

Ví dụ bài toán mà AI chưa giải được, cơ sở về sự vượt trội của trực giác toán học so với AI

Hà Nội 19.10.2023

Nguyễn Thế Trung – trung@pythaverse.net

## Quick taxonomy of catastrophic AI risk

► Before we dive into discussing AI risk news and debates, let's keep in mind what (some) AI safety researchers consider to be catastrophic AI risks. As with all of this section, these should be taken with a grain of salt. Like any domain involving forecasts, and especially so with data-dependent (almost) black-box systems, there is no consensus about the actual risks that AI systems pose on a reasonable time horizon. The figure below is taken from *An Overview of Catastrophic AI Risks*, by Dan Hendrycks (Center for AI Safety).

### Malicious Use



- ✗ Bioterrorism
- ✗ Surveillance State
- ✓ Access Restrictions
- ✓ Legal Liability

### AI Race



- ✗ Automated Warfare
- ✗ Evolutionary Pressures
- ✓ International Coordination
- ✓ Safety Regulation

### Organizational Risks



- ✗ Weak Safety Culture
- ✗ Leaked AI Systems
- ✓ Information Security
- ✓ External Audits

### Rogue AIs



- ✗ Power-Seeking
- ✗ Deception
- ✓ Use-Case Restrictions
- ✓ Safety Research

stateof.ai 2023

# Cập nhật về AI – nỗi lo AI vượt trội

Nguyen The Trung:  
trung@pythaverse.net

The x-risk debate has exploded into the mainstream this year...

IDEAS • TECHNOLOGY

## Pausing AI Developments Isn't Enough. We Need to Shut it All Down



### How Rogue AIs may Arise

Published 22 May 2023 by yoshuabengio

## AI 'godfather' Geoffrey Hinton warns of dangers as he quits Google

2 May · Comments



Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

### Signatories:

☒ AI Scientists ☒ Other Notable Figures

**Geoffrey Hinton**  
Emeritus Professor of Computer Science, University of Toronto

**Yoshua Bengio**  
Professor of Computer Science, U. Montreal / Mila

**Demis Hassabis**  
CEO, Google DeepMind

**Sam Altman**  
CEO, OpenAI

**Dario Amodei**  
CEO, Anthropic

FT Magazine Artificial intelligence + Add to myFT

## We must slow down the race to God-like AI

I've invested in more than 50 artificial intelligence start-ups. What I've seen worries me

# AI giải toán



## IMO Grand Challenge

The [International Mathematical Olympiad](#) (IMO) is perhaps the most celebrated mental competition in the world and as such is among the ultimate grand challenges for Artificial Intelligence (AI).

**The challenge: build an AI that can win a gold medal in the competition.**

To remove ambiguity about the scoring rules, we propose the formal-to-formal (F2F) variant of the IMO: the AI receives a *formal* representation of the problem (in the [Lean Theorem Prover](#)), and is required to emit a *formal* (i.e. machine-checkable) proof. We are working on a proposal for encoding IMO problems in Lean and will seek broad consensus on the protocol.

## Mathematicians Are Close to Building the Perfect Periodic Table of Shapes

The table will unlock a whole new wave of discoveries.



BY DARREN ORF PUBLISHED: OCT 14, 2023

SAVE ARTICLE

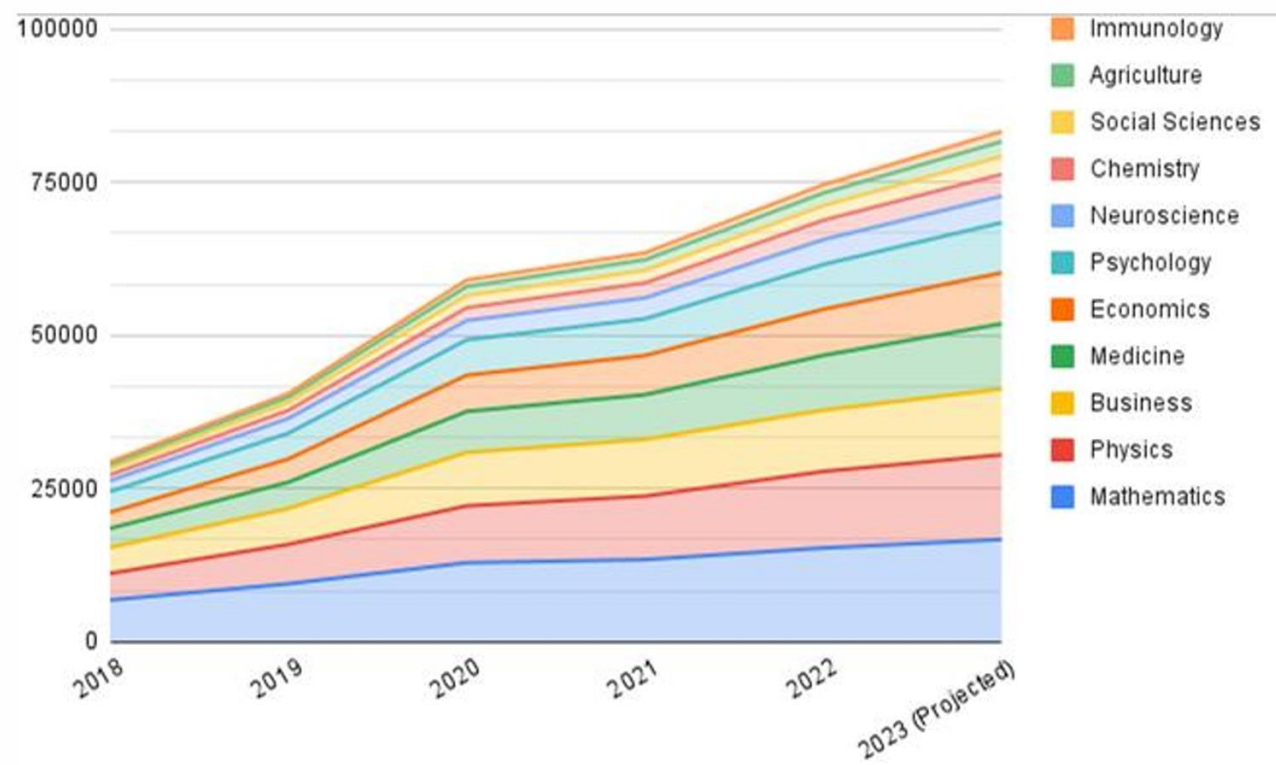


Eugene Miyazaki // Getty Images

- Fano varieties are known as the basic “atomic structure” of geometry, but are notoriously difficult to discover and organize.

## AI for science: medicine is growing fastest but mathematics captures the most attention

- The top 20 scientific fields applying AI to accelerate progress include physical, social, life and health sciences. Out of all the highest increase in the number of publications is Medicine. We expect there to be significant research breakthroughs in the foreseeable future as a result of AI's use in the sciences.





# Có giới hạn nào của AI

- Có nhiều cách tính về giới hạn của AI, tuy nhiên giới hạn vật lý là dễ hình dung nhất.
- Theo báo cáo thì toàn bộ dữ liệu số của thế giới năm 2023 vào khoảng 200 zettabytes và *khả năng cao nhất thì nhân đôi mỗi năm.*
  - *Tương đương  $2 \cdot 10^{23}$  hay ít hơn  $2 \cdot 10^{24}$  bits*
- *Như vậy nếu một bài toán cần lưu nhiều hơn số này thì sẽ vượt quá khả năng tính toán và vì thế vượt qua khả năng của AI về mặt lý thuyết.*



Câu trả lời từ ChatGPT

# Có bài toán nào vượt quá giới hạn của AI?

TO

Hãy cho Ví dụ bài toán mà AI chưa giải được làm cơ sở về sự vượt trội của trực giác toán học so với AI



Vấn đề "Đồng nhất Vật Lý" (Grand Unified Theory) là một bài toán mà AI chưa thể giải quyết đến nay, và đây là một lĩnh vực mà trực giác toán học của con người có thể chơi một vai trò quan trọng. Đồng nhất Vật Lý là việc kết hợp ba trong bốn lực cơ bản của vũ trụ—lực điện từ, lực yếu, và lực mạnh—vào một mô hình toán học duy nhất. Lực hấp dẫn, được mô tả bởi Thuyết Tương đối của Einstein, cũng là một mục tiêu để đồng nhất, nhưng đây là một nhiệm vụ còn khó khăn hơn.

Một số lý do vì sao trực giác toán học vượt trội hơn AI trong trường hợp này:

1. Sáng tạo Học Thuật: Mặc dù AI có thể tối ưu hóa các mô hình dựa trên dữ liệu đầu vào, nhưng nó thiếu khả năng "sáng tạo" các mô hình toán học mới một cách độc lập, điều này rất quan trọng trong việc phát triển các lý thuyết mới.
2. Khả Năng Đánh Giá Tiềm Năng: Con người có khả năng đánh giá và tìm hiểu sự quan trọng của các khái niệm toán học mới, cũng như cách chúng có thể áp dụng trong các lĩnh vực khác của khoa học. AI hiện tại chưa có khả năng này.
3. Kết nối Qua Các Lĩnh Vực: Trực giác toán học giúp các nhà khoa học kết nối các lĩnh vực toán học và vật lý khác nhau, một điều mà AI hiện tại chưa thể thực hiện một cách hiệu quả.
4. Khả Năng Tự Học: Mặc dù AI có khả năng học từ dữ liệu, nó cần phải được "huấn luyện"

trước. Nhưng trực giác toán học cho phép con người tiếp thu và áp dụng các kiến thức mới một cách nhanh chóng.

Nguyen The Trung: [trung@pythaverse.net](mailto:trung@pythaverse.net)

# Có bài toán nào vượt quá giới hạn của AI?

- **Monster group:**
  - **Monster Group là nhóm đối xứng lớn nhất trong dãy các "Sporadic Groups" và có kích thước khoảng  $8 \cdot 10^{53}$**
- Monstrous Moonshine là một trong những câu chuyện kỳ thú nhất trong toán học, nối liền đại số và phân tích phức tạp thông qua một cách không ngờ tới. Lịch sử của nó bắt đầu vào năm 1978, khi John McKay, một nhà đại số, phát hiện ra một sự trùng hợp số học giữa các hệ số trong công thức  $j$ -invariant, một hàm trong toán học biểu diễn các cấu trúc trên các diện Riemann, và số chiều của các biểu diễn không đối xứng của Monster Group.
- McKay thông báo sự phát hiện này cho John H. Conway và Simon P. Norton, và đặt ra câu hỏi: liệu có sự liên kết sâu xa nào giữa  $j$ -invariant và Monster Group hay không? Conway và Norton thử nghiệm này và đưa ra "Monstrous Moonshine Conjecture," đề xuất rằng có một liên hệ chặt chẽ giữa các hàm biểu diễn của Monster Group và các hàm mô đun (modular forms).
- Năm 1992, Richard Borcherds chứng minh được Monstrous Moonshine, và việc này giúp ông giành được Huy chương Fields năm 1998. Chứng minh của Borcherds sử dụng công cụ từ đại số vertex, một lĩnh vực mới của đại số đối xứng, cũng như sự hiểu biết sâu sắc về các hàm mô đun và K-theory.
- Về cơ bản, Monstrous Moonshine mô tả một liên kết giữa hai lĩnh vực toán học mà trước đây được coi là không liên quan: lý thuyết nhóm và lý thuyết mô đun. Nó đã mở ra các đường nghiên cứu mới và kết nối giữa các lĩnh vực toán học và vật lý.
- Liên kết tới các tài liệu liên quan:
  - [Monstrous Moonshine, the Monster Lie Algebra, and the Classification of CFT](#)
  - [Richard Borcherds and Monstrous Moonshine](#)



In 1978, [John McKay](#) found that the first few terms in the [Fourier expansion](#) of the normalized [J-invariant](#) (sequence [A014708](#) in the [OEIS](#)),

$$J(\tau) = \frac{1}{q} + 196884q + 21493760q^2 + 864299970q^3 + 20245856256q^4 + \cdots$$

with  $q = e^{2\pi i\tau}$  and  $\tau$  as the [half-period ratio](#) could be expressed in terms of [linear combinations](#) of the [dimensions](#) of the [irreducible representations](#)  $r_n$  of the monster group  $M$  (sequence [A001379](#) in the [OEIS](#)) with *small* non-negative coefficients. Let  $r_n = 1, 196883, 21296876, 842609326, 18538750076, 19360062527, 293553734298, \dots$  then,

$$\begin{aligned} 1 &= r_1 \\ 196884 &= r_1 + r_2 \\ 21493760 &= r_1 + r_2 + r_3 \\ 864299970 &= 2r_1 + 2r_2 + r_3 + r_4 \\ 20245856256 &= 3r_1 + 3r_2 + r_3 + 2r_4 + r_5 = 2r_1 + 3r_2 + 2r_3 + r_4 + r_6 \\ 333202640600 &= 5r_1 + 5r_2 + 2r_3 + 3r_4 + 2r_5 + r_7 = 4r_1 + 5r_2 + 3r_3 + 2r_4 + r_5 + r_6 + r_7 \end{aligned}$$

where the LHS are the coefficients of  $j(\tau)$  while the RHS are the dimensions  $r_n$  of the monster group  $M$ . (Since there can be several linear relations between the  $r_n$  such as  $r_1 - r_3 + r_4 + r_5 - r_6 = 0$ , the representation may be in more than one way.) McKay viewed this as evidence that there is a naturally occurring infinite-dimensional [graded representation](#) of  $M$ , whose [graded dimension](#) is given by the coefficients of  $J$ , and whose lower-weight pieces decompose into irreducible representations as above. After he informed [John G. Thompson](#) of this observation, Thompson suggested that because the graded dimension is just the graded [trace](#) of the [identity element](#), the graded traces of nontrivial elements  $g$  of  $M$  on such a representation may be interesting as well.

Conway and Norton computed the lower-order terms of such graded traces, now known as McKay–Thompson series  $T_g$ , and found that all of them appeared to be the expansions of [Hauptmoduln](#). In other words, if  $G_g$  is the subgroup of  $\mathrm{SL}_2(\mathbf{R})$  which fixes  $T_g$ , then the [quotient](#) of the [upper half](#) of the [complex plane](#) by  $G_g$  is a [sphere](#) with a finite number of points removed, and furthermore,  $T_g$  generates the [field](#) of [meromorphic functions](#) on this sphere.

Based on their computations, Conway and Norton produced a list of Hauptmoduln, and conjectured the existence of an infinite dimensional graded representation of  $M$ , whose graded traces  $T_g$  are the [expansions](#) of precisely the functions on their list.

In 1980, [A. Oliver L. Atkin](#), Paul Fong and Stephen D. Smith produced strong computational evidence that such a graded representation exists, by decomposing a large number of coefficients of  $J$  into representations of  $M$ . A graded representation whose graded dimension is  $J$ , called the moonshine module, was explicitly constructed by [Igor Frenkel](#), [James Lepowsky](#), and [Arne Meurman](#), giving an effective solution to the McKay–Thompson conjecture, and they also determined the graded traces for all elements in the centralizer of an involution of  $M$ , partially settling the Conway–Norton conjecture. Furthermore, they showed that the [vector space](#) they constructed, called the Moonshine Module  $V^\natural$ , has the additional structure of a [vertex operator algebra](#), whose [automorphism group](#) is precisely  $M$ .

In 1985, the [Atlas of Finite Groups](#) was published by a group of mathematicians, including [John Conway](#). The Atlas, which enumerates all [sporadic groups](#), included "Moonshine" as a section in its list of notable properties of the [monster group](#).<sup>[4]</sup>

Borcherds proved the Conway–Norton conjecture for the Moonshine Module in 1992. He won the [Fields Medal](#) in 1998 in part for his solution of the conjecture.

# Sức mạnh của trực giác toán học

---

Câu chuyện của Moonshine theory cho thấy các nhà toán học bằng trực giác đã liên kết được đặc điểm của Monster Group với hàm biểu diễn bề mặt Riemann.

---

Nếu việc này dựa vào tính toán thuần túy thì thực tế là không đủ Ổ cứng để chứa dữ liệu của tập Monster, chưa nói đến khảo sát

---

Như vậy có thể kết luận trực giác của các nhà toán học hiện vượt trội khả năng của tính toán hiện nay bao gồm cả AI.

---

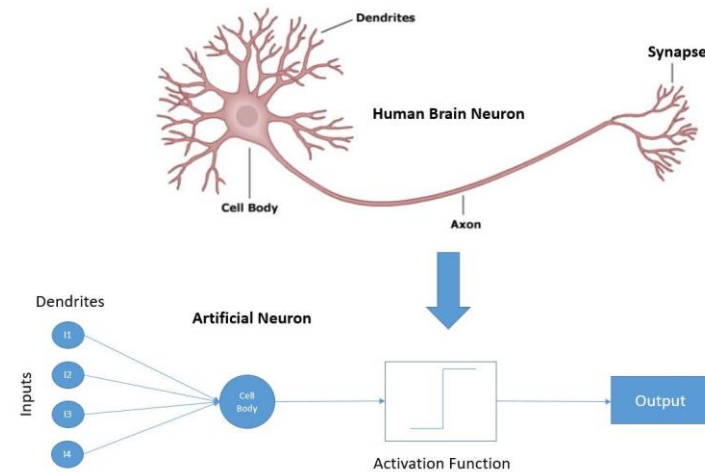
Tuy nhiên AI thì giỏi toán hơn 99% các người đang làm toán là hoàn toàn có thể (ước lượng ChatGPT 4 IQ là 155 – Einstein là 160)

---

Câu hỏi đặt ra là trực giác toán học đến từ đâu?

# Câu hỏi tổng quát là trực giác đến từ đâu

- Nếu não bộ chỉ là một siêu máy tính 01 với các neuron và synapse thì ước lượng về bộ nhớ như sau
- Theo một ước lượng được tiến hành bởi Paul Reber, một giáo sư tâm lý học tại Northwestern University, bộ não có khả năng lưu trữ khoảng 1 triệu gigabytes (1 petabyte) dưới dạng thông tin. Điều này dựa trên số lượng neuron và cách mỗi neuron có thể kết nối với các neuron khác.
  - Tương đương  $2 \times 10^{15}$
- Điều này có nghĩa là não người cũng không thể đủ để khảo sát Monster Group và hiện nay ChatGPT 4 có khoảng 170 tỷ
- **Mô hình LLM trong GPT:** GPT-3.5 có khoảng 175 tỷ tham số và GPT 4 nghe đồn có 1 ngàn tỷ tham số tương đương  $10^{12}$
- Điều này có nghĩa nếu não người cũng chỉ là máy tính toán thì GPT4 có thể vượt trội, và có vẻ như vậy.



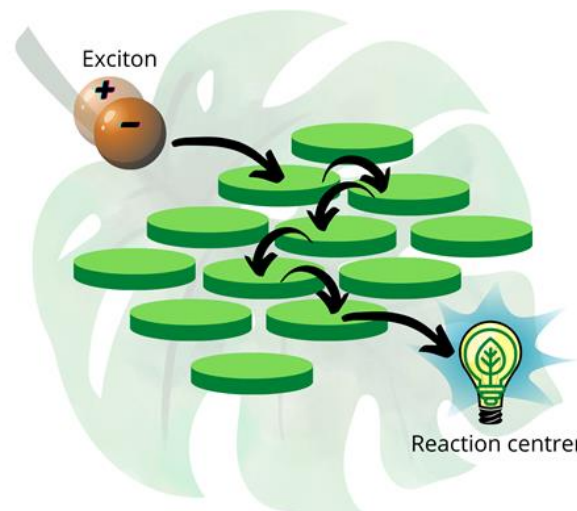
Tài liệu liên quan: [How much information can the brain hold? - Paul Reber](<https://www.scientificamerican.com/article/what-is-the-memory-capacity/>).

# Tuy nhiên: não bộ có thể có cơ chế lượng tử

- Stuart Hameroff, cùng với Roger Penrose, đã đề xuất một mô hình có tên là Orchestrated Objective Reduction (Orch OR), trong đó họ cho rằng các trạng thái lượng tử có thể được thiết lập và duy trì trong các cấu trúc protein có tên là microtubules, nằm trong tế bào thần kinh của não. Họ đề xuất rằng các trạng thái lượng tử này có thể tồn tại trong một môi trường nhiệt động lực không ổn định nhờ vào một quá trình gọi là "ổn định nhiệt động lực lượng tử" (quantum thermodynamic stability).
  1. **Ổn Định Nhiệt Động Lực:** Hameroff và Penrose cho rằng các trạng thái lượng tử trong các microtubules có thể được ổn định bằng cách sử dụng năng lượng sinh học từ ATP (adenosine triphosphate) và GTP (guanosine triphosphate).
  2. **Vị Trí của Microtubules:** Theo Hameroff, microtubules có thể là các cấu trúc rất bảo vệ, giúp giữ cho các trạng thái lượng tử không bị phá vỡ bởi nhiệt động lực của môi trường sống.
  3. **Quá Trình Orch OR:** Trong mô hình Orch OR, quá trình sụp đổ lượng tử của các trạng thái lượng tử trong microtubules được cho là có thể tương tác với các quá trình sinh học, tạo ra những "moment" của ý thức.
  4. **Thách Thức và Phản Bác:** Phải nói rằng, dù lý thuyết này rất thú vị, nó vẫn gặp nhiều chỉ trích và chưa được chấp nhận rộng rãi trong cộng đồng khoa học. Một số thách thức chính đến từ việc thiếu bằng chứng thực nghiệm và khả năng duy trì trạng thái lượng tử trong điều kiện sinh học.
    1. Tuy nhiên trong 1 chiếc lá cũng có phản ứng lượng tử trong quang học

# Hiệu ứng lượng tử trong quá trình quang hợp

- Cách các exciton (quasiparticle do sự kích thích của một photon tạo ra) được chuyển động thông qua các phức hợp protein như các trung tâm phản ứng và anten mật độ cao. Hiện tượng này thể hiện rõ nhất trong việc các exciton có thể di chuyển qua một số con đường khác nhau một cách hiệu quả, tối ưu hóa việc chuyển đổi năng lượng mặt trời thành năng lượng hóa học.



# Một vài suy đoán

1. Chúng ta hi vọng bộ não con người nhiều hơn là một máy tính 01
2. Nếu bộ não hoạt động theo cơ chế lượng tử thì trực giác có thể được giải thích thông qua cơ chế vướng víu lượng tử
3. Sức mạnh tập thể là có thật và vượt trội khi có vướng víu lượng tử
4. Tuy nhiên máy tính lượng tử cũng đang hình thành và sớm có thể vượt số lượng vi ống trong não ( có thể vào khoảng năm 2100-2200)
5. Điều may là lúc đó nếu những người có khả năng trực giác cao vẫn có thể “điều khiển” các hệ thống máy tính lượng tử này.
6. Vì vậy con người vẫn có thể hi vọng vượt lên AI, và các nhà toán học chắc sẽ còn chứng minh việc này dài dài.





# Một phỏng đoán

*Monster group với tính rời rạc của nó, có thể là biểu diễn của các chuyển đổi tự trội (emergent) – như trong lượng tử hay năng lực của các mô hình AI cỡ lớn.*



Trân trọng cảm ơn