MedWiss.Online

Mensch-KI-Kollektive stellen die besseren medizinischen Diagnosen

Künstliche Intelligenz (KI) kann Ärztinnen und Ärzte bei der Diagnosefindung wirksam unterstützen. Sie macht andere Fehler als Menschen – und diese Komplementarität stellt eine bislang ungenutzte Stärke dar. Ein internationales Team unter Leitung des Max-Planck-Instituts für Bildungsforschung zeigt nun erstmals systematisch, dass die Kombination aus menschlicher Expertise und KI-Modellen zu den genauesten offenen Diagnosen führt.

Diagnosefehler gehören zu den folgenschwersten Problemen im medizinischen Alltag. KI-Systeme – insbesondere sogenannte große Sprachmodelle (Large Language Models) wie ChatGPT-4, Gemini oder Claude 3 – eröffnen neue Möglichkeiten, medizinische Diagnosen effizient zu unterstützen. Diese Systeme bergen jedoch auch erhebliche Risiken – beispielsweise können sie „halluzinieren" und falsche Informationen generieren. Zudem reproduzieren sie bestehende gesellschaftliche oder medizinische Vorurteile (Bias) und machen Fehler, die für den Menschen oft nicht nachvollziehbar sind.

Ein internationales Forschungsteam unter der Leitung des Max-Planck-Instituts für Bildungsforschung hat gemeinsam mit Partnern vom Human Diagnosis Project (San Francisco) und dem Institute for Cognitive Sciences and Technologies des italienischen Nationalen Forschungsrats (CNR-ISTC, Rom) untersucht, wie Mensch und KI optimal zusammenarbeiten können. Das Ergebnis: Hybride Diagnose-Kollektive – Gruppen aus menschlichen Fachkräften und KI-Systemen – sind deutlich genauer als nur menschliche Kollektive oder KI-Kollektive. Dies gilt insbesondere auch dann, wenn es nicht um einfache Ja-Nein-Entscheidungen geht, sondern um komplexe diagnostische Fragestellungen mit einer Vielzahl möglicher Lösungen. „Unsere Ergebnisse zeigen, dass die Zusammenarbeit zwischen Menschen und KI-Modellen ein großes Potenzial zur Verbesserung der Patientensicherheit hat", sagt Erstautor Nikolas Zöller. Er ist Postdoktorand am Forschungsbereich Adaptive Rationalität des Max-Planck-Instituts für Bildungsforschung.

Realitätsnahe Simulationen mit mehr als 2.100 klinischen Fallbeispielen

Die Forschenden griffen auf Daten des Human Diagnosis Project zurück, das klinische Fallvignetten– das sind kurze Beschreibungen realitätsnaher Patientenbeschwerden – und die zugehörigen korrekten Diagnosen bereitstellt. In der Studie wurden mehr als 2.100 dieser Vignetten genutzt und die Diagnosen von medizinischen Fachkräften mit jenen von fünf führenden KI-Modellen verglichen. Im zentralen Experiment wurden verschiedene Diagnosekollektive simuliert: Einzelpersonen, menschliche Kollektive, KI-Modelle, Kollektive von KI-Modellen und gemischte Mensch-KI-Kollektive. Insgesamt analysierten die Forschenden mehr als 40.000 Diagnosen. Jede wurde nach internationalen medizinischen Standards (SNOMED CT) klassifiziert und bewertet.

Mensch und Maschine ergänzen sich – auch in ihren Fehlern

Die Studie zeigt: Wenn mehrere KI-Modelle kombiniert wurden, erhöhte sich die Diagnosequalität. Das KI-Kollektiv lag im Durchschnitt über dem Niveau von 85 Prozent der menschlichen Diagnostikerinnen und Diagnostiker. Es gab jedoch zahlreiche Fälle, in denen Menschen besser

abschnitten. Interessanterweise kannten Menschen oft die richtige Diagnose, wenn die KI versagte.

Die größte Überraschung: Die Kombination beider Welten führte zu einer deutlichen Steigerung der Genauigkeit. Selbst das Hinzufügen eines einzelnen KI-Modells zu einer Gruppe von Diagnostikerinnen und Diagnostikern – oder umgekehrt – verbesserte das Ergebnis erheblich. Die zuverlässigsten Ergebnisse wurden durch kollektive Entscheidungen erzielt, an denen mehrere Menschen und mehrere KIs beteiligt waren.

Die Erklärung ist, dass Mensch und KI systematisch unterschiedliche Fehler machen. Wenn die KI in manchen Fällen versagte, konnte eine menschliche Fachkraft den Fehler ausgleichen – und umgekehrt. Diese sogenannte Fehlerkomplementarität macht hybride Kollektive so leistungsstark.

„Es geht nicht darum, den Menschen durch Maschinen zu ersetzen. Vielmehr sollten wir Künstliche Intelligenz als ergänzendes Werkzeug begreifen, das in der kollektiven Entscheidungsfindung sein volles Potenzial entfaltet“, sagt Co-Autor Stefan Herzog, Senior Research Scientist am Forschungsbereich Adaptive Rationalität des Max-Planck-Instituts für Bildungsforschung.

Die Forschenden betonen jedoch auch die Grenzen ihrer Arbeit. So wurden ausschließlich textbasierte Fallvignetten untersucht, nicht jedoch echte Patientinnen und Patienten in realen klinischen Situationen. Ob sich die Ergebnisse direkt auf die Praxis übertragen lassen, müssen Folgestudien zeigen. Ebenso konzentrierte sich die Studie ausschließlich auf die Diagnose, nicht auf die Behandlung, und eine korrekte Diagnose garantiert nicht unbedingt eine optimale Behandlung.

Zudem bleibt die Frage offen, wie KI-basierte Unterstützungssysteme in der Praxis von medizinischem Personal sowie von Patientinnen und Patienten angenommen werden. Die potenziellen Risiken von Voreingenommenheit (Bias) und Diskriminierung durch KI sowie durch menschliche Fachkräfte, insbesondere in Bezug auf ethnische, soziale oder geschlechtsspezifische Unterschiede, bedürfen weiterer Forschung.

Breite Einsatzmöglichkeiten für hybride Mensch-KI-Kollektive

Die Studie ist Teil des Projekts „Hybrid Human Artificial Collective Intelligence in Open-Ended Decision Making” (HACID), das im Rahmen von Horizon Europe finanziert wird und die Entwicklung zukünftiger klinischer Entscheidungsunterstützungssysteme durch die intelligente Integration von menschlicher und künstlicher Intelligenz fördern soll. Die Forschenden sehen besonderes Potenzial in Regionen mit eingeschränktem Zugang zu medizinischer Versorgung. Hybride Mensch-KI-Kollektive könnten in solchen Gebieten einen entscheidenden Beitrag zu mehr Gerechtigkeit im Gesundheitswesen leisten.
„Der Ansatz lässt sich auch auf andere kritische Bereiche übertragen – wie das Rechtssystem, die Katastrophenhilfe oder die Klimapolitik –, also überall dort, wo komplexe, risikoreiche Entscheidungen getroffen werden müssen. Das HACID-Projekt entwickelt beispielsweise auch Instrumente zur Verbesserung der Entscheidungsfindung im Bereich der Klimaanpassung“, sagt Vito Trianni, Mitautor und Koordinator des HACID-Projekts.

In Kürze:

– Hybride Diagnose-Kollektive aus Menschen und KI erzielen deutlich genauere Diagnosen als medizinische Fachkräfte oder KI-Systeme allein, weil sie systematisch unterschiedliche Fehler machen, die sich gegenseitig aufheben.
– In der Studie wurden mehr als 2.100 realitätsnahe medizinische Fallvignetten mit über 40.000 ärztlichen und maschinellen Diagnosen analysiert und miteinander verglichen.
– Schon das Hinzufügen eines KI-Modells zu einem menschlichen Kollektiv – oder umgekehrt –

verbesserte die Diagnosequalität spürbar; hybride kollektive Entscheidungen, die von mehreren Menschen und Maschinen getroffen wurden, erzielten die besten Ergebnisse.
– Diese Ergebnisse unterstreichen das Potenzial für mehr Patientensicherheit und eine gerechtere Gesundheitsversorgung, insbesondere in unterversorgten Regionen. Allerdings sind weitere Untersuchungen zur praktischen Umsetzung und zu ethischen Aspekten erforderlich.

Originalpublikation:

Zöller, N., Berger, J., Lin, I., Fu, N., Komarneni, J., Barabucci, G., Laskowski, K., Shia, V., Harack, B., Chu, E. A., Trianni, V., Kurvers, R. H. J. M., & Herzog, S. M. (2025). Human-AI collectives most accurately diagnose clinical vignettes. Proceedings of the National Academy of Sciences of the United States of America, 122(24), Article e2426153122. https://doi.org/10.1073/pnas.2426153122