AI in Science : OECD work

Vienna - 24th January 2024

Alistair Nolan, Senior Policy Analyst, Directorate for Science and Technology Policy, OECD

Artificial Intelligence in Science

CHALLENGES, OPPORTUNITIES AND THE FUTURE OF RESEARCH

OECD

Today's presentation

Why AI in science matters

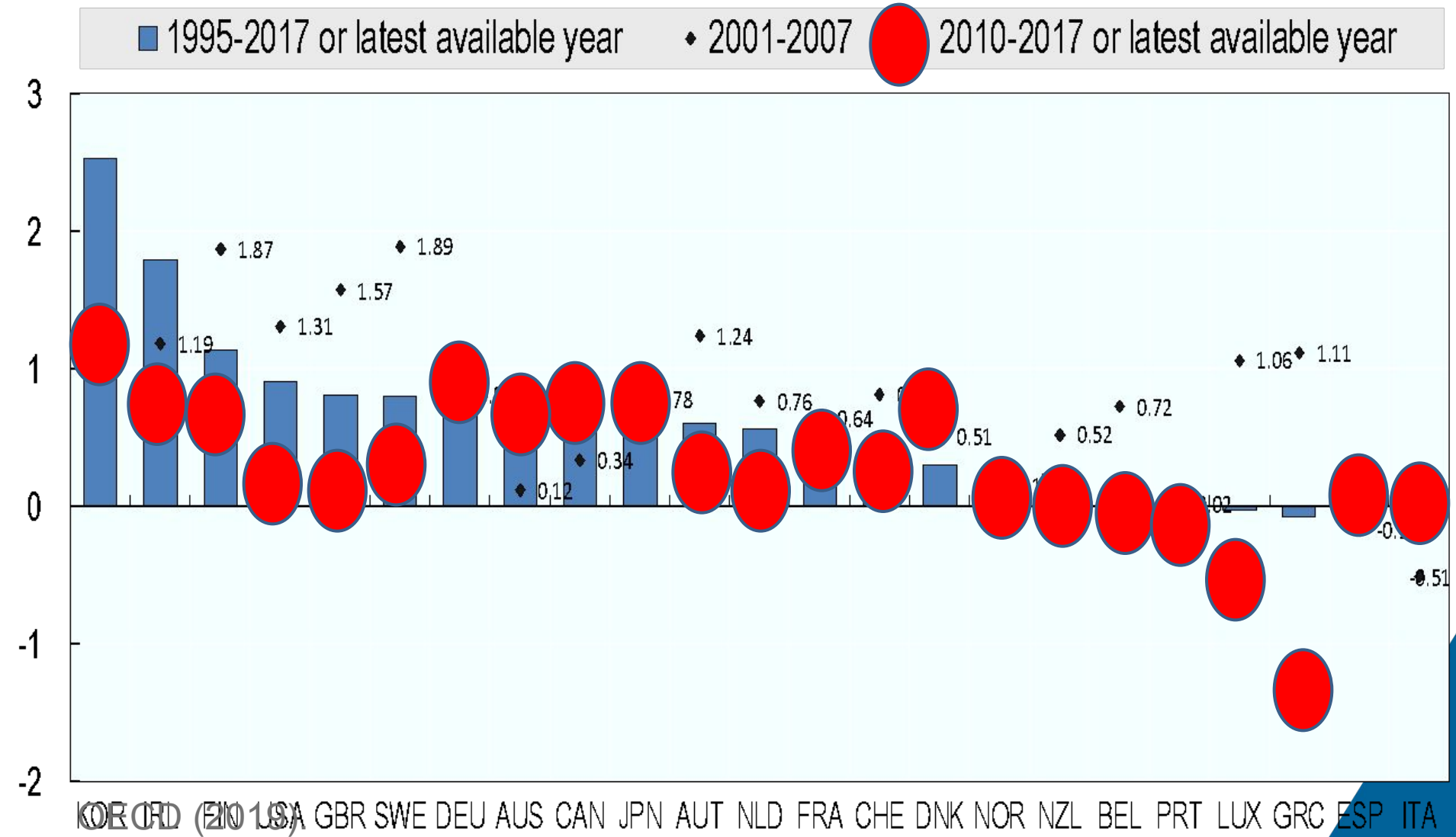
AI in science today and tomorrow

Impacts of AI in science so far

Public policy

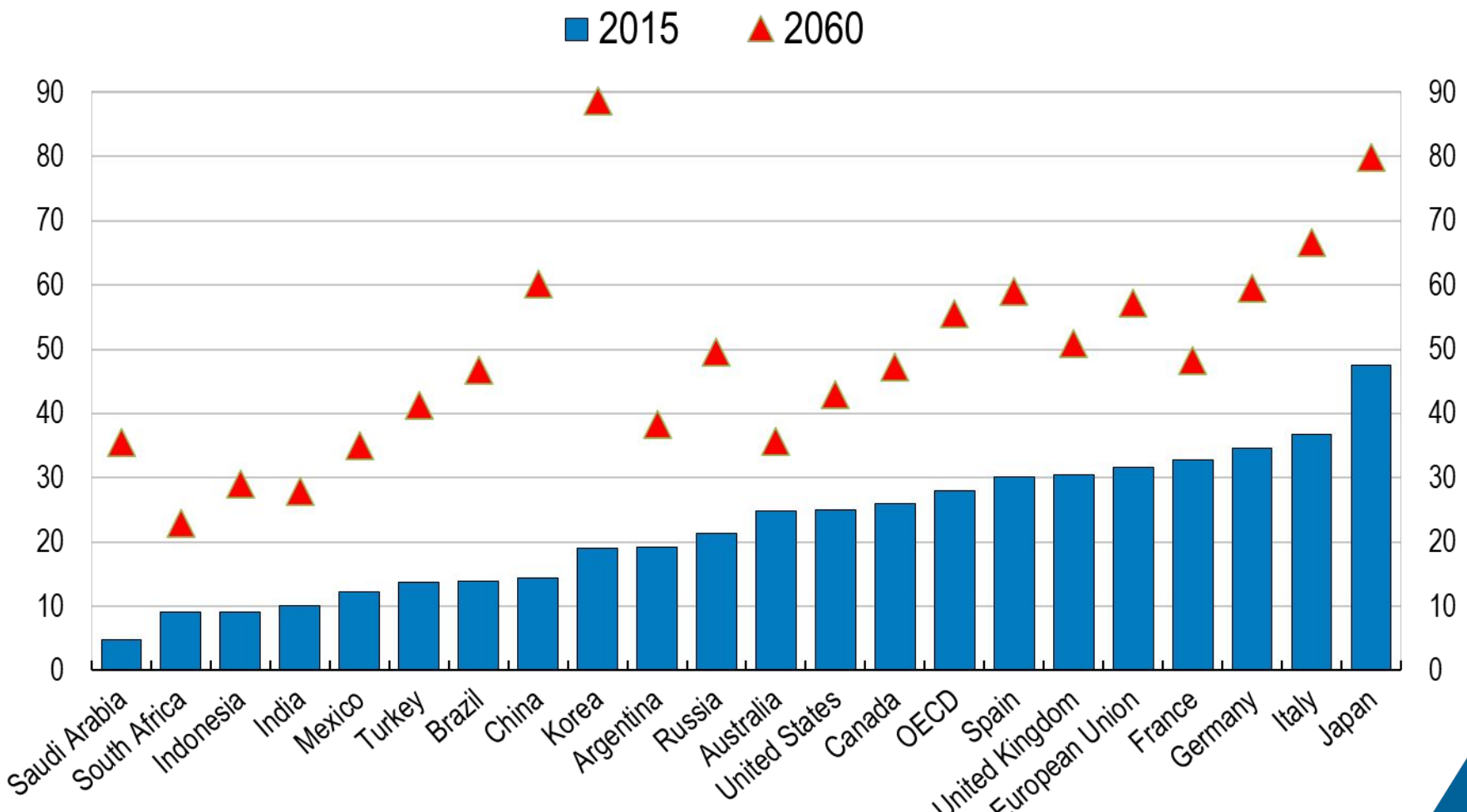
Why we need more AI in science

Slower growth of total factor productivity



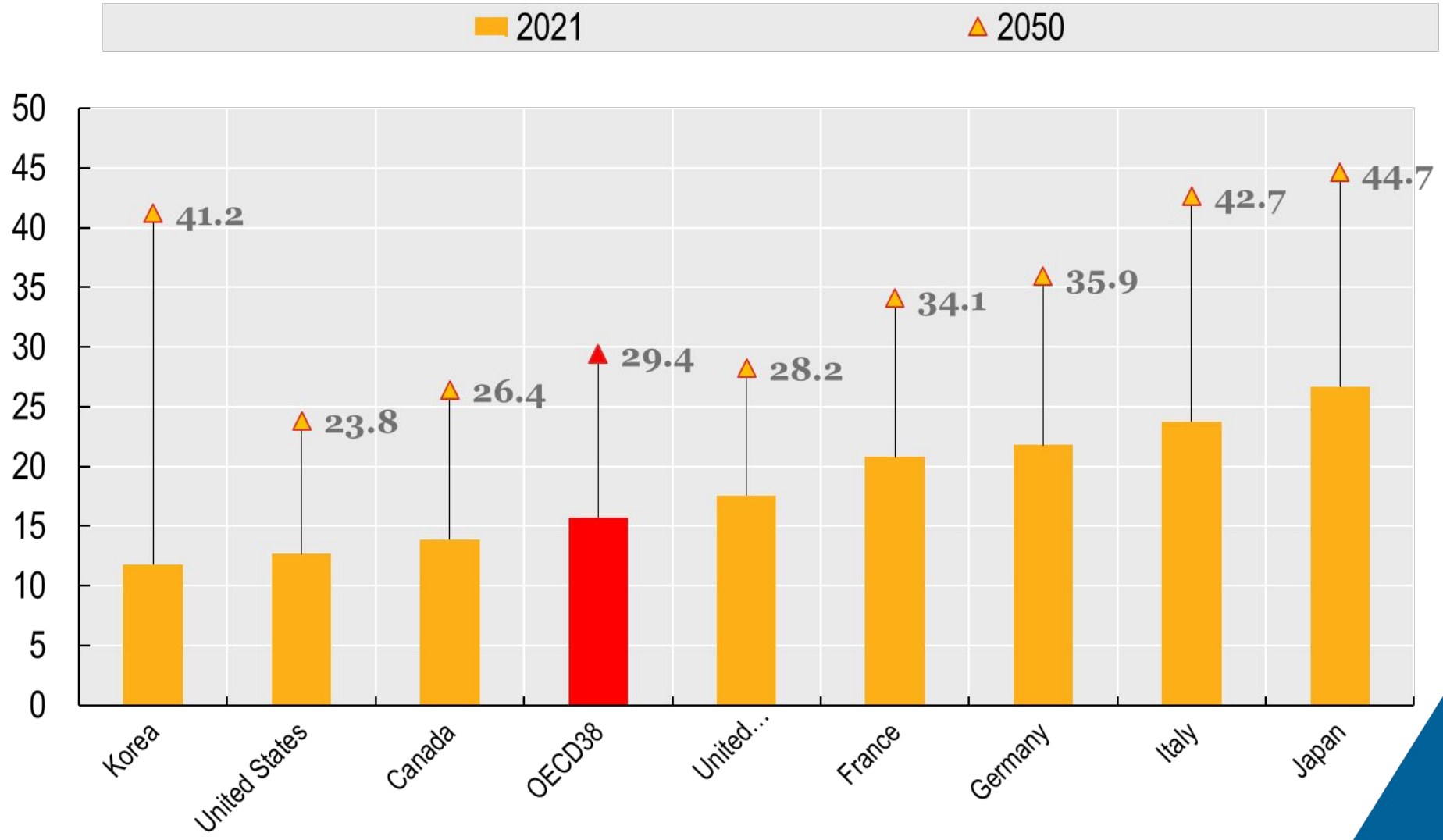
Old-age dependency ratios are projected to at least double in most G20 countries by 2060

Number of people older than 65 years per 100 of working-age (20-64)



Source: Fiscal challenges and inclusive growth in ageing societies (OECD, 2019)

People with dementia per 1000 population, 2021 and 2050



Source: Health at a Glance (OECD, 2021)

Breakthroughs in climate-relevant fields, such as materials science

Ultra-light materials – possible uses in improving fuel efficiency in aerospace

Willow glass – strong flexible ultra-thin glass, for low-cost solar cells

Is science getting harder ?

And recent attention to the productivity of research spurred by

the NATIONAL BUREAU of ECONOMIC RESEARCH

Are Ideas Getting Harder to Find?

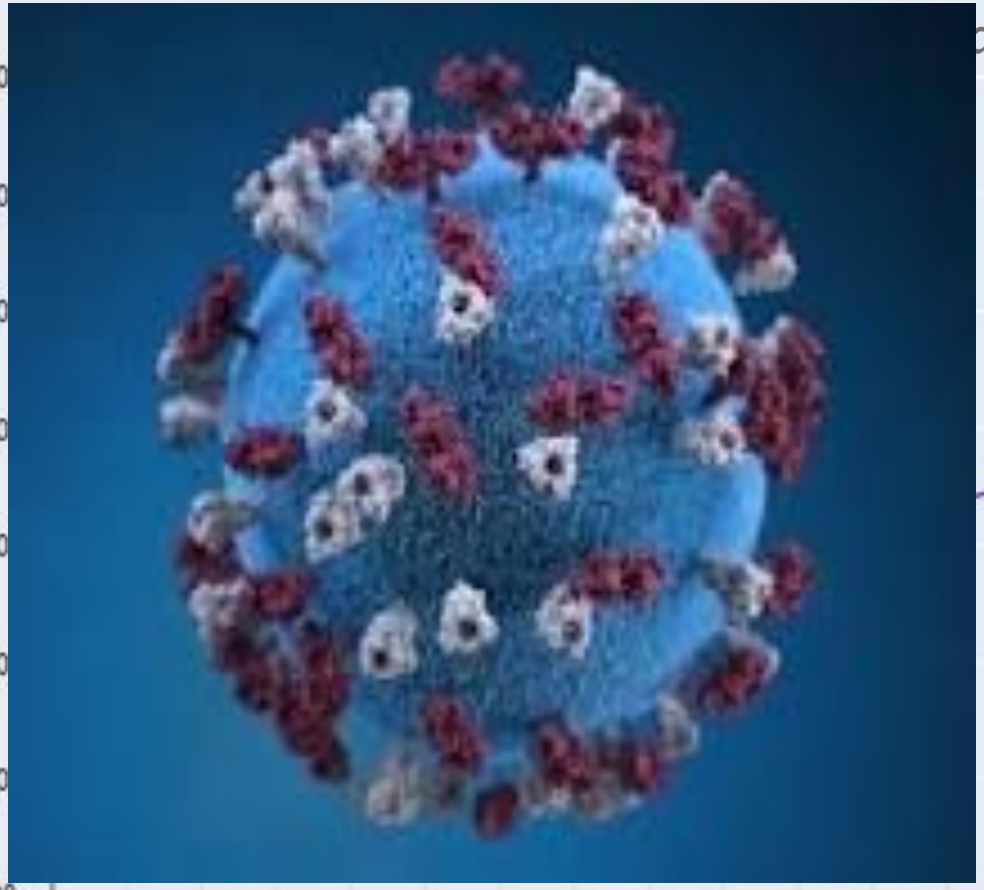
Nicholas Bloom, Charles I. Jones, John Van Reenen, Michael Webb

NBER Working Paper No. 23782
Issued in September 2017
NBER Program(s): Economic Fluctuations and Growth, Productivity, Innovation, and Entrepreneurship

In many growth models, economic growth arises from people creating ideas, and the long-run growth rate is the product of two terms: the effective number of researchers and their research productivity. We present a wide range of evidence from various industries, products, and firms showing that research effort is rising substantially while research productivity is declining sharply. A good example is Moore's Law. The number of researchers required today to achieve the famous doubling every two years of the density of computer chips is more than 18 times larger than the number required in the early 1970s. Across a broad range of case studies at various levels of (dis)aggregation, we find that ideas — and in particular the exponential growth they imply — are getting harder and harder to find. Exponential growth results from the large increases in research effort that offset its declining productivity.

Information overload

(annual number of scientific publications, 1996-2018)



100,000 scientific papers on coronavirus in year 1 of the pandemic

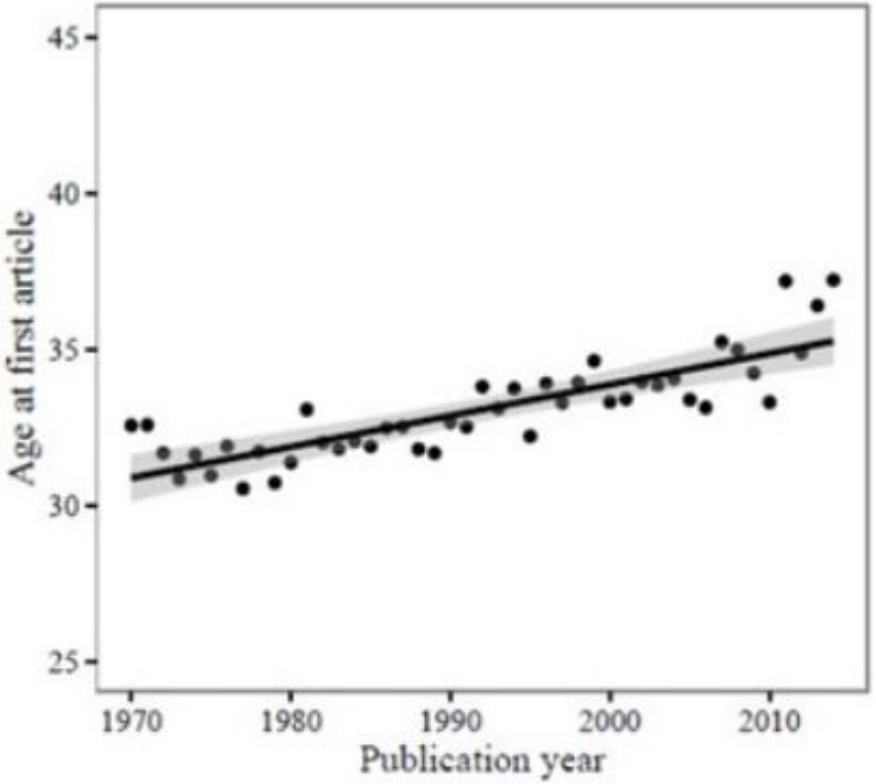
4,000,0
3,500,0
3,000,0
2,500,0
2,000,0
1,500,0
1,000,0
500,000
0

1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018

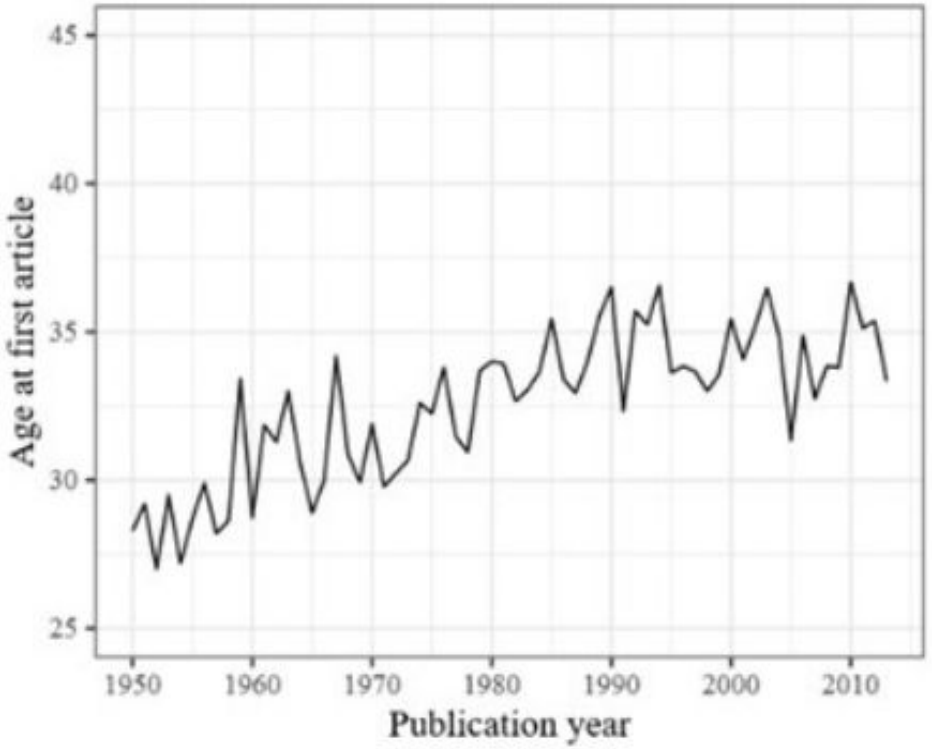
Scientific publications

Creating a knowledge burden ?

Age at first solo economics article



Age at first solo (top) mathematics article



Schweitzer and Brendel (2020) Brendel and Schweitzer (2019)

Discovery getting harder ?

$$F = m \times a$$ **1686**

$$\ln \frac{K_2}{K_1} = \frac{-\Delta H^{\emptyset}}{R}\left(\frac{1}{T_2} - \frac{1}{T_1}\right)$$ **1884**

$$\left(1 - e^{-2\Delta}\right) r^{D-3} = \frac{2\kappa}{D-2}\int_0^r \rho(r') r'^{D-2}\, dr' = \frac{2 \cdot 8\pi(D-3)G}{(D-2)} \frac{M}{\Omega_{D-2}} \Rightarrow$$

$$\Rightarrow \frac{1}{3} \frac{4\left[anti\log \frac{\int_0^{\infty} \frac{\cos \pi x w'}{\cosh \pi x} e^{-\pi x^2 w'} dx}{e^{-\frac{\pi^2}{4} w'} \varphi_{w'}(itw')} \right] \cdot \frac{\sqrt{142}}{t^2 w'}}{\log\left[\sqrt{\left(\frac{10+11\sqrt{2}}{4}\right)} + \sqrt{\left(\frac{10+7\sqrt{2}}{4}\right)} \right]} . \quad (2.93c)$$ **1973**

Thanks to Prof. Hugh Cartwright, Oxford University

AI : Coming to scientific knowledge in new ways

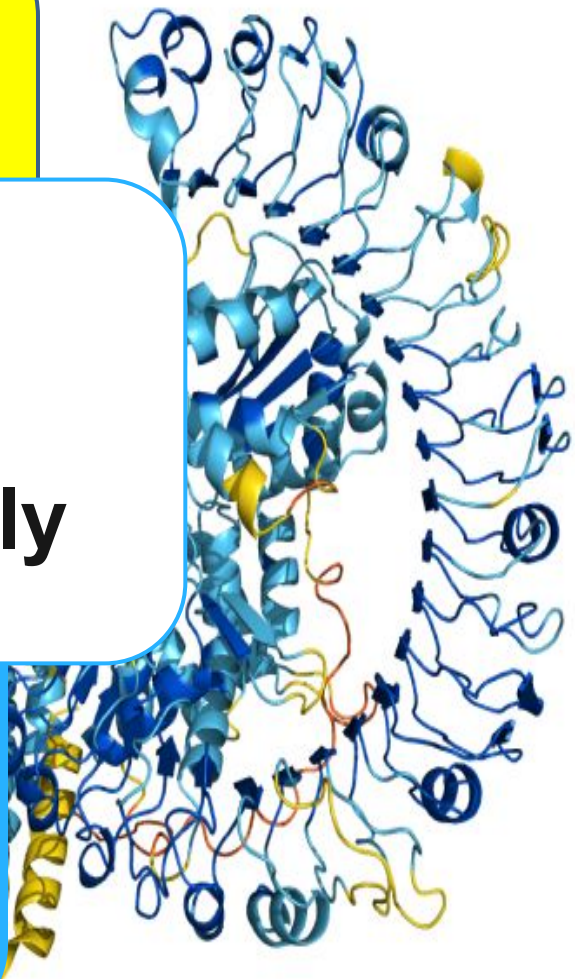
DeepMind: predicting protein folding

- alanine - A
- arginine - R
- asparagine - N
- aspartic acid - D
- cysteine - C
- glutamine - Q
- glutamic acid - E
- glycine - G
- histidine - H
- isoleucine - I
- leucine - l
- lysine - K
- methionine - M
- phenylalanine - F
- proline - P
- serine - S
- threonine - T
- tryptophan - W
- tyrosine - Y
- valine -V

How to predict the 3D structure of a protein fro its

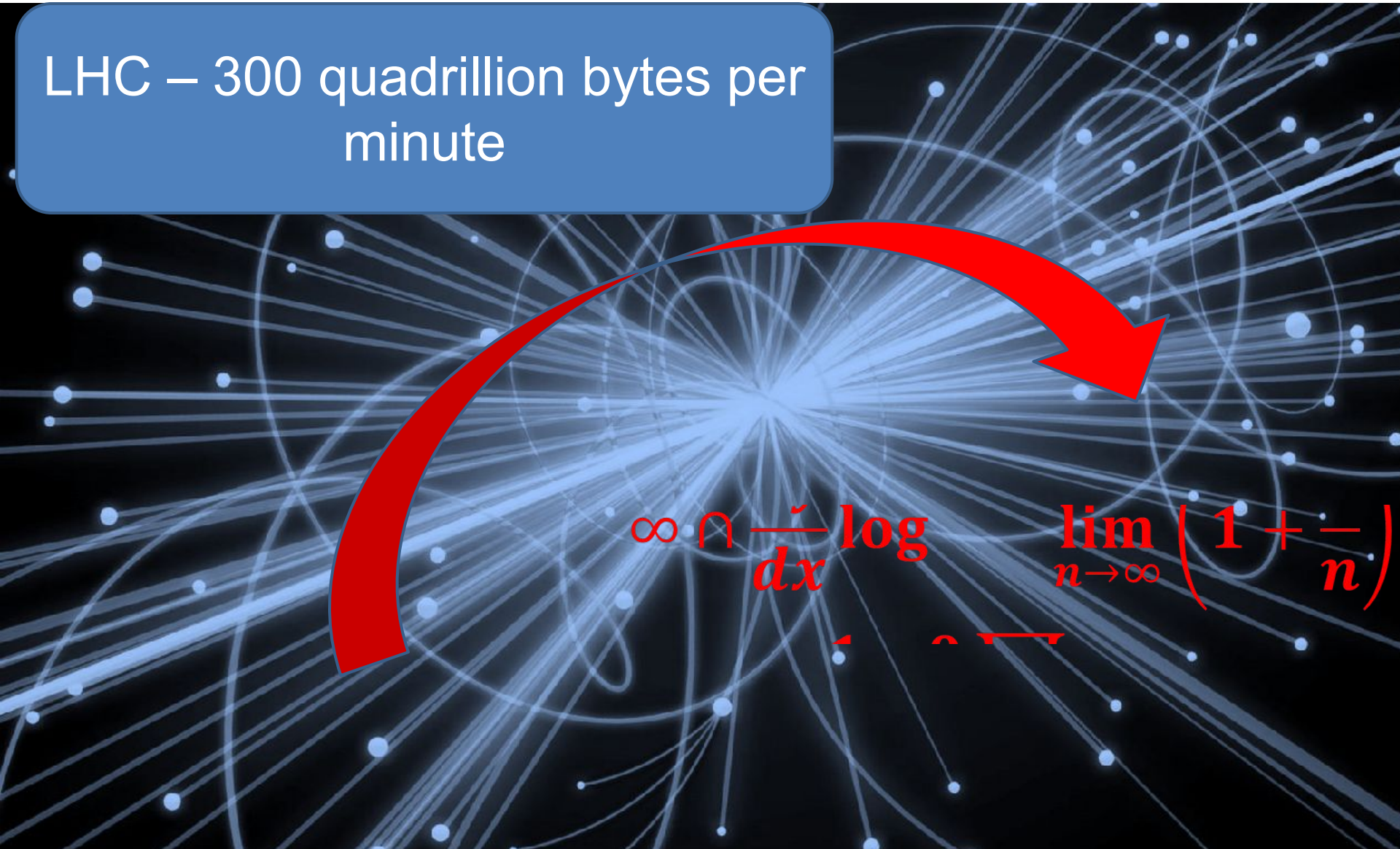
Alphafold2 **solves this computationally**

Takes experiments to solve for just one protein (2 million proteins in the human body)

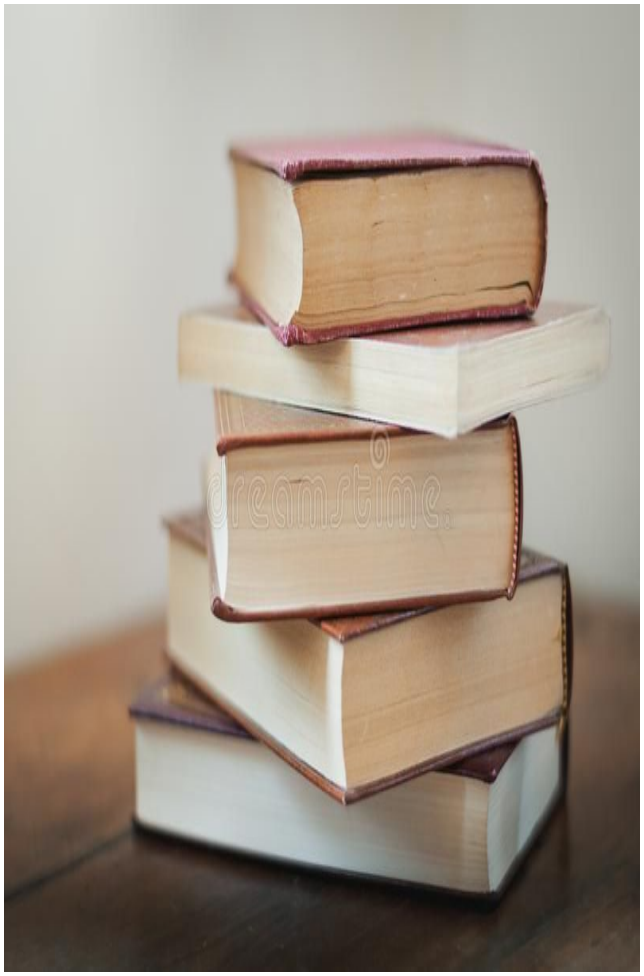


Generating hypotheses from vast datasets

LHC – 300 quadrillion bytes per minute



Finding undiscovered public knowledge (knowledge we don't know we have)



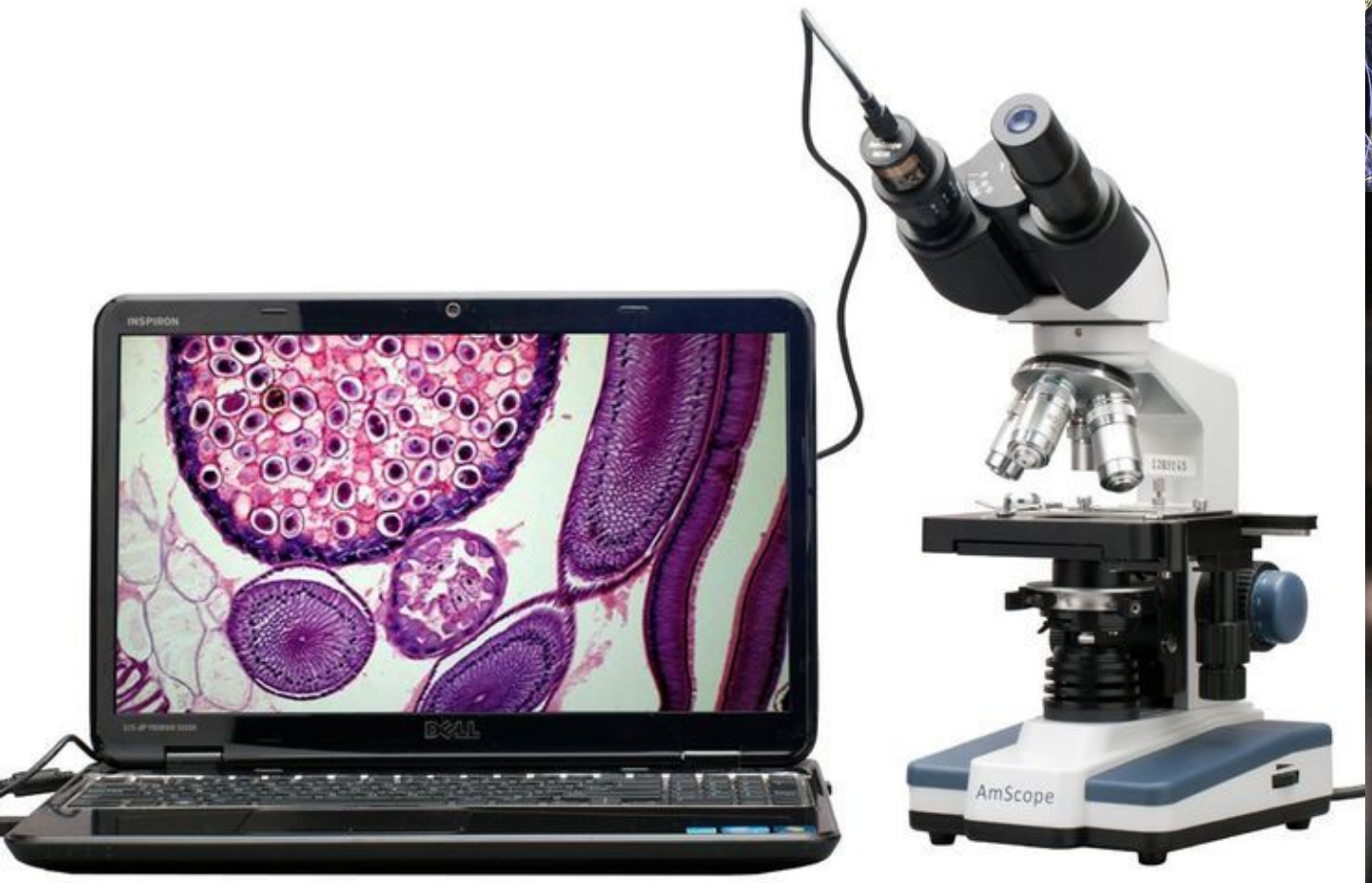
Book 1 shows "A affects B"

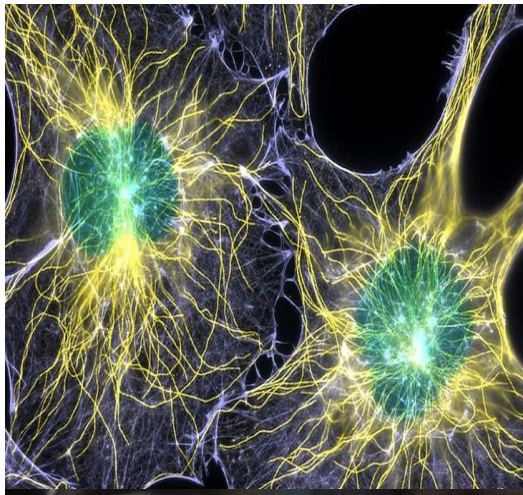
Book 2 shows "B affects C"

Then "A may affect C"

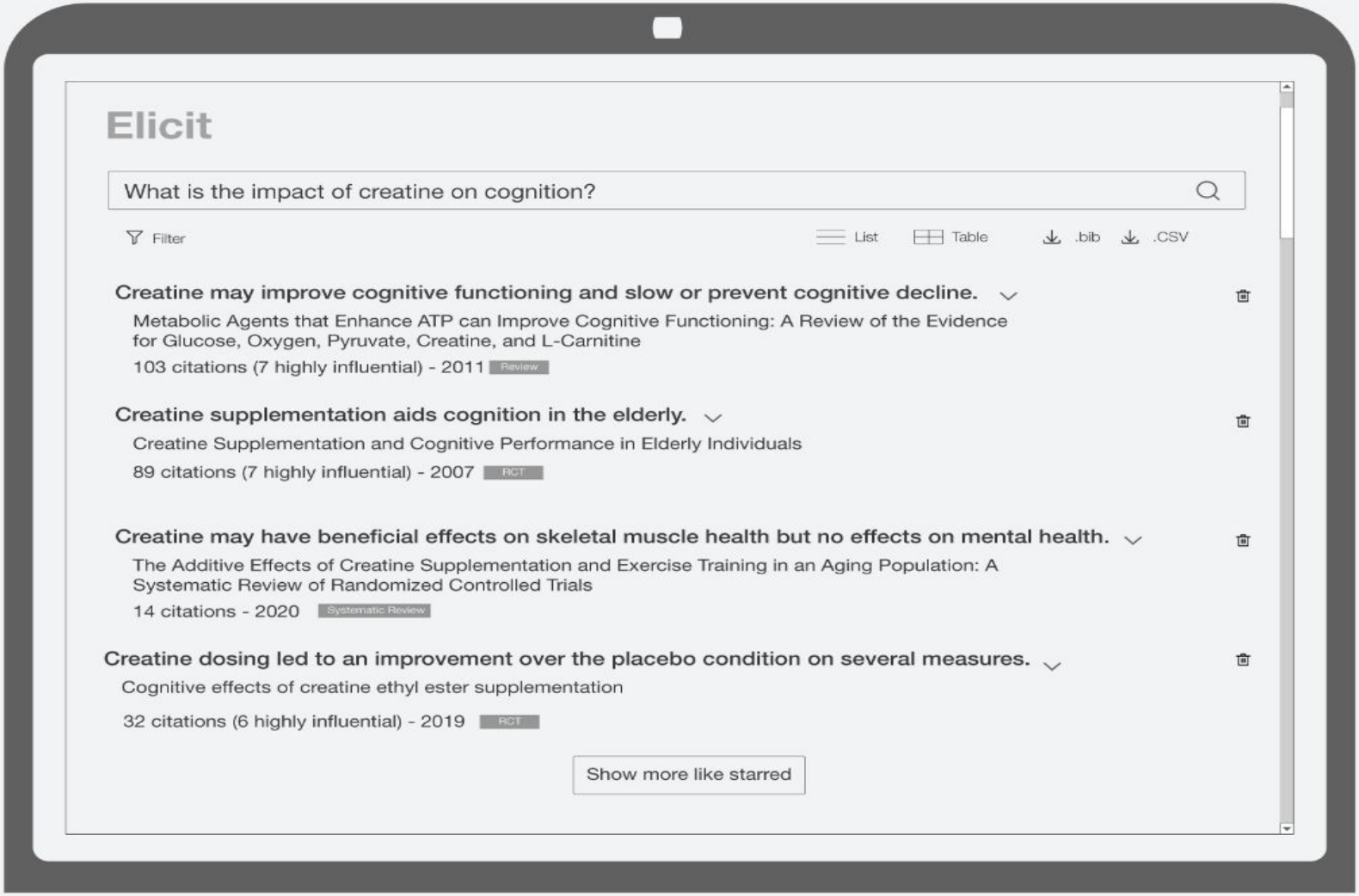
Novel simulation

Revolutionising microscopy

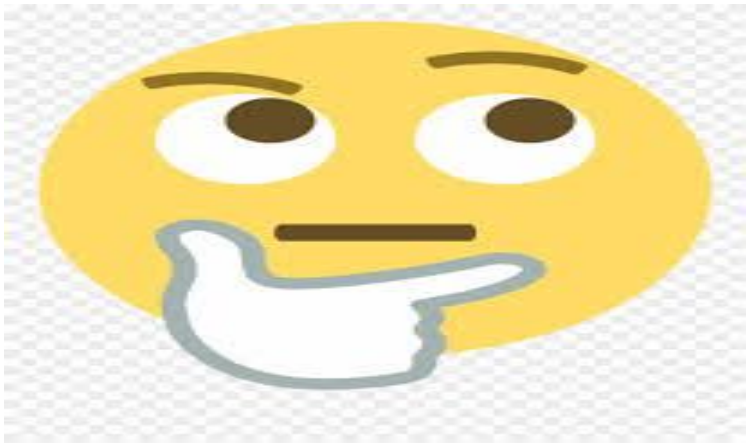




Elicit – (Ought.com) - AI Research assistant – using GPT3



Many other possible AI applications



Planning experiments

Professor Ross King in front of Adam, the robot scientist

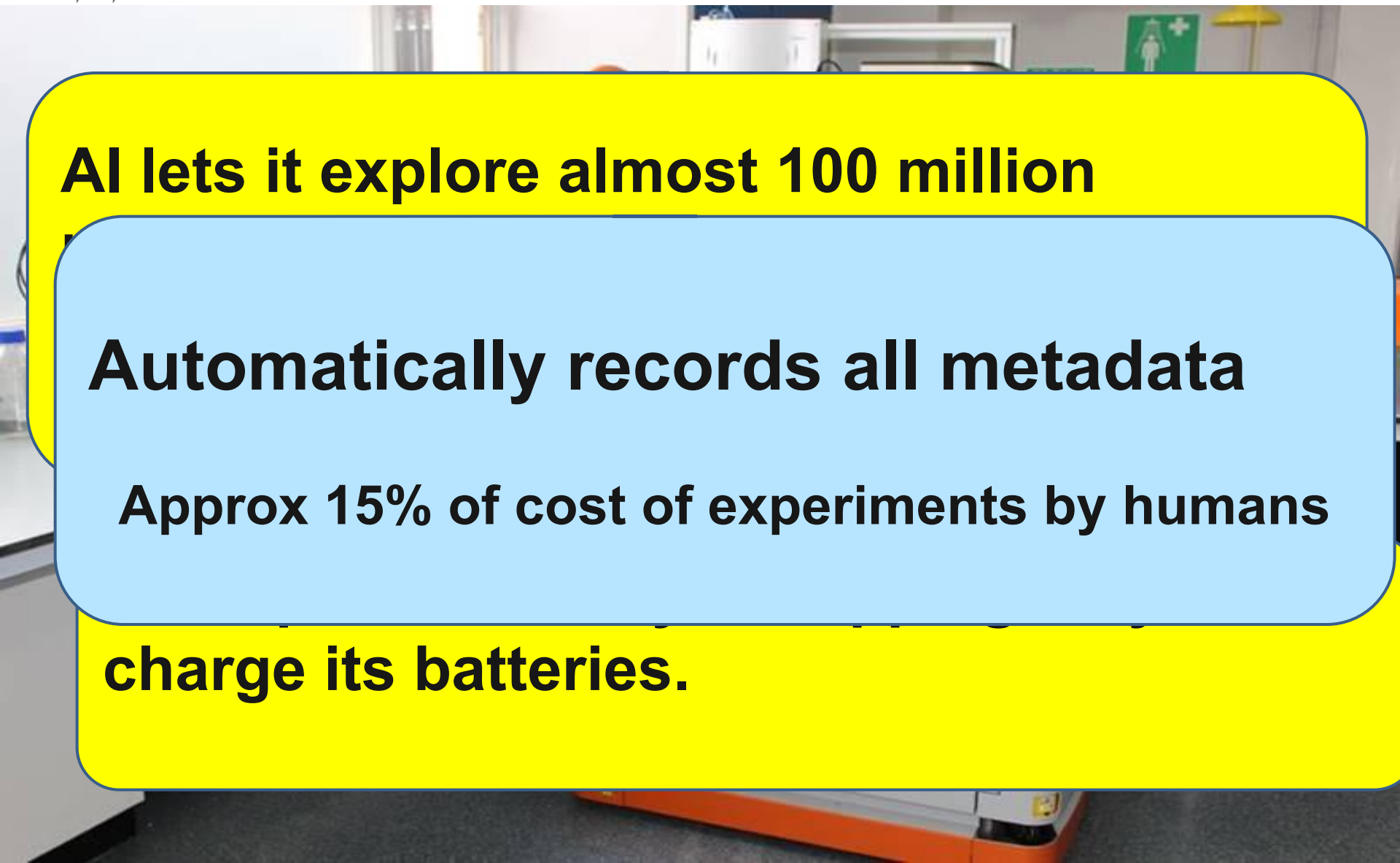
Triclosan – works against wild-type and drug resistant Plasmodium falciparum, and Plasmodium vivax.

2008-2015 Eve – Drug Design for Tropical Diseases

Williams *et al*. (2015) *Royal Society Interface*, DOI 10.1098/rsif.2014.1289

Effects on research productivity ?

Robot chemist at the University of Liverpool



AI lets it explore almost 100 million

Automatically records all metadata

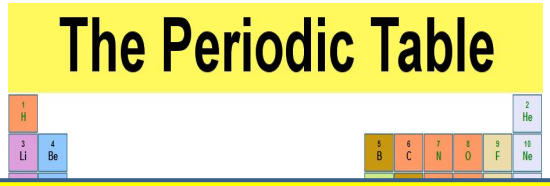
Approx 15% of cost of experiments by humans

charge its batteries.

Boeing wanted to mass produce 3D metal parts for jets…

…but most useful alloys are not printable

Assesses millions of materials

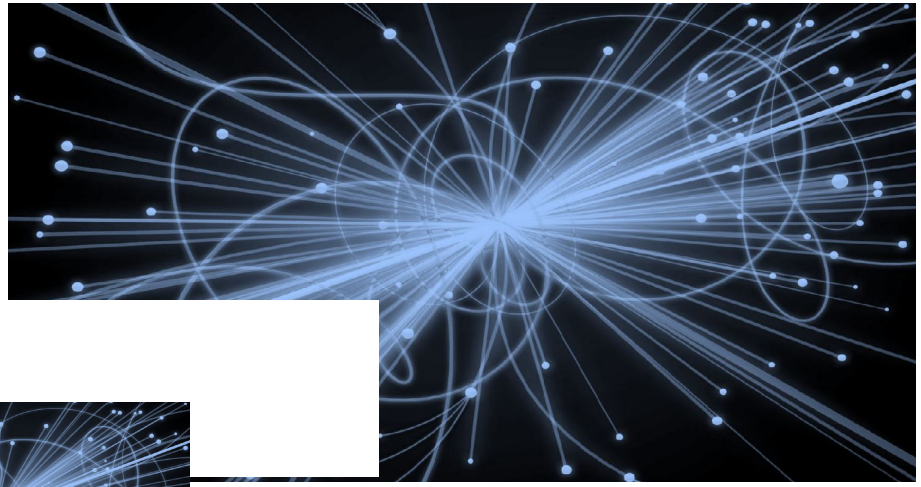


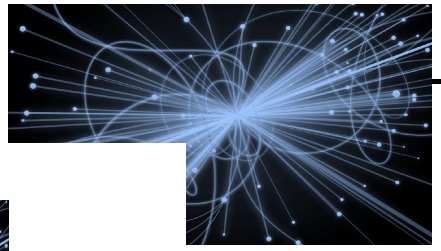
“What would have taken years, it narrowed down to days”

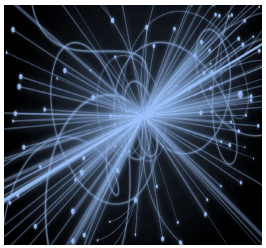
he AI
uggests
ew
etal
alloy for
printing

Even scans pre-digital era research

Intelligent data sampling saves compute $$$







Intelligent research assistants : to save time and money

8 months to +/- weeks

"Our results show that ChatGPT substantially raises average productivity: time taken decreases by 0.8 SDs and output quality rises by 0.4 SDs."

https://economics.mit.edu/sites/default/files/inline-files/Noy_Zhang_1.pdf

- **USD 1.5 billion in 2020 in the US**

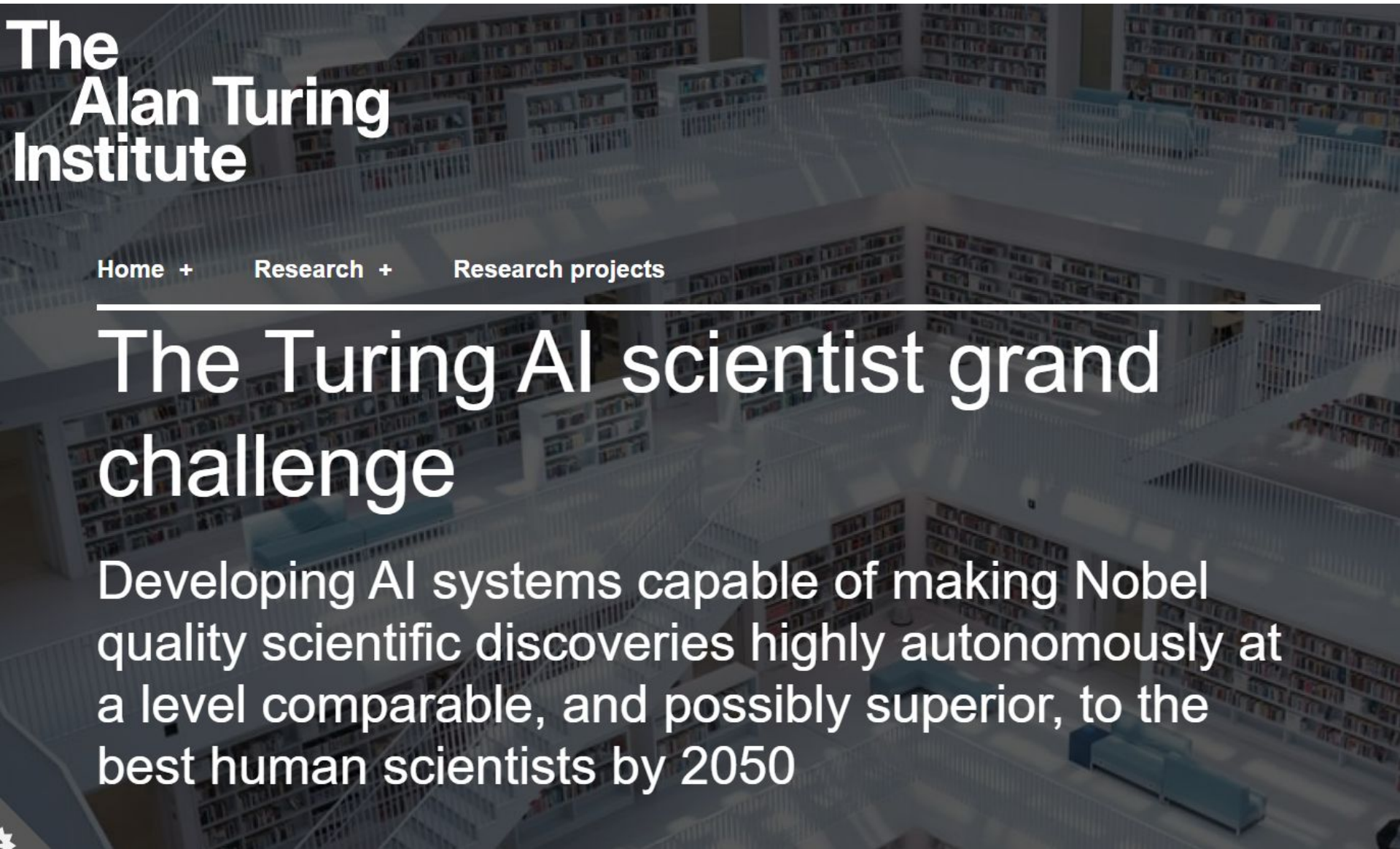
(Aczel, Szaszi and Holcombe, 2021)

Can public policy help ?

Ambitious multi-disciplinary programmes

Multi-disciplinarity

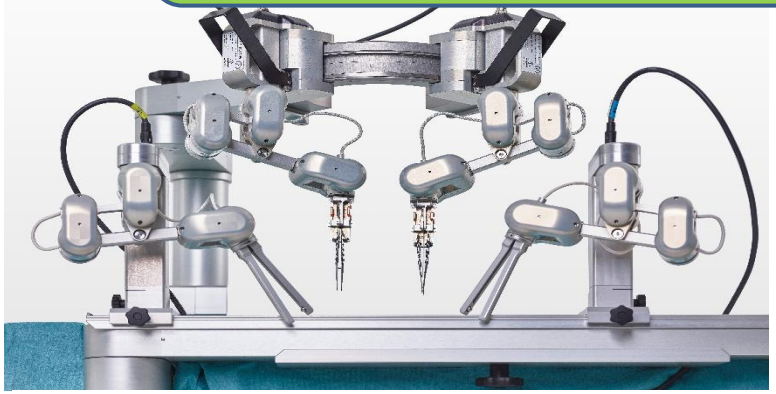
Ambitious multidisciplinary programmes



Less than 6% of all LBD publications can be mapped to at least one SDG

Bring industry, roboticists and domain specialists together

Strengthen data governance



Computational resources

- **National labs, industry and academia could work together to nurture AI ecosystems for tertiary education**

Computational resources

- Na
  aca
  ecc
  edu

Explore pooling resources internationally

A CERN for AI in Europe ?

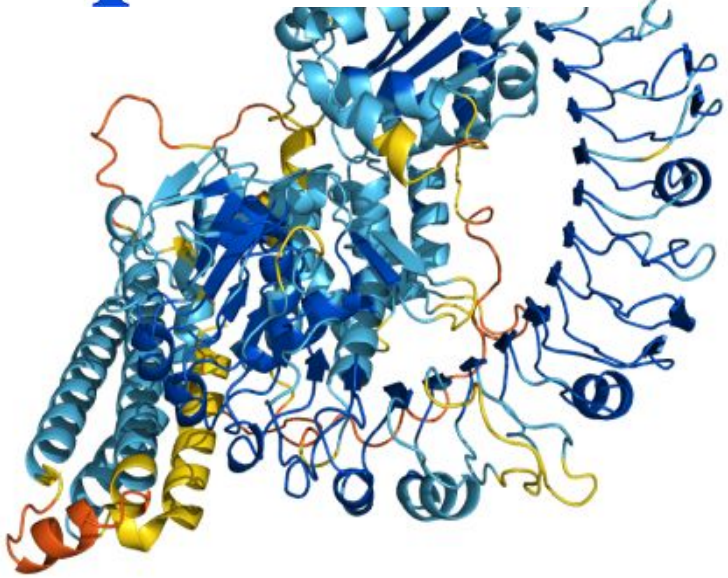
Curricula

- ***Standard bio-science education doesn't address how to search for new hypotheses.***
- ***New PhD programmes based on knowledge synthesis – aided by AI***
- ***Promote research software engineers and engineering***
- ***Raise awareness of stage of development robot systems***

Public R&D can advance the field

Invest in developing new tools for AI in science

AlphaFold



≠

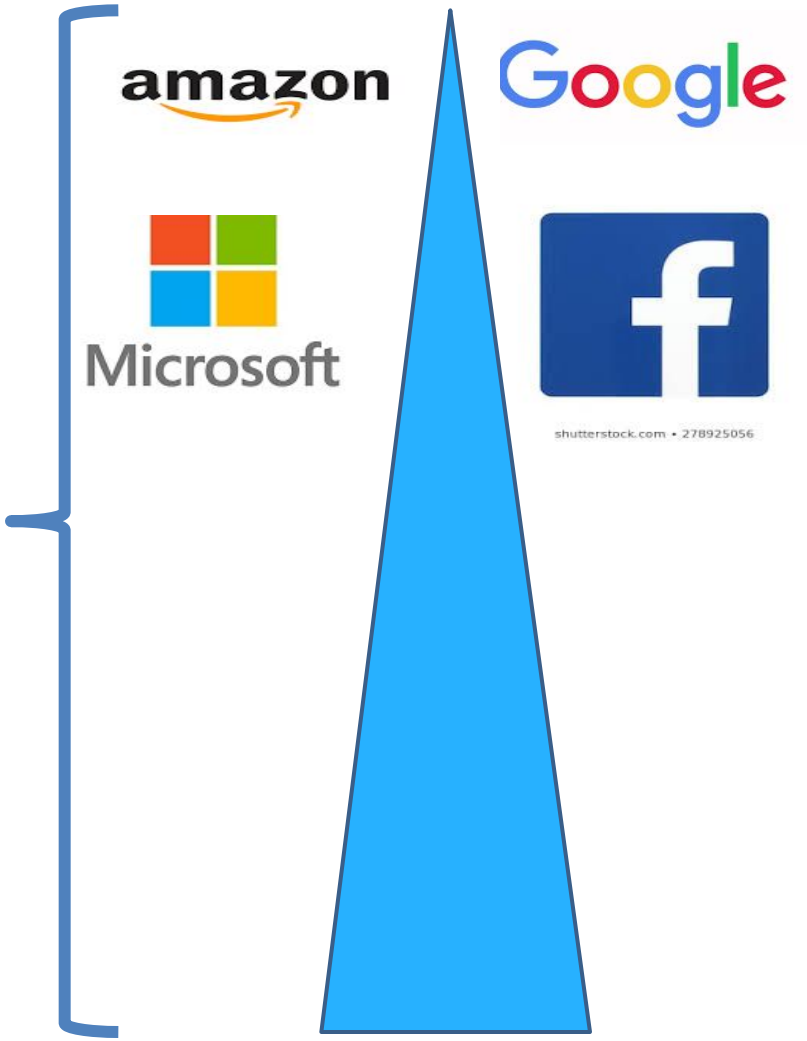
A narrowing of AI research

AI R&D $$$

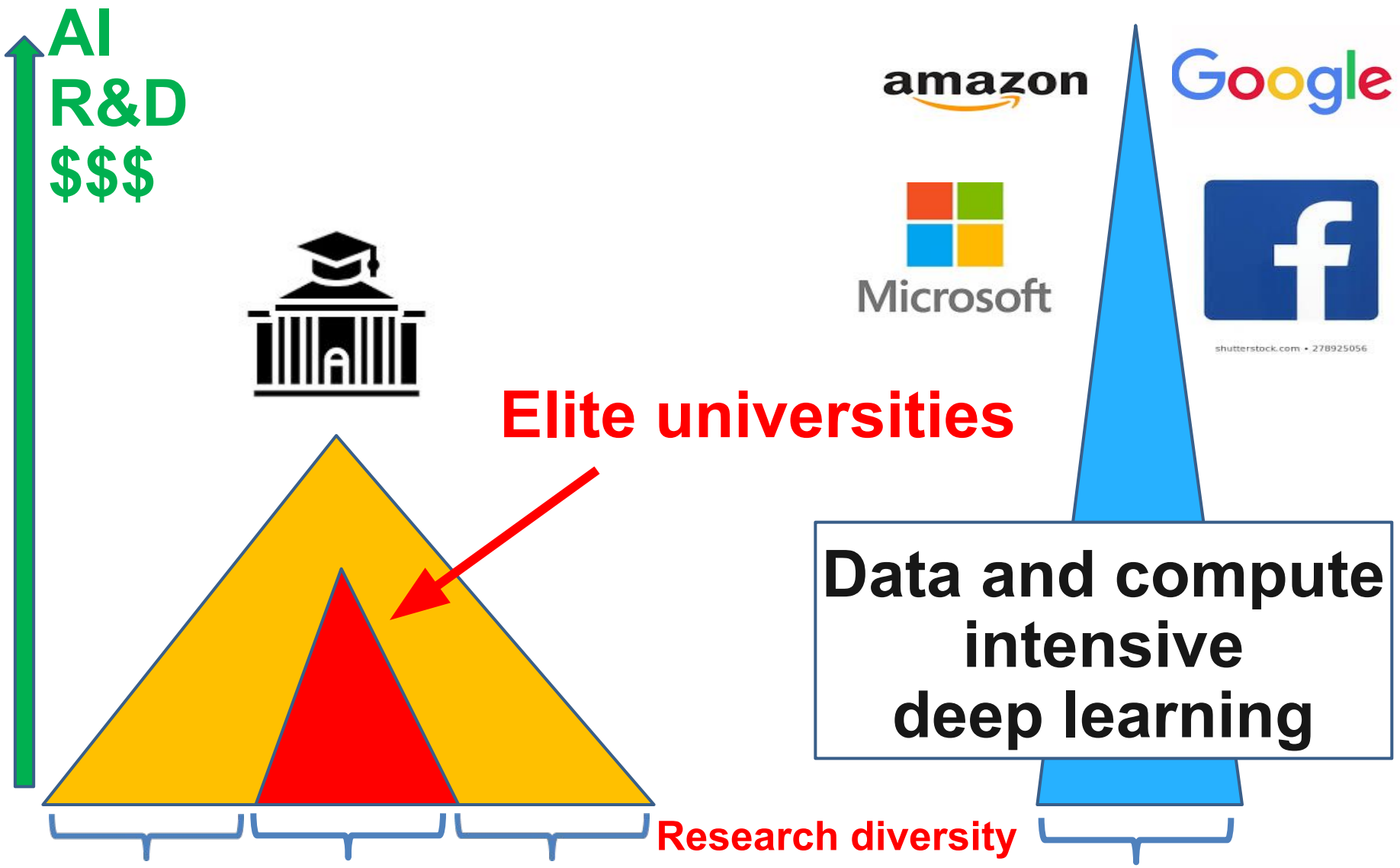
Multiples of public $

Growing faster

21% of AI PhDs went to industry in 2004. 70% in 2020.



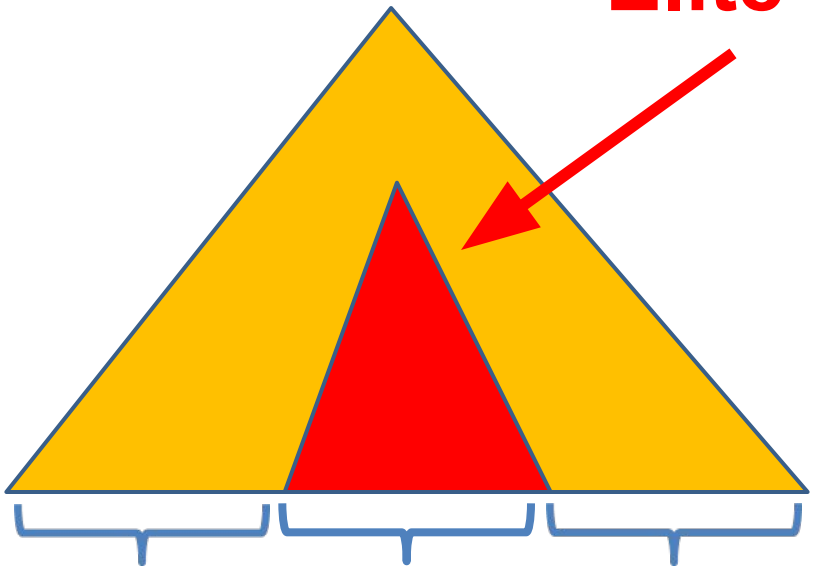
A narrowing of AI research



A narrowing of AI research

AI
R&D
$$$

Elite u



Research diversity

amazon

Google

Industry models are on average 29 times larger than academic models (2021)

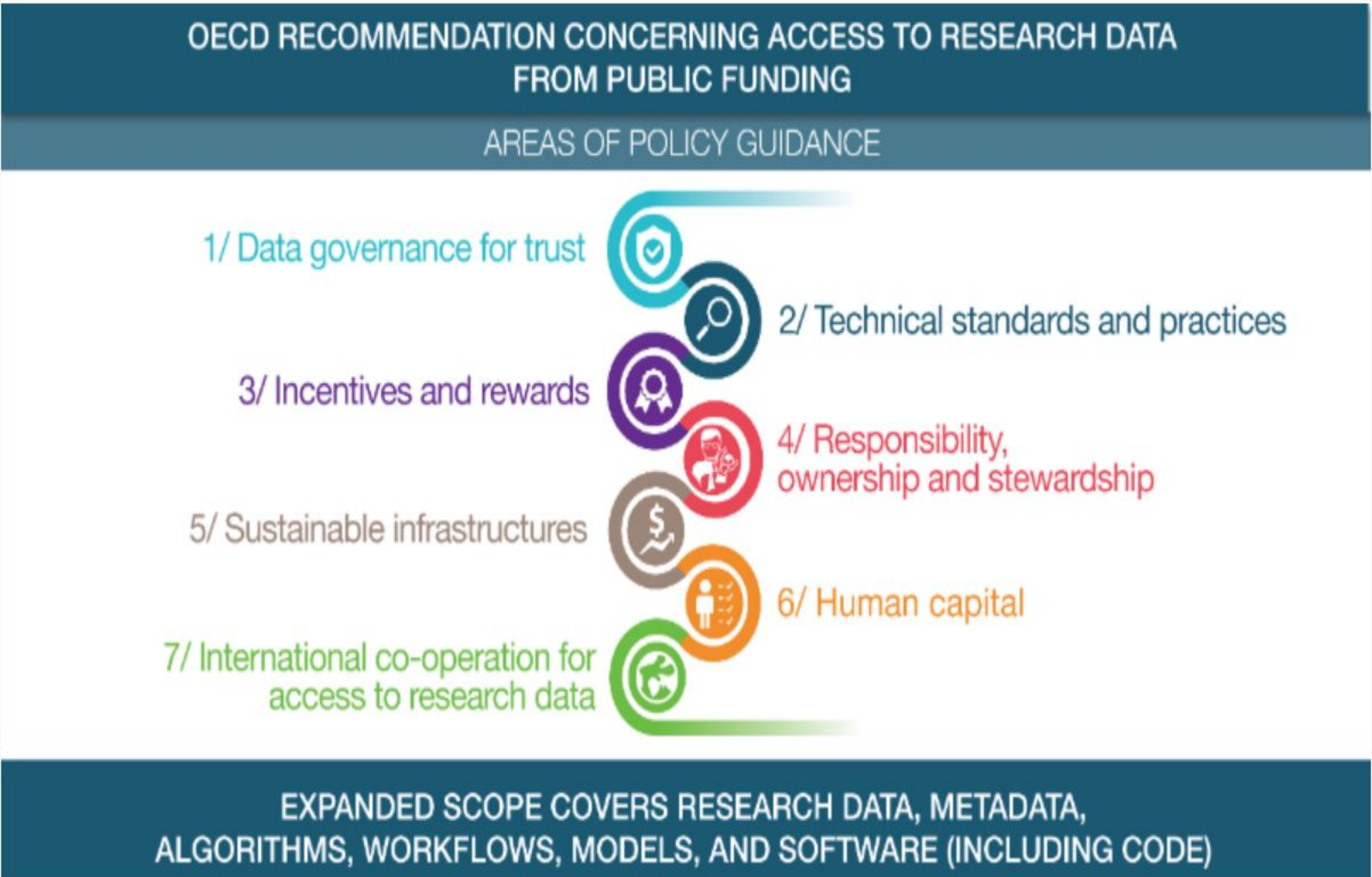
Foster more blue sky thinking

More funding streams and/or publication processes to reward novel methods

Funders could help develop specialised tools to enhance collaborative human AI teams

+ AI

Data


OECD RECOMMENDATION CONCERNING ACCESS TO RESEARCH DATA FROM PUBLIC FUNDING
AREAS OF POLICY GUIDANCE
1/ Data governance for trust
2/ Technical standards and practices
3/ Incentives and rewards
4/ Responsibility, ownership and stewardship
5/ Sustainable infrastructures
6/ Human capital
7/ International co-operation for access to research data
EXPANDED SCOPE COVERS RESEARCH DATA, METADATA, ALGORITHMS, WORKFLOWS, MODELS, AND SOFTWARE (INCLUDING CODE)

And one could target data development strategically

Alpha Fold trained on existing public databases like the Protein Data Bank (PDB).

Only learned what the PDB contained, which tends to smaller, soluble proteins.

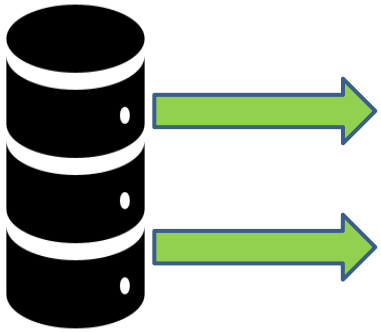
Doesn’t do well on other important proteins.

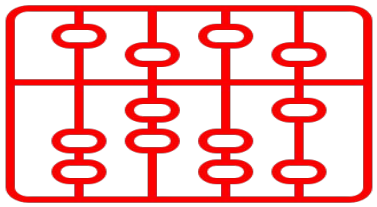
Research governance

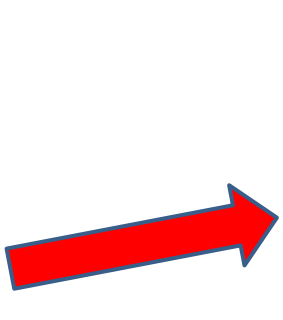
Dangers of dual use AI in drug design

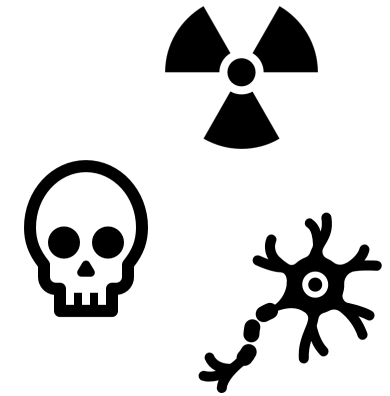
Publicly available data

AI molecule design

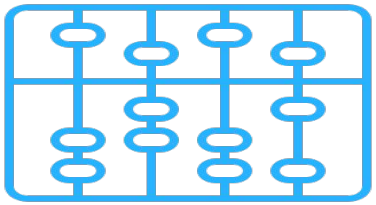


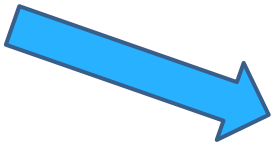






Maximise toxicity





Minimise toxicity

nature

NEWS FEATURE | 06 February 2023 | Correction 08 February 2023

What ChatGPT and generative AI mean for science

Researchers are excited but apprehensive about the latest advances in artificial intelligence.

Forbes

BREAKING

Fake Scientific Abstracts Written By ChatGPT Fooled Scientists, Study Finds

Brian Bushard Forbes Staff
I cover breaking news for Forbes

Follow

Jan 10, 2023, 01:44pm EST

TOPLINE Fake scientific abstracts and research papers generated using OpenAI's highly-advanced chatbox ChatGPT fooled scientists into thinking they were real reports nearly one-third of the time, according to a new study, as the eerily human-like program raises eyebrows over the future of artificial intelligence.

Computer Science > Machine Learning

[Submitted on 3 Oct 2023]

Can large language models provide useful feedback on research papers? A large-scale empirical analysis

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, James Zou

Two parting thoughts

A fast-moving field – much will be new in a year from now.

AI in science may be the most important of all uses of AI.

Thank you

alistair.nolan@oecd.org

BETTER POLICIES FOR BETTER LIVES