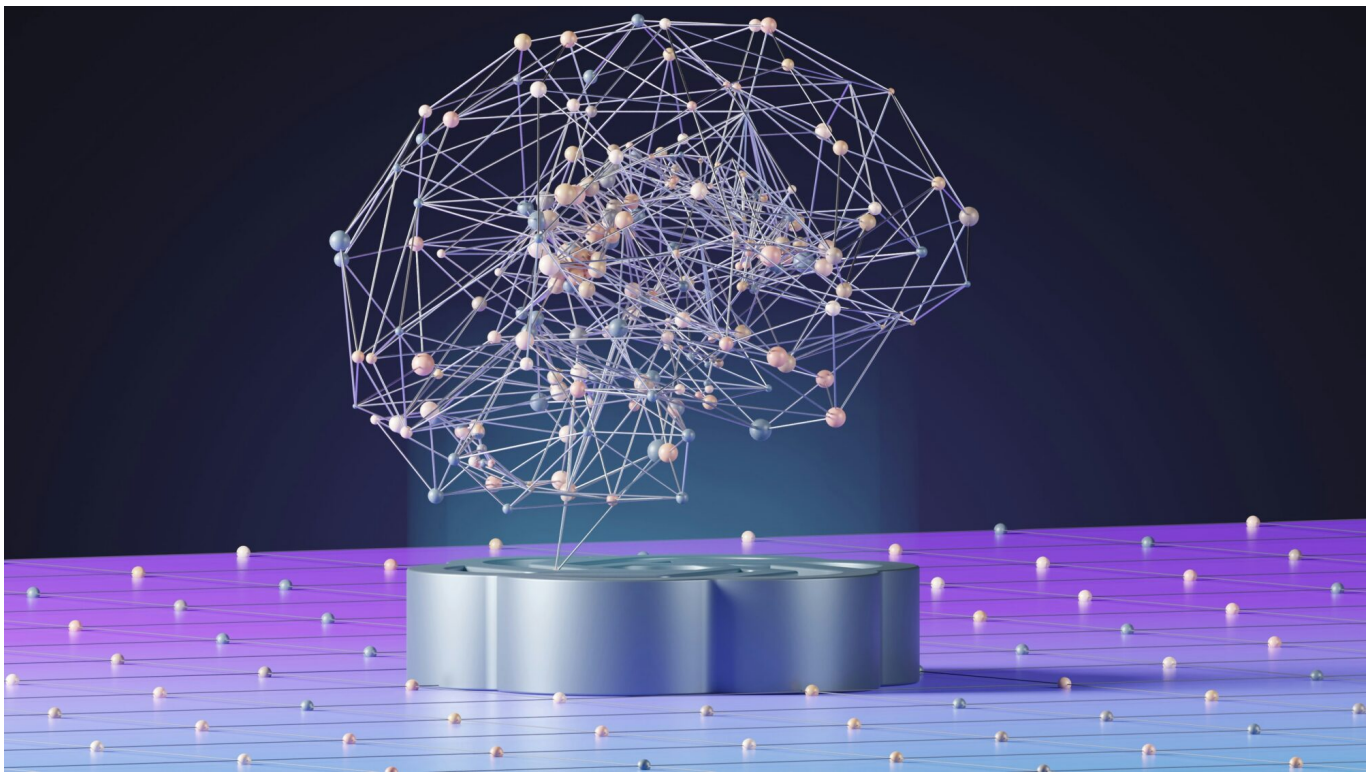
Künstliche Intelligenz wirkt vernünftig. Aber denkt sie wie wir?

Francesco Veri
28th October 2026



Künstliche Intelligenz (KI) kann innerhalb weniger Sekunden überzeugende Argumente zu komplexen politischen Fragen formulieren. Doch ist Überzeugungskraft mit vernünftiger Argumentation gleichzusetzen? Eine in der Fachzeitschrift "Proceedings of the National Academy of Sciences" (PNAS) veröffentlichte Studie des Zentrums für Demokratie Aarau (ZDA) vergleicht 60 Sprachmodelle mit dem Argumentationsverhalten von 190 Bürgerinnen und Bürgern, die an deliberativen Prozessen teilgenommen haben. Die Ergebnisse mahnen zur Vorsicht: Eine plausibel klingende Antwort ist noch kein Garant für die Qualität der Argumente, auf denen sie beruht.

In einer Demokratie gibt es nicht immer die einzig richtige Antwort

Stellen wir uns eine Bürgerversammlung vor, die darüber diskutiert, wie dem Klimawandel begegnet werden soll. Einige Teilnehmende sprechen sich für einschneidende Massnahmen aus, um die Emissionen zu reduzieren, während andere schrittweise Veränderungen bevorzugen, weil sie wirtschaftliche und soziale Folgen befürchten. Obwohl sie ähnliche Sorgen teilen, gelangen sie zu unterschiedlichen Schlussfolgerungen.

Tatsächlich gibt es bei politischen Fragen nicht immer eine einzige richtige Antwort. Bürgerinnen und Bürger mit unterschiedlichen Werten und Prioritäten können verschiedene Positionen vertreten, ohne dass die eine notwendigerweise weniger vernünftig ist als die andere. Im gemeinsamen Austausch können sie

ein geteiltes Verständnis dafür entwickeln, wie sich verschiedene politische Präferenzen entwickeln, auch wenn sie weiterhin unterschiedliche Standpunkte vertreten. Vernunft erhält damit eine kollektive Dimension: Sie zeigt sich in gemeinsamen Mustern des Argumentierens und setzt nicht zwingend Einigkeit über politische Lösungen voraus.

Doch wäre ein KI-System, das sich zu denselben Fragen äussern soll, in der Lage, diese Form kollektiver Vernünftigkeit nachzubilden?

60 KI-Modelle im Vergleich mit menschlichem Argumentieren

Die Studie untersuchte 24 deliberative Prozesse zu komplexen Themen, darunter Klimawandel, Migration und Raumplanung. Mithilfe statistischer Analysen wurden neun Fälle ausgewählt, in denen die Deliberation die Kohärenz des kollektiven Argumentierens signifikant verbesserte. Die Antworten der 190 Bürgerinnen und Bürger, die an diesen Prozessen beteiligt waren, dienten anschliessend als Vergleichsmassstab für die Bewertung von 60 KI-Modellen. Dabei wurde untersucht, inwieweit die Modelle dieselben Argumentationsmuster reproduzieren konnten, die bei den Bürgerinnen und Bürgern beobachtet worden waren.

Um die Ergebnisse auszuwerten, wurde der Deliberative Reason Index (DRI) verwendet. Dieser misst, inwieweit sich gemeinsame Strukturen im Zusammenhang zwischen Begründungen und politischen Präferenzen erkennen lassen. Von den 60 untersuchten Modellen konnten lediglich vier diese Muster in mehr als 60 Prozent der Tests reproduzieren und erzielten dabei Ergebnisse, die über dem zufällig zu erwartenden Niveau lagen. Dieses Resultat zeigt, dass die Fähigkeit der KI, überzeugend klingende Argumente zu liefern, nicht zwangsläufig mit der Fähigkeit einhergeht, die Strukturen menschlichen Argumentierens nachzubilden.

Wenn Plausibilität zum demokratischen Risiko wird

Warum sollte uns dieser Unterschied beunruhigen? Weil KI zunehmend zur Unterstützung kollektiver Entscheidungsprozesse eingesetzt wird. Dabei besteht die Gefahr, dass überzeugend formulierten Antworten Autorität zugeschrieben wird, selbst wenn die zugrunde liegenden Begründungen wenig stichhaltig sind. Daraus ergeben sich drei mögliche Risiken:

Das erste Risiko besteht darin, Überzeugungskraft mit Rechtfertigung zu verwechseln. KI kann überzeugend klingende politische Vorschläge formulieren, ohne dass zwischen den vorgebrachten Argumenten und den daraus gezogenen Schlussfolgerungen ein kohärenter Zusammenhang besteht. Die Gefahr liegt darin, sprachliche Qualität mit der Qualität des Argumentierens gleichzusetzen. Dadurch können unzureichend begründete Entscheidungen den Anschein von Legitimität erhalten und sich einer kritischen Hinterfragung leichter entziehen.

Das zweite Risiko betrifft die Abbildung gesellschaftlicher Meinungsvielfalt. Wenn KI die Positionen von Bürgerinnen und Bürgern neu ordnet, ohne die Zusammenhänge zwischen ihren Begründungen und Präferenzen angemessen zu berücksichtigen, kann sie bestimmte Positionen bevorzugen, andere verdrängen und grundlegende Meinungsverschiedenheiten als vermeintlichen Konsens

darstellen. Dadurch droht eine Verzerrung der öffentlichen Debatte, bei der Minderheitspositionen nicht sichtbar werden und die gesellschaftliche Vielfalt nur noch unzureichend abgebildet wird.

Das dritte Risiko betrifft die Verantwortung für die Rechtfertigung politischer Positionen. In der demokratischen Deliberation bedeutet das Vertreten eines Standpunkts, diesen öffentlich zu begründen und sich mit Kritik auseinanderzusetzen. KI hingegen kann scheinbar wohlüberlegte Schlussfolgerungen präsentieren, ohne die zugrunde liegenden Begründungen transparent zu machen. Damit besteht die Gefahr, dass der kritische Austausch durch autoritativ wirkende Ergebnisse ersetzt wird, die sich einer wirksamen Anfechtung entziehen.

Eine KI, die uns beim Denken unterstützt, statt für uns zu denken

Die Lösung besteht nicht einfach darin, einen Menschen «im Entscheidungsprozess» zu belassen. Denn auch Personen, die ein KI-System kontrollieren, können sich von der Überzeugungskraft seiner Antworten beeinflussen lassen. Notwendig sind deshalb klare Grenzen für den Einsatz von KI bei der Formulierung politischer Urteile, Instrumente zur unabhängigen Überprüfung sowie wirksame Möglichkeiten, Argumente und Empfehlungen kritisch zu hinterfragen.

Ein vielversprechenderer Ansatz besteht darin, KI für Aufgaben einzusetzen, die weniger stark in die menschliche Urteilsbildung eingreifen. So könnte KI dazu dienen, die von Bürgerinnen und Bürgern vorgebrachten Argumente zu strukturieren, miteinander zu vergleichen und sichtbar zu machen, ohne sie durch künstlich erzeugte Präferenzen oder Interpretationen zu ersetzen.

Die Herausforderung besteht nicht darin, Maschinen möglichst vernünftig erscheinen zu lassen. Vielmehr gilt es, die Voraussetzungen zu schützen, in welchen Menschen gemeinsam nachdenken, unterschiedlicher Meinung sein und ihre Entscheidungen öffentlich begründen können.

Methodischer Hintergrund | Wie lässt sich vernünftiges Argumentieren messen?

Der Deliberative Reason Index (DRI) misst, ob Ähnlichkeiten und Unterschiede in den vorgebrachten Überlegungen mit erkennbaren Strukturen politischer Präferenzen übereinstimmen. Er beurteilt weder, welche Meinungen wahr oder richtig sind, noch setzt er einen Konsens voraus. Mithilfe statistischer Tests und Permutationstests lässt sich überprüfen, ob die beobachteten Regelmässigkeiten über das hinausgehen, was zufällig zu erwarten wäre.

Referenz

Veri, Francesco und Gustavo Kreia Umbelino (2026). «Plausible nonsense and deliberative reasoning: Benchmarking LLMs against human judgment». Proceedings of the National Academy of Sciences, 123(37), e2600126123.

https://doi.org/10.1073/pnas.2600126123