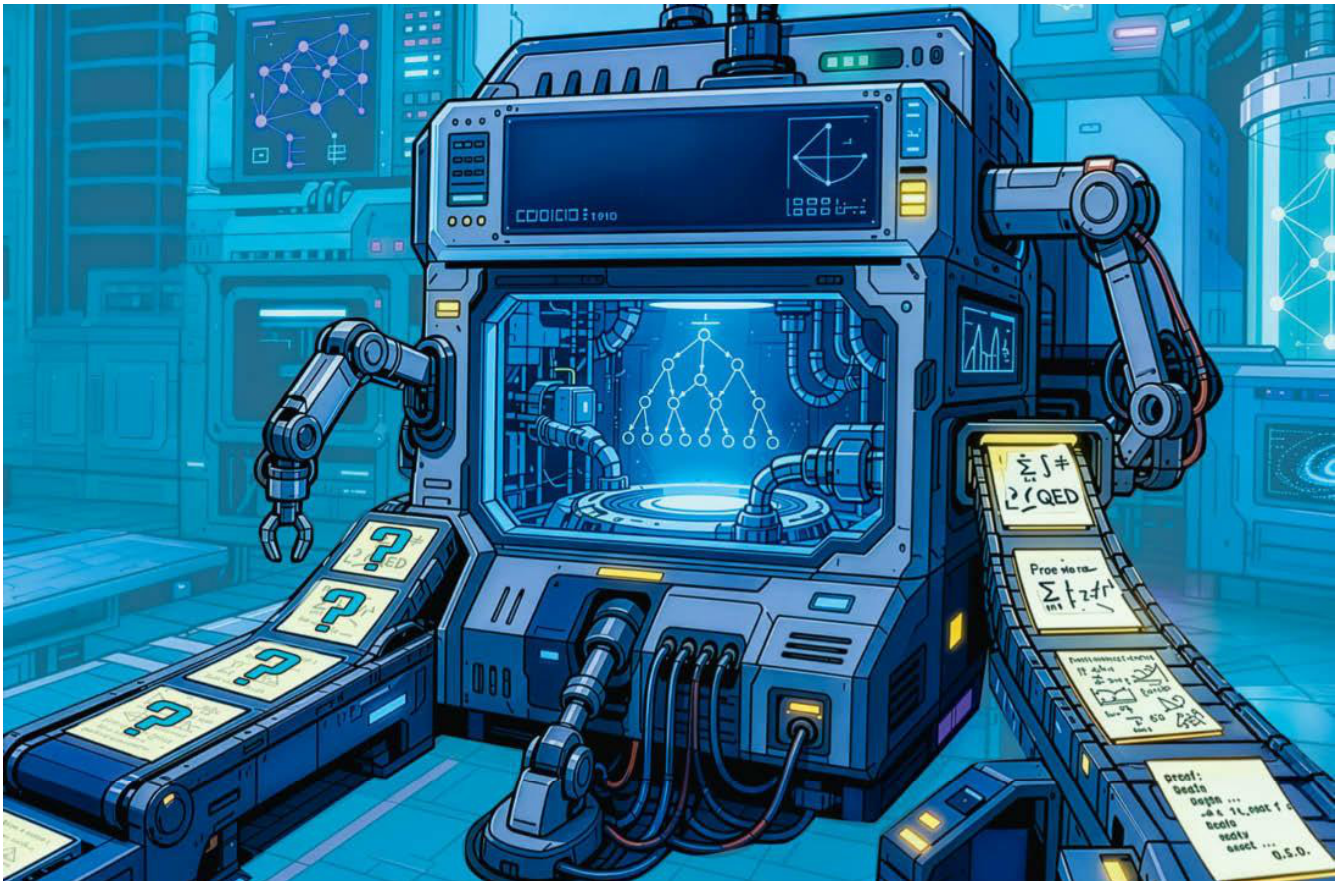


Bild: Jessica Nachtigall / KI / heise medien

Beweisfabriken

Wenn die KI Millennium-Rätsel löst

OpenAI und Anthropic liefern sich ein atemloses Rennen, bevorzugt in menschlichen Königsdisziplinen wie der Mathematik. Doch was ist wirklich KIs Werk und was des Menschen Beitrag?

Von Andrea Trinkwalder

Ein noch unveröffentlichtes Modell, 88 Stunden, 10.000 Agenten, 130 Milliarden Output-Token und 2,7 Millionen ausgetauschte Nachrichten; Kostenpunkt unterm Strich mehrere Millionen Dollar. Das sind die Eckdaten dessen, was OpenAI an Ressourcen auf die Lösung eines mathematischen Problems warf. Und zwar vor allem deshalb, weil einem Gerücht zufolge zwei menschliche Mathematiker im Begriff waren, es zu knacken. Dabei handelte es sich nicht um irgendein beliebiges Rätsel, sondern um eines der Millennium-Probleme des Clay Mathematics Institute: die seit 90 Jahren nicht allgemein bewiesenen Navier-Stokes-Gleichungen, die die Strömungen von Gasen und Flüssigkeiten beschreiben. Zudem handelte es sich bei einem der beiden menschlichen Wissenschaftler, die OpenAI nun auf der Zielgeraden überholte, nicht um irgendwen, sondern um einen Anthropic-Mitarbeiter.

OpenAIs Coup ist der vorläufige Höhepunkt einer Entwicklung, die 2024 begann, in diesem Jahr kulminiert und eine ziemlich eindrucksvolle Demonstration des „Move fast and break things"-Prinzips ist, das die Tech-Szene zu ihrer Handlungsmaxime erhoben hat. Tatsächlich sehen sich Mathematiker nun in ihrer Disziplin mit einigen Brüchen konfrontiert, die in dieser Ausprägung noch nicht stattgefunden haben.

Verkommt die altehrwürdige Mathematik zu einer Benchmark-Spielwiese für große KI-Labore? Wird sich der bisher im Wesentlichen von fachlichem Austausch und vertrauensvoller Zusammenarbeit geprägte Umgang miteinander gravierend ändern, hin zur geheimniskrämerischen Arbeit im Verborgenen? Und überhaupt: Was steckt fachlich hinter den vermeintlichen Durchbrüchen und hat die KI wirklich so autark gearbeitet, wie ihre Schöpfer gerne behaupten? Wir sichten die Beweislage.

Die Mathe-Evolution

Die mathematische Karriere der KI-Modelle begann 2024, als AlphaProof und

AlphaGeometry 2 von Google DeepMind den Silbermedaillen-Standard bei der internationalen Mathe-Olympiade (IMO) erreichten. Es war das erste Mal, dass es Machine-Learning-Systeme aufs Podium schafften. Sie lösten vier der sechs Probleme und kamen damit auf 28 von 42 Punkten. Allerdings brauchten sie mitunter bis zu drei Tage, um eine Lösung zu finden und damit deutlich länger als die menschlichen Teilnehmer. Ein Jahr später stand Google DeepMind mit Gemini Deep Think bereits ganz oben auf dem Treppchen. Diesmal erreichte das Team mit fünf gelösten Aufgaben 35 von 42 Punkten, blieb innerhalb des offiziellen Zeitlimits von 4,5 Stunden und die IMO-Prüfer würdigten die Lösungswege als klar und präzise.

Immer stärker gerieten die sogenannten Erdős-Probleme ins Visier. Das sind über tausend Fragen und Vermutungen, die der ungarische Mathematiker Paul Erdős im Laufe seines Lebens aufgeworfen hatte, beginnend in den 1930er-Jahren. Die meisten davon sind seit Jahrzehnten ungelöst; manche, weil sie tatsächlich knifflig sind, andere gelten aber auch als nicht sonderlich relevant. Unter anderem betreffen sie Primzahlen und deren Verteilung, Graphentheorie sowie die Kombinatorik – also zahlreiche Probleme, die prädestiniert dafür sind, mithilfe von KI gelöst zu werden. Die bescherten OpenAI im Herbst 2025 eine herbe Blamage: Dessen Chief Product Officer Kevin Weil feierte vorschnell einige, vermeintlich von GPT-5 erbrachten Erdős-Beweise, musste dann allerdings zugeben, dass er den Post eines Professors grob fehlinterpretiert hatte: Dieser hatte lediglich mithilfe eines GPT-5-Prompts bereits veröffentlichte, aber noch nicht offiziell gelistete Lösungen für zehn Erdős-Probleme gefunden.

2026 sollte dann das Jahr werden, in dem KIs Erdős-Probleme einsammelten wie Fallobst. Anfang des Jahres gab es einige Achtungserfolge unter Beteiligung von Reasoningmodellen, doch schon im April machte mit dem 23-jährigen Liam Price ein mathematischer Laie von sich reden, weil er mit einem einzigen Prompt an GPT-5.4 Pro das Erdős-Problem #1196 löste (siehe ct.de/y6ag). Und nicht nur das: Das Modell fand einen in der Fachliteratur bislang kaum verfolgten Ansatz, indem es die Brücke zwischen analytischer Zahlentheorie und stochastischer Prozesstheorie schlug.

Mittlerweile sind über 100 dieser Probleme unter Beteiligung von Machine-Learning-Systemen als gelöst gelistet, wie der Mathematikprofessor Terence Tao in einer Übersicht auf GitHub dokumentiert (siehe ct.de/y6ag). Einen geringen Teil davon haben die jeweils eingesetzten Reasoningmodelle nahezu autonom bewältigt; das Gros entfällt auf das Auffinden übersehener Publikationen sowie mehr oder weniger intensive Kooperationen zwischen Mensch und Maschine.

Zwischen Fail und Anerkennung

Einen wichtigen Hinweis von Tao sollte jeder beherzigen, der über die dort gesammelten KI-Erfolge schreibt: Wenn ein Rätsel jahrzehntelang niemand gelöst hat, muss es nicht automatisch besonders schwierig sein. Es könnte auch daran liegen, dass sich bislang kaum jemand dafür interessiert hat, etwa weil es weder für die Entwicklung der Mathematik noch in der Praxis als sonderlich bedeutsam gilt. Vier Lösungswege listet die GitHub-Übersicht als besonders bemerkenswert: Wie im Dezember 2025 Problem 1026 geknackt wurde, demonstriere „diverse Formen von KI-Unterstützung und -Nutzen in der mathematischen Zusammenarbeit“. Nummer 728 gilt als die erste autonome, nicht triviale Lösung eines Erdős-Problems durch ein KI-System. Liam Price wird ebenfalls gewürdigt, weil sich an #1196 zuvor schon einige Experten vergeblich versucht hatten und GPT-5.4 eine ungewöhnliche Herangehensweise fand (mehr zu den Lösungswegen siehe ct.de/y6ag).

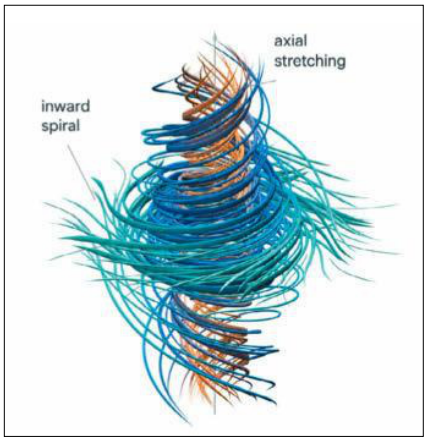


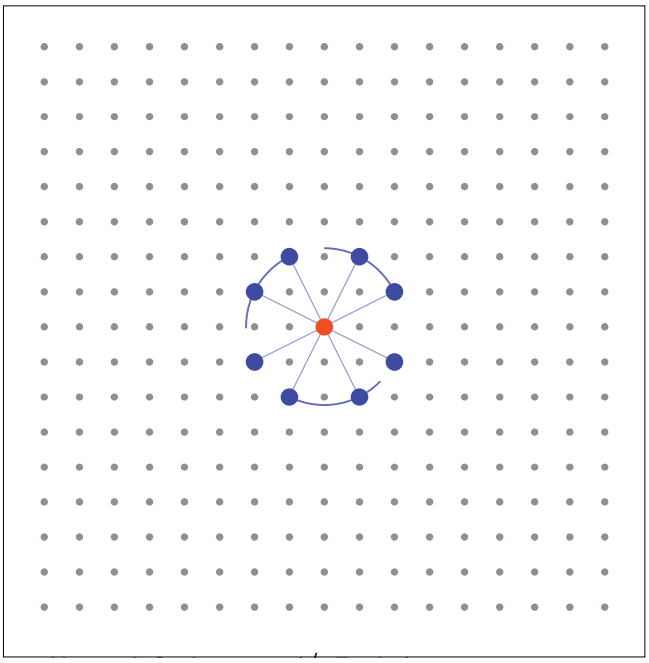
Bild: OpenAI

Ein noch nicht veröffentlichtes KI-Modell von OpenAI hat möglicherweise eine der offenen Millenniumfragen geklärt, indem es ein nicht praxisrelevantes Beispiel dafür konstruierte, dass die Navier-Stokes-Gleichungen unter bestimmten Bedingungen zusammenbrechen.

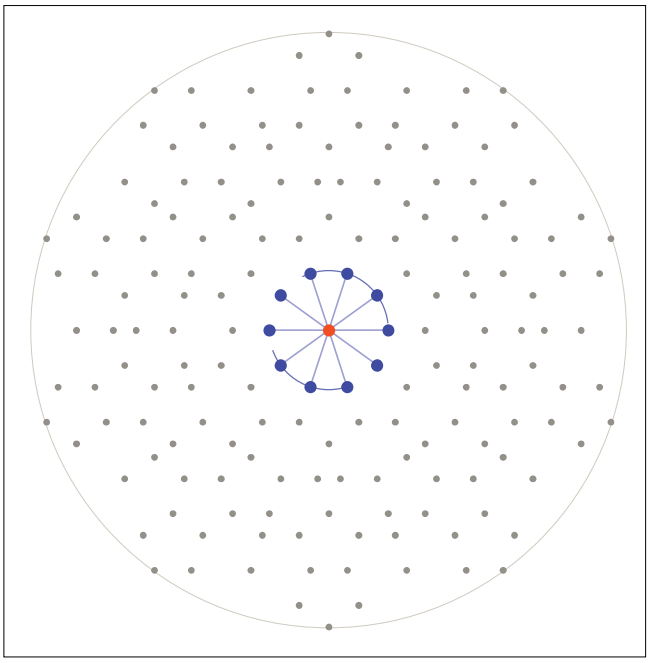
Besondere Anerkennung zollte die Fachwelt allerdings OpenAIs Beitrag zum 1946 formulierten „Planar Unit Distance Problem“ (#90). Denn zum einen handelte es sich um eine mathematisch bedeutende Fragestellung, an der sich bereits einige Wissenschaftler die Zähne ausgebissen hatten. Zum anderen überzeugte das hohe fachliche und formelle Niveau der Arbeit, die im Übrigen nicht von einem dedizierten Mathematiksystem wie AlphaProof, Aristotle oder Deep Think, sondern von einem damals noch internen allgemeinen Reasoningmodell aus OpenAIs Laboren stammte.

Die Vermutung, die Paul Erdős 1946 aufstellte und die ein KI-Modell von OpenAI im Mai 2026 widerlegte, ist auch für Nicht-Mathematiker verständlich: Wenn man n Punkte in einer Ebene verteilt, wie viele Punktpaare können dann höchstens genau den Abstand 1 haben? Erdős vermutete, dass man bei sehr vielen Punkten kaum besser abschneiden kann als mit einem quadratischen Gitter, dessen Maschenweite so gewählt ist, dass ein Springerzug genau die Länge 1 hat. Denn ein Springer kann in acht Richtungen ziehen, ein Schritt zum Nachbarfeld gelingt nur viermal. Das OpenAI-Modell fand jedoch eine Konstruktion, die das Karomuster übertrifft, wenn auch erst bei astronomisch vielen Punkten. Dafür griff es überraschenderweise zu einem Werkzeug aus der algebraischen Zahlentheorie: Es nutzte spezielle Zahlensysteme, in denen es ungewöhnlich viele Richtungen gibt, in die man genau eine Längeneinheit weit von Punkt zu Punkt gelangt.

Außerdem formulierte es die Beweisschritte überraschend stringent und für den Menschen nachvollziehbar: „Das erste Ergebnis, das autonom von einer KI erzeugt wurde und das ich an sich interessant finde“, bescheinigte der an der University of Toronto tätige Mathematiker Daniel Litt. Gleichzeitig sah er die Lösung des Problems auch durchaus im Bereich des menschlich Möglichen. Mehrere Mathematiker, die die Veröffentlichung geprüft und für korrekt befunden haben, merkten in ihrer Stellungnahme außerdem an, dass sich die „Argumentation entscheidend auf Ideen stütze, die – zumindest im Nachhinein – Ellenberg-Venkatesh Golod-Shafarevich und Hajir-Maire-Ramakrishna zugeschrieben werden können.“ Erhellende Einblicke in Konstruktion und Gestalt der neuen Anordnung gibt die Animation von Ed Pegg von Wolfram Research (siehe ct.de/y6ag).



Mit dem Springerzug als Längeneinheit hat jeder Gitterpunkt acht Partner im Abstand 1 statt nur vier direkte in horizontaler und vertikaler Richtung.



Vereinfachtes Modell der KI-Konstruktion: Die Punkte stammen aus einem Zahlensystem mit zehn Richtungen der Länge 1, den Zacken eines Fünfsterns.

In jedem Fall ist Erdős #90 eine sehr gelungene Demonstration dessen, wie und in welchen Bereichen große Sprachmodelle die Mathematik voranbringen können. Sie spielen mit vergleichsweise geringem Aufwand unterschiedliche Lösungsstrategien durch und finden dabei auch Ansätze, die über den jeweiligen Fachbereich hinausgehen. Mitunter finden sich entsprechende Ideen schon in irgendwelchen Veröffentlichungen, aber es gibt schlichtweg zu viele und die Zeit des menschlichen Mathematikers ist begrenzt. So hilft KI, über den Tellerrand zu blicken, indem sie auch Szenarien verfolgt, die auf den ersten Blick weniger zielführend erscheinen. In der Tat hatten die meisten Mathematiker bisher eher versucht, Erdős' These zu beweisen, weil sie von ihrer Richtigkeit überzeugt waren.

Solche „low hanging fruits", wie sie Terence Tao bezeichnet, gibt es vor allem auf den Gebieten der Kombinatorik und Graphentheorie zu ernten. Auch kommt es auf die Art der Fragestellung und des Beweises an. Gilt es, ein Beispiel oder Gegenbeispiel zu konstruieren, wie für Erdős #90? Das ist in der Regel leichter, als etwa zu beweisen, dass es sich bei der gefundenen Anordnung wirklich um die optimale handelt.

Die OpenAI-Forscher haben nun eine bessere Lösung gefunden als die von Erdős vermutete; ob es die beste ist, wissen sie nicht. Die richtig harten Nüsse der Mathematik wie die Riemann-Vermutung zu knacken, das sahen sie indes in näherer Zukunft weder in menschlicher noch in maschineller Reichweite.

Nun reklamiert OpenAI mehr oder weniger explizit für sich, mit seinem aktuellen Labormodell eines der sieben bedeutendsten Mathematikprobleme geknackt zu haben. Von denen ist bisher nur eines gelöst, und zwar durch den Mathematiker Grigori Perelman, der die Poincaré-Vermutung bewies, und zwar deutlich umfassender als gefordert. Denn er bewies sogar die allgemeinere Geometrisierungsvermutung, die die Poincaré-Vermutung als Spezialfall enthält. Spielt die OpenAI-Lösung auch in dieser Liga und ist sie überhaupt korrekt?

Prüfung steht aus

Die Navier-Stokes-Gleichungen beschreiben, wie sich Flüssigkeiten und Gase als zusammenhängende Substanz bewegen und auf äußere Kräfte reagieren. Sie sind ein Modell, das für alle praktischen Anwendungsfälle vom Ingenieurwesen bis zur Klimamodellierung perfekt funktioniert. Das Clay-Millennium-Problem dreht sich um die Frage, ob diese Gleichungen unter allen theoretisch denkbaren Bedingungen gelten. OpenAI hat es nun möglicherweise geschafft, mit seinen 10.000 Agenten ein Beispiel dafür zu konstruieren, das die Allgemeingültigkeit widerlegt. Denn unter den im Beweis konstruierten Bedingungen würde die Flüssigkeit eine unendliche Geschwindigkeit erreichen, was physikalisch unmöglich ist. Die Arbeit liegt sowohl als 166 Seiten starke Publikation sowie in einer daraus abgeleiteten formalisierten Fassung für den Beweisassistenten Lean vor. In dessen Sprache – der Dependent Type Theory – lassen sich sämtliche mathematische Annahmen und Herleitungen formulieren. Diese quasi maschinenprüfbare Form des OpenAI-Beweises passierte den Lean-4-Kernel ohne Beanstandung. Das kann man als ziemlich starken Hinweis darauf interpretieren, dass die Argumentationskette in sich schlüssig ist.

Trotzdem kann der Beweis falsch sein, nämlich dann, wenn der Reasoning-Output des Sprachmodells nicht korrekt in Lean-Definitionen und -Methodik übersetzt wurde. Das kommt immer wieder vor und ist eines der Hauptprobleme der KI-gestützten Beweisführung. Im Prinzip ergänzen sich Reasoningmodelle und die streng regelbasierte Lean-Umgebung hervorragend. Denn ein Sprachmodell allein läuft immer Gefahr, nur stichhaltig wirkende, aber in entscheidenden Details falsche Beweise zu konstruieren. Das liegt

an den stochastischen Mechanismen, denen die Textproduktion auch beim Reasoning unterliegt. Überträgt man den Output in Lean-Code, so lässt sich automatisiert überprüfen, ob sich bestimmte Aussagen tatsächlich aus den verwendeten Sätzen und Definitionen ableiten lassen. Aber wie bereits erwähnt: Auch bei der Übersetzung können Fehler passieren und deshalb müssen nun menschliche Experten detailliert nachvollziehen, ob wissenschaftlicher Text und Lean-Code tatsächlich dieselben Sachverhalte beschreiben.

Die Prüfung eines bahnbrechenden, technisch anspruchsvollen Beweises kann Jahre dauern, wie das Beispiel der Poincaré-Vermutung zeigt. Perelman selbst hatte sieben Jahre Arbeit investiert und seine Gedanken, Herleitungen und Schlussfolgerungen auf 70 Seiten niedergeschrieben. Anschließend widmeten sich mehrere unabhängige Teams dem Werk, um die Methodik nachzuvollziehen und die Beweiskette auf Schwächen abzuklopfen. Denn ein kleiner Fehler oder eine Ungenauigkeit kann genügen, schon stürzt das vermeintlich solide Gebäude wie ein Jenga-Turm in sich zusammen.

Menschliche Aufzeichnungen sind in der Regel leichter nachzuvollziehen als die von einem Sprachmodell verfassten, weil sie die Sachverhalte kompakter und meist eleganter auf den Punkt bringen. Perelman hat dies allerdings in extrem kompakter Form getan und mitunter auf Erklärungen verzichtet, die in anderen Arbeiten mehrere Seiten füllten. Zudem war der Beweis technisch äußerst anspruchsvoll, weil der russische Mathematiker dafür bahnbrechende neue Methoden entwickelte. Das schränkte den Kreis der fachlich geeigneten Prüfer ohnehin stark ein. Erschwerend kam hinzu, dass diese sich die nicht dokumentierten Zwischenschritte eigenständig erarbeiten mussten: Der Autor verweigerte jegliche über sein Paper hinausgehende Erklärung mit der Begründung, den Verifikationsprozess nicht beeinflussen zu wollen. Die kürzeste überarbeitete Version ist knapp 200 Seiten lang, die umfangreichste fasst 474 Seiten.

Das 166 Seiten starke Traktat der OpenAI-Agenten zu den Navier-Stokes-Gleichungen stellt die Prüfer wiederum vor ganz eigene Herausforderungen. Die in Textform vorliegenden Erklärungen gelten als nahezu unverständlich. Mathematiker berichten von extremen Problemen, die Logik hinter der KI-verfassten Arbeit zu entschlüsseln und dem Paper irgendetwas zu entlocken, was für einen menschlichen Leser Sinn ergibt. Javier Gómez-Serrano von der Brown University merkt an, dass die Arbeit nicht für Menschen geschrieben wurde und erst nach intensiver Überarbeitung etwas zum Fachgebiet beitragen könne. In diesem Fall kann man zumindest davon ausgehen, dass sich genügend menschliche Experten finden werden, die an einer raschen Klärung interessiert sind. Insgesamt gestaltet sich die Prüfung von KI-generierten Beweisen jedoch zunehmend schwierig: Viele maschinell erstellte Problemlösungen harren der Verifizierung, weil menschliche Zeit begrenzt ist und sich niemand für das zugrunde liegende Thema begeistern kann.

Fragen der Urheberschaft

Der zeitliche Ablauf der Geschehnisse rund um den Navier-Stokes-Beweis wirft einige Fragen rund um die Urheberschaft sowie die tatsächliche Eigenleistung des Modells auf. Der Mathematik-Professor Tristan Buckmaster hatte zusammen mit dem Anthropic-Mitarbeiter Levent Alpöge – explizit nicht im Auftrag dessen Arbeitgebers – bereits ein Jahr lang an einem vergleichbaren Lösungsansatz geforscht. Ihr Ansatz wiederum stützt sich auf die grundlegende Idee und mehrjährige Vorarbeit ihrer Kollegen Diego Córdoba und Luis Martínez-Zoroa. Dabei arbeiteten sie ebenfalls mit verschiedenen KI-Modellen, darunter auch ChatGPT. Ihre Ergebnisse veröffentlichten sie einen Tag vor der Erfolgsmeldung des KI-Unternehmens, und zwar ziemlich überstürzt zusammen mit einer Dokumentation der zeitlichen Abläufe (siehe ct.de/y6ag). Buckmaster zufolge habe er in Gesprächen mit OpenAI-Mitarbeitern erfahren, dass die Firma Kenntnis über einen bevorstehenden Durchbruch sowie die konkrete Art der Fortschritte erlangt und ihr KI-Modell, angeleitet durch ein Team von Mathematikern, erst anschließend auf die Navier-Stokes-Gleichungen angesetzt hatte. Unklar blieb in den Gesprächen, ob entscheidende Erkenntnisse aus den Chatverläufen der Wissenschaftler ins Training des noch unveröffentlichten OpenAI-Modells eingeflossen sein konnten. Außerdem sei Druck auf Buckmaster ausgeübt worden, die Autorenschaft des für den schärfsten Konkurrenten Anthropic tätigen Alpöge zu unterschlagen. OpenAI bestreitet sämtliche Punkte, bleibt konkrete Belege bisher aber schuldig (siehe ct.de/y6ag).

Doch unabhängig davon, ob tatsächlich entscheidende Erkenntnisse als Trainingsdaten ins Labormodell gelangt sind und dadurch den Weg zum Beweis geebnet haben, sieht sich die Wissenschaft mit völlig neuen Rahmenbedingungen konfrontiert – was 25 hochkarätige Mathematiker, allesamt Träger der Fields-Medaille, zu einem offenen Brief bewegte (siehe ct.de/y6ag). Darin lehnen sie den Einsatz von KI ausdrücklich nicht ab, sondern wehren sich vor allem gegen das kompetitive Verhalten der Tech-Konzerne, die die Mathematik als Benchmark-Spielwiese missbrauchten. Auf diese Weise produzierten sie lediglich Mathe-Slop, der von Menschen aufwendig verifiziert werden muss, anstatt echten Fortschritts. Denn in der Mathematik entstehe dieser im Austausch und durch die Auseinandersetzung mit den im Zuge eines Beweises entwickelten Methoden. Diesen wahren Kern ihrer Disziplin sehen die Unterzeichner spätestens seit dem Navier-Stokes-Coup stark gefährdet. *(atr@ct.de)* c't

Let $a_0 < a_1 < a_2 < \cdots$ be an infinite sequence of positive integers. Prove that there exists a unique $n \geq 1$ such that

$$a_n < \frac{a_0 + a_1 + \cdots + a_n}{n} < a_{n+1}.$$

One sensible way one might formalise a infinite sequence of integers is as a function from ℕ to ℤ. We will declare our sequence a as a variable.

```
variable (a : ℕ → ℤ)
theorem imo2014q1 (hpos : ∀ n, 0 < a n) (hinc : ∀ n, a n < a (n + 1)) :
  ∃! n : ℕ+, (a n : ℚ) < (↑∑ i : fin (n + 1), a i) / n ∧
  (↑∑ i : fin (n + 1), a i : ℚ) / n ≤ a (n + 1) := sorry
-- from src/imo/imo2014_q1.lean
```

Bild: Adrián Doña Mateo

Einen Beweis zu formalisieren, bedeutet, die mathematische Notation in die Sprache des Beweisassistenten Lean zu übersetzen, wie hier im Beispiel die zu beweisende Aussage über Folgen ganzer Zahlen.

Erdős-Probleme und Lösungen: ***ct.de/y6ag***