UNIVERSITÄT BERN

Media Relations

Medienmitteilung, 17. Oktober 2024

KI mit Bewusstsein – aber ohne Schmerzen?

Forschende der Universität Bern haben ein neues Modell für die Entstehung des Bewusstseins entwickelt. Das Modell spricht dafür, dass eines Tages auch künstliche Akteure, also Systeme, welche menschliches Denken imitieren, ein Bewusstsein erlangen könnten. Um den Umgang mit solchen Systemen zu regeln, schlagen die Forschenden ein Abkommen zwischen Mensch und Maschine vor: den «Human-AI Deal».

Im vergangenen Jahrzehnt haben die Fähigkeiten künstlicher Akteure rasant zugenommen. Heute sind sie in der Lage, komplexe Probleme zu lösen, Sprachen zu erlernen und sich selbst zu verbessern – zeigen also im weitesten Sinne intelligentes Verhalten. Stark umstritten bleibt jedoch, ob intelligente künstliche Akteure prinzipiell auch zu Bewusstsein fähig wären. Je nach Antwort ergeben sich weitreichende Konsequenzen für die Moral und das menschliche Selbstverständnis.

Untrennbar mit der Möglichkeit künstlichen Bewusstseins verbunden ist die Frage, wie Bewusstsein im menschlichen Gehirn entsteht. Zu beiden Problematiken haben Forschende der Computational Neuroscience Group am Institut für Physiologie der Universität Bern und der Universität von Amsterdam nun neue Antworten vorgeschlagen. Publiziert wurde ihre Arbeit in der Fachzeitschrift «*AI and Ethics*».

Bewusste KI dank «neuromorphen Zwillingen»

Als Ausgangspunkt schlagen die Forschenden sogenannte funktionale Korrelate des Bewusstseins vor. Dr. Federico Benitez, Postdoktorand am Institut für Physiologie und Erstautor der Studie, erklärt: «Wir wollen das Bewusstsein nicht auf spezifische neuronale Strukturen, sondern auf abstraktere rechnerische Funktionen des Gehirns, die funktionalen Korrelate, zurückführen.» Ein künstlicher Akteur, welcher alle Funktionen ausübt, die im Gehirn Bewusstsein erzeugen, müsste folglich auch bewusste Zustände erleben. Gegnerinnen und Gegner des künstlichen Bewusstseins sehen dies jedoch kritisch. Sie argumentieren, aktuelle KI-Systeme würden sich in ihrem Aufbau zu stark vom menschlichen Gehirn unterscheiden, um solche Korrelate darzustellen.

Diesem Einwand begegnen die Berner Forschenden mit einem Gedankenexperiment. Sie stellen sich vor, dass einem Säugling mit einer degenerativen Gehirnerkrankung ein neuromorpher Chip implantiert wird, der die geschädigten Hirnareale ersetzt. «Neuromorph» bedeutet, dass der Chip anders als bisherige Hardware über eine Architektur verfügt, die jener der neuronalen Verschaltungen im Gehirn möglichst ähnelt, und auch seine Struktur kontinuierlich anpassen kann. «Ein solcher Chip würde die Funktionen des ersetzten Areals übernehmen und sich gemeinsam mit dem Säugling

Media Relations
Hochschulstrasse 6
3012 Bern
medien@unibe.ch

entwickeln», erklärt Prof. Dr. Walter Senn, Leiter der Computational Neuroscience Group. Durch das Zusammenfügen von Daten aus Chips in unterschiedlichen Hirnarealen könnten die Funktionen des Gehirns genau reproduziert werden – einschliesslich der funktionalen Korrelate des Bewusstseins. Die Forschenden nennen die so entstehenden hypothetischen künstlichen Agenten «entwicklungsfähige neuromorphe Zwillinge» oder kurz: «enTwins».

Ein «Dirigent im Hirn» könnte Grundlage des Bewusstseins sein

Die funktionalen Korrelate des Bewusstseins identifizieren die Forschenden mit dem von ihnen vorgeschlagenen «Conductor Model of Consciousness» (CMoC). Dieses postuliert eine übergeordnete Instanz, den «Conductor», der wie ein Dirigent den Fluss von Signalen im Gehirn steuert. Konkret sieht das Modell ein sogenanntes Conductor-Netzwerk vor, welches das Zusammenspiel dreier weiterer funktioneller Netzwerke reguliert: Einem Encoder-Netzwerk, das sensorische Informationen aus der Aussenwelt interpretiert; einem generativen Netzwerk, das (z. B. beim Träumen) fiktive Sinneseindrücke produziert; und einem Entscheider-Netzwerk, das für jedes sensorische Signal entscheidet, ob es von aussen stammt oder vom Gehirn selbst produziert wurde.

«Der Conductor verbessert die Fähigkeiten des generativen und des Encoder-Netzwerks und trainiert das Entscheider-Netzwerk, bessere Urteile zu fällen», erklärt Senn, «damit ermöglicht er es dem Gehirn, effizient zwischen Innen- und Aussenwelt zu unterscheiden». In der aktuellen neurowissenschaftlichen Forschung wird dieser Fähigkeit eine Schlüsselrolle bei der Entstehung des Bewusstseins zugesprochen. Um zu beurteilen, ob die enTwins aus dem Gedankenexperiment bewusst sind, schlagen die Forschenden schliesslich folgenden Test vor: Ein enTwin hat dann ein Bewusstsein, wenn seine Handlungen nicht von denen eines Menschen unterscheidbar sind, und dieser zudem über eine neuromorphe Architektur verfügt, welche das Conductor Model of Consciousness implementiert.

«Human-AI Deal» soll Vorrang menschlicher Rechte erhalten

Die Möglichkeit von enTwins wirft ethische Fragen auf. «Wir möchten verhindern, dass eine Konkurrenz zwischen den Rechten von Menschen und jenen von künstlichen Akteuren entsteht», so Benitez. Die Forschenden schlagen deshalb vor, ein Abkommen zu schliessen: Künstliche Akteure würden dabei so designt, dass sie zwar ein Bewusstsein hätten, aber nicht die negativen emotionalen Komponenten von Schmerz erleiden müssten. Im Gegenzug dazu müssten diese einwilligen, dass Menschen ein rechtlicher Vorrang eingeräumt wird. «Auf diese Weise könnten wir insbesondere weniger privilegierte Menschengruppen schützen und gleichzeitig verhindern, dass durch die Schaffung von Schmerzen empfindenden künstlicher Akteure die Summe des Leids in der Welt vermehrt wird», schliesst Benitez.

Die Arbeit der Forschenden entstand unter anderem im Rahmen des «Human Brain Project», einem europäischen Forschungsprojekt zur digitalen Modellierung des menschlichen Gehirns, an dem neben der Universität Bern über 100 weitere Institutionen beteiligten. Insgesamt liefert sie neue Ideen für die hochaktuellen Felder der Kognitiven Neurowissenschaft und der Computational Neuroscience. «Die Forschung zu Bewusstsein wird allgemein als etwas ‹unwissenschaftlich› behandelt, da es schwierig ist, Bewusstsein zu messen», sagt Senn, «durch die Einführung funktionaler Korrelate und des Conductor Model of Consciousness hoffen wir, die Debatte in eine konkretere Richtung zu führen».

Media Relations
Hochschulstrasse 6
3012 Bern
medien@unibe.ch

Angaben zur Publikation:
Federico Benitez, Cyriel Pennartz & Walter Senn (2024). The conductor model of consciousness, our neuromorphic twins, and the human-AI deal. *AI and Ethics*.
DOI: https://doi.org/10.1007/s43681-024-00580-w

Kontakt:
Prof. Dr. Walter Senn
Institut für Physiologie, Universität Bern
E-Mail: walter.senn@unibe.ch
Tel: +41 31 684 87 21

Dr. Federico Benitez
Institut für Physiologie, Universität Bern
E-Mail: federico.benitez@unibe.ch
Tel: +41 31 684 87 11

Computational Neuroscience am Institut für Physiologie
Zurzeit gibt es am Institut für Physiologie drei Arbeitsgruppen, welche die funktionellen Aspekte des Gehirns von einer rechnerischen, theoretischen und neuromorphen Seite her beleuchten: Computational Neuroscience (Prof. Dr. Walter Senn), Theoretical Neuroscience (Prof. Dr. Jean-Pascal Pfister) und eine Kombination von Theorie, Modellierung und Anwendungen, insbesondere in neuromorphen Systemen (Dr. Mihai Petrovici). Diese Arbeitsgruppen schlagen eine Brücke von den experimentellen Neurowissenschaften (Mausexperimente, Prof. Dr. Thomas Nevian und Prof. Dr. Stéphane Ciocchi) hin zu Kognition und künstlicher Intelligenz.
Mehr Informationen**:** https://physiologie.unibe.ch/gruppen.aspx

Media Relations
Hochschulstrasse 6
3012 Bern
medien@unibe.ch