt sich im Handeln
Im Zentrum des Forschungsprojekts steht deshalb die Frage, ob Vertrauen daran erkennbar wird, wie stark Menschen kognitive Prozesse auslagern. Wer einer KI vertraut, behält weniger Informationen selbst aktiv im Blick oder im Gedächtnis, sondern verlässt sich stärker auf das System. Dieses sogenannte Offloading ist im Alltag längst verbreitet – vom Navigationssystem im Auto bis zur automatisierten Bildauswertung in der Medizin. Im Labor lässt sich dieser Mechanismus in kontrollierten visuellen Aufgaben gezielt untersuchen.

EEG macht Veränderungen in Aufmerksamkeit und Gedächtnis sichtbar
Gemessen wird mit dem Elektroenzephalogramm (EEG), das Gehirnaktivität mit Millisekundenauflösung erfasst. Im Fokus stehen zwei etablierte Marker: die N2pc als Maß für visuelle selektive Aufmerksamkeit und die CDA (Contralateral Delay Activity) als Maß für visuelles Kurzzeitgedächtnis. Die Forscher*innen gehen davon aus: Wenn Vertrauen Offloading begünstigt, dann muss sich das in diesen Signalen zeigen. Genau dadurch lässt sich erfassen, ob Menschen kognitive Ressourcen bei sich behalten oder stärker an die KI delegieren – und zwar kontinuierlich, ohne die Aufgabe zu unterbrechen.

Experimente mit KI-Partnern

In den Studien arbeiten Teilnehmer*innen mit einem einfachen probabilistischen Algorithmus, der gezielt als KI-Partner präsentiert wird. Dieses KI-Framing ermöglicht es, die Zusammenarbeit unter kontrollierten Bedingungen zu untersuchen.
Zum Einsatz kommen visuelle Suchaufgaben sowie Change-Detection-Aufgaben. Dabei erfassen die Forscher*innen parallel per EEG, ob Aufmerksamkeit und Gedächtnisarbeit eher bei den Menschen verbleiben oder zunehmend auf die KI verlagert werden.

Von der Grundlagenforschung zur adaptiven KI
Das Projekt ist in vier aufeinander aufbauende Arbeitspakete gegliedert. Zunächst wird geprüft, ob N2pc und CDA kognitives Offloading gegenüber einer KI zuverlässig abbilden. Anschließend werden diese Maße mit etablierten Vertrauensmodellen verknüpft. Untersucht wird dabei unter anderem, welche Rolle die wahrgenommene Leistungsfähigkeit der KI, das empfundene Risiko oder die Transparenz des Systems spielen.
Ein weiterer Schwerpunkt liegt auf Vertrauensdynamiken: Wie reagieren Menschen auf plötzliche Fehler eines Systems? Und durch welche Strategien lässt sich verlorenes Vertrauen wiederherstellen? Im letzten Schritt untersucht das Team adaptive KI-Systeme, die sich an das Verhalten der Nutzer*innen anpassen und in Phasen hoher kognitiver Beanspruchung gezielt unterstützen.

Erste Studien liefern zentrale Hinweise
Zwei aktuelle Publikationen aus dem Fachgebiet bilden die Grundlage des Projekts. Die Studie „Measuring trust in artificial intelligence with the N2pc component“, erschienen im Journal NeuroImage, zeigt, dass die N2pc-Amplitude systematisch mit der Zuverlässigkeit einer KI variiert. Arbeiteten Menschen mit einer verlässlichen KI, fiel die N2pc geringer aus – ein Hinweis auf stärkeres Offloading und eine Entlastung eigener Aufmerksamkeitsressourcen. Bei einer fehleranfälligen KI war die N2pc dagegen größer, was auf eine stärkere eigene Kontrolle hindeutet.
Eine zweite Studie, veröffentlicht im Journal Computers in Human Behavior: Artificial Humans unter dem Titel „Implicit neural measures of trust in artificial intelligence“, zeigt, dass auch die CDA Vertrauensverläufe abbilden kann: beim Aufbau von Vertrauen, bei Vertrauensbrüchen und bei dessen Wiederherstellung.

Relevanz für sicherheitskritische Anwendungen
Die Ergebnisse des Forschungsprojekts sind dort besonders relevant, wo Menschen mit KI in sensiblen oder sicherheitskritischen Kontexten zusammenarbeiten – etwa in der medizinischen Diagnostik, im Verkehr oder in der industriellen Überwachung. In solchen Bereichen kann fehlkalibriertes Vertrauen gravierende Folgen haben: Übervertrauen erhöht das Risiko, Fehler ungeprüft zu übernehmen; Misstrauen wiederum verhindert, dass leistungsfähige Systeme ihr Potenzial entfalten. Das TU-Berlin-Projekt schafft dafür eine neue Grundlage. Es entwickelt objektive neuronale Marker, mit denen sich Vertrauen in KI kontinuierlich und ohne Unterbrechung der Aufgabe erfassen lässt. Langfristig könnte das helfen, KI-Systeme nicht nur besser zu bewerten, sondern auch gezielter zu gestalten. Ziel ist eine belastbare Vertrauenskultur im Umgang mit KI.

Zu den Studien:

„Measuring trust in artificial intelligence with the N2pc component“
https://www.sciencedirect.com/science/article/pii/S1053811926000388?via%3Dihub

„Implicit neural measures of trust in artificial intelligence“
https://www.sciencedirect.com/science/article/pii/S2949882126000253?via%3Dihub

Weitere Informationen erteilen Ihnen gern:
Prof. Dr. Eva Wiese und Dr. Tobias Feldmann-Wüstefeld
Fachgebiet Kognitionspsychologie und Kognitive Ergonomie
Fakultät V – Verkehrs- und Maschinensysteme

Tel.: 030 314-73777
E-Mail: eva.wiese@tu-berlin.de und feldmann-wuestefeld@tu-berlin.de