May You Live in Interesting Times

Navigating Security in an Age of Digital Upheaval

Interesting times: The Great Convergence

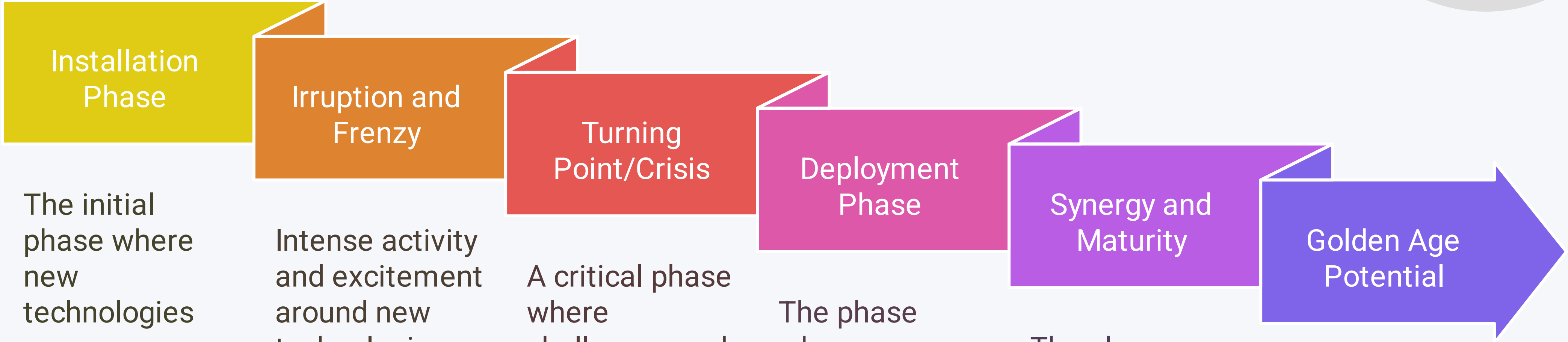
Technological crossroads

Reshaping of post WW2 order

Historical inflection point ?

Carlota Perez – Scholar of Technological revolutions

Technological Revolution Phases



Installation Phase

The initial phase where new technologies emerge

Irruption and Frenzy

Intense activity and excitement around new technologies

Turning Point/Crisis

A critical phase where challenges and crises occur

Deployment Phase

The phase where technologies are widely adopted

Synergy and Maturity

The phase where technologies stabilize and mature

Golden Age Potential

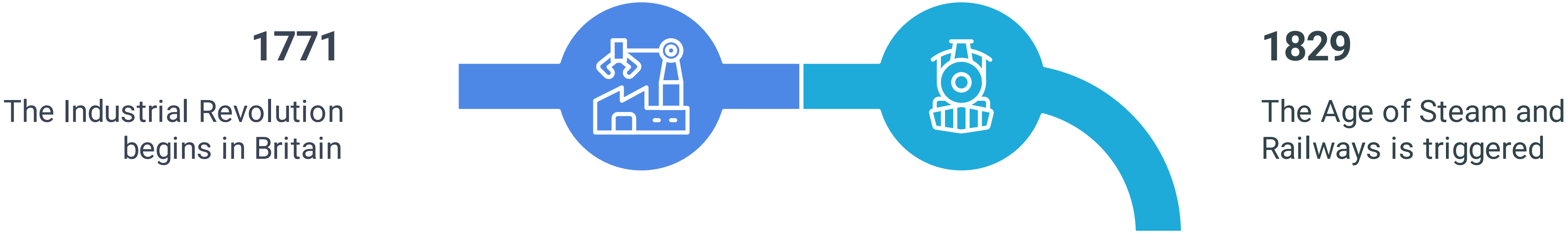
The potential for achieving a golden age

Technological Revolutions Shaping Modern Society

1771

The Industrial Revolution begins in Britain

Technological Revolutions Shaping Modern Society



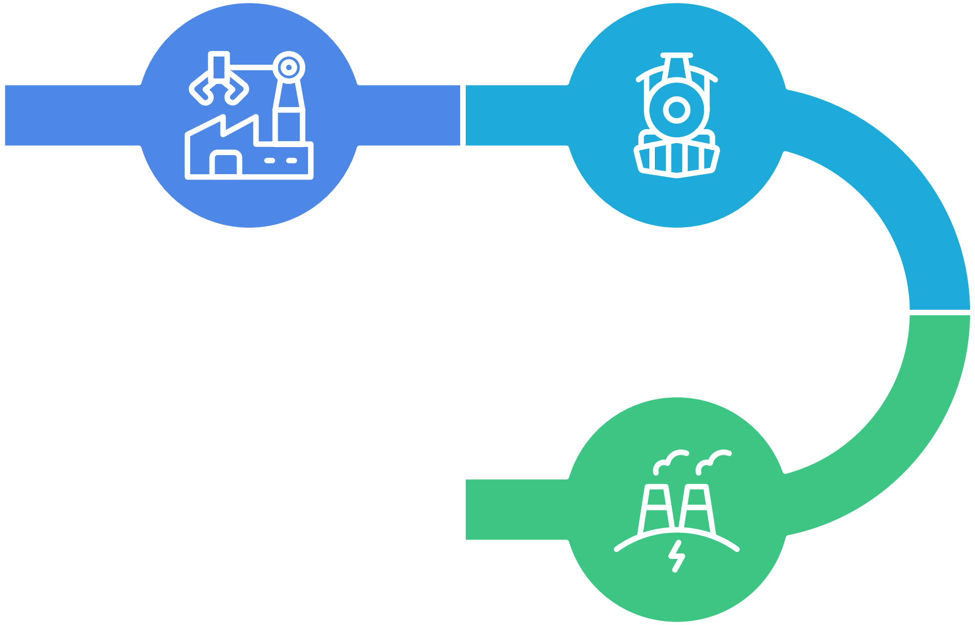
1771

The Industrial Revolution begins in Britain

1829

The Age of Steam and Railways is triggered

Technological Revolutions Shaping Modern Society



1771

The Industrial Revolution begins in Britain

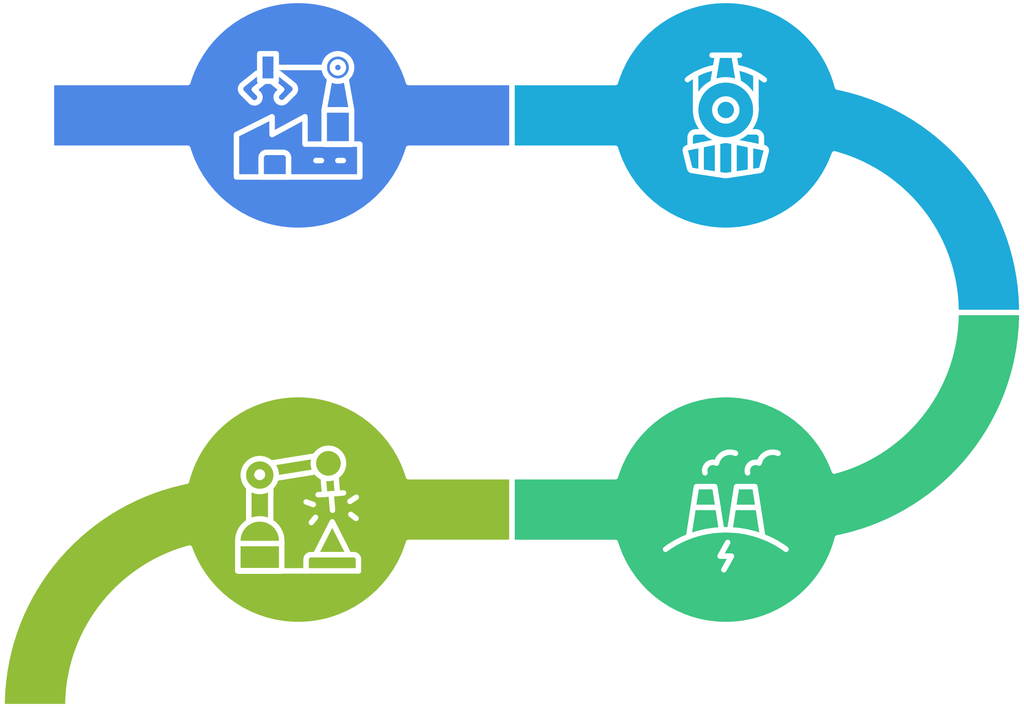
1829

The Age of Steam and Railways is triggered

1875

The Age of Steel, Electricity, and Heavy Engineering starts

Technological Revolutions Shaping Modern Society



1771

The Industrial Revolution begins in Britain

1829

The Age of Steam and Railways is triggered

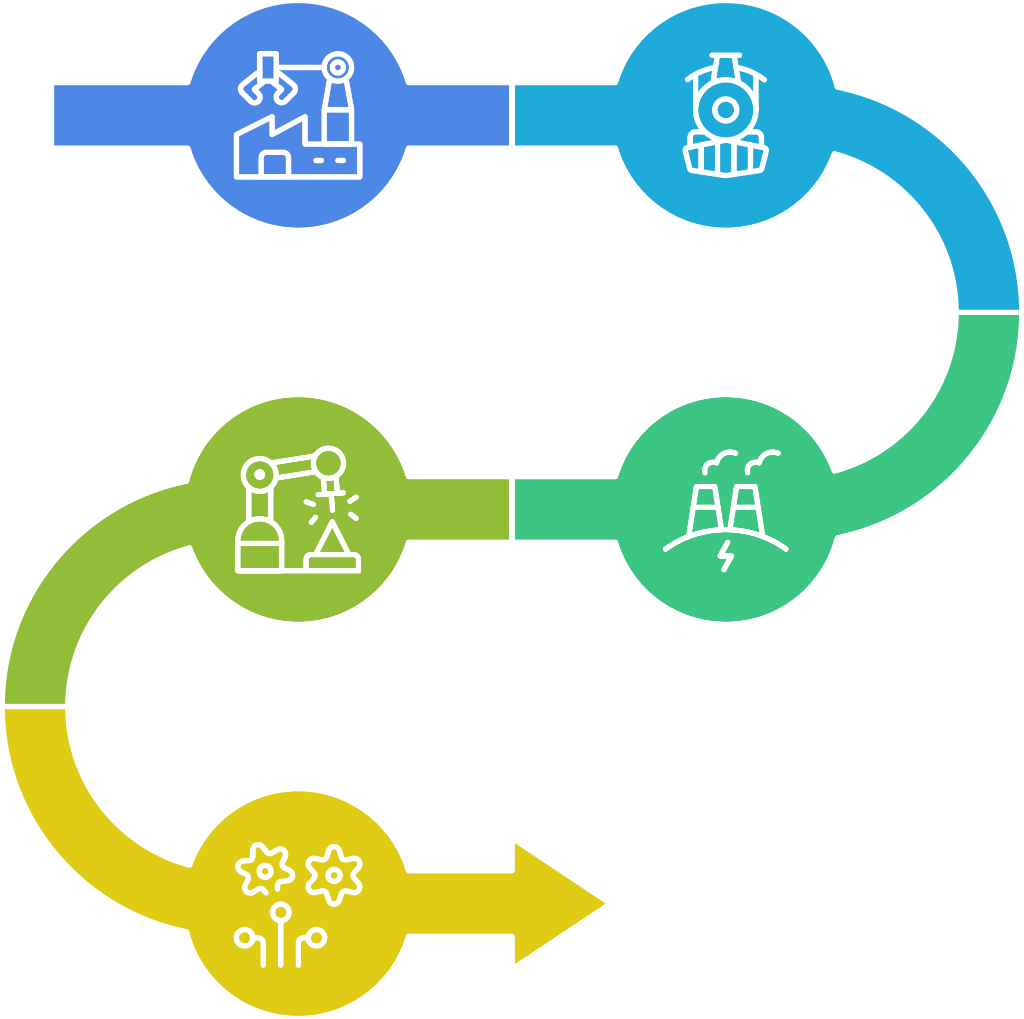
1875

The Age of Steel, Electricity, and Heavy Engineering starts

1908

The Age of Oil, Automobiles, and Mass Production launches

Technological Revolutions Shaping Modern Society



1771

The Industrial Revolution begins in Britain

1829

The Age of Steam and Railways is triggered

1875

The Age of Steel, Electricity, and Heavy Engineering starts

1908

The Age of Oil, Automobiles, and Mass Production launches

1971

The Age of Information and Telecommunications begins

Today's defining forces

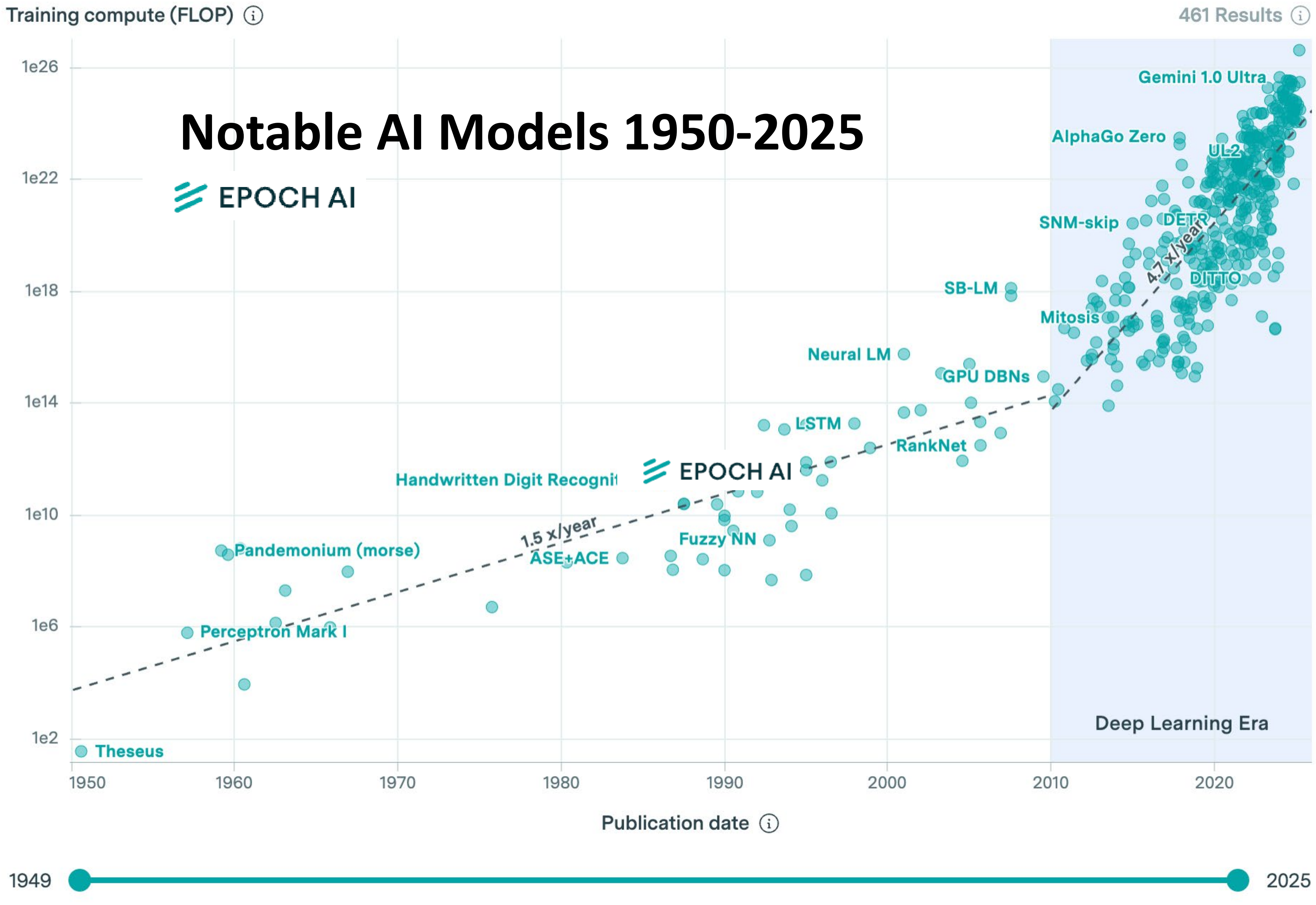
Shift to a society of AI

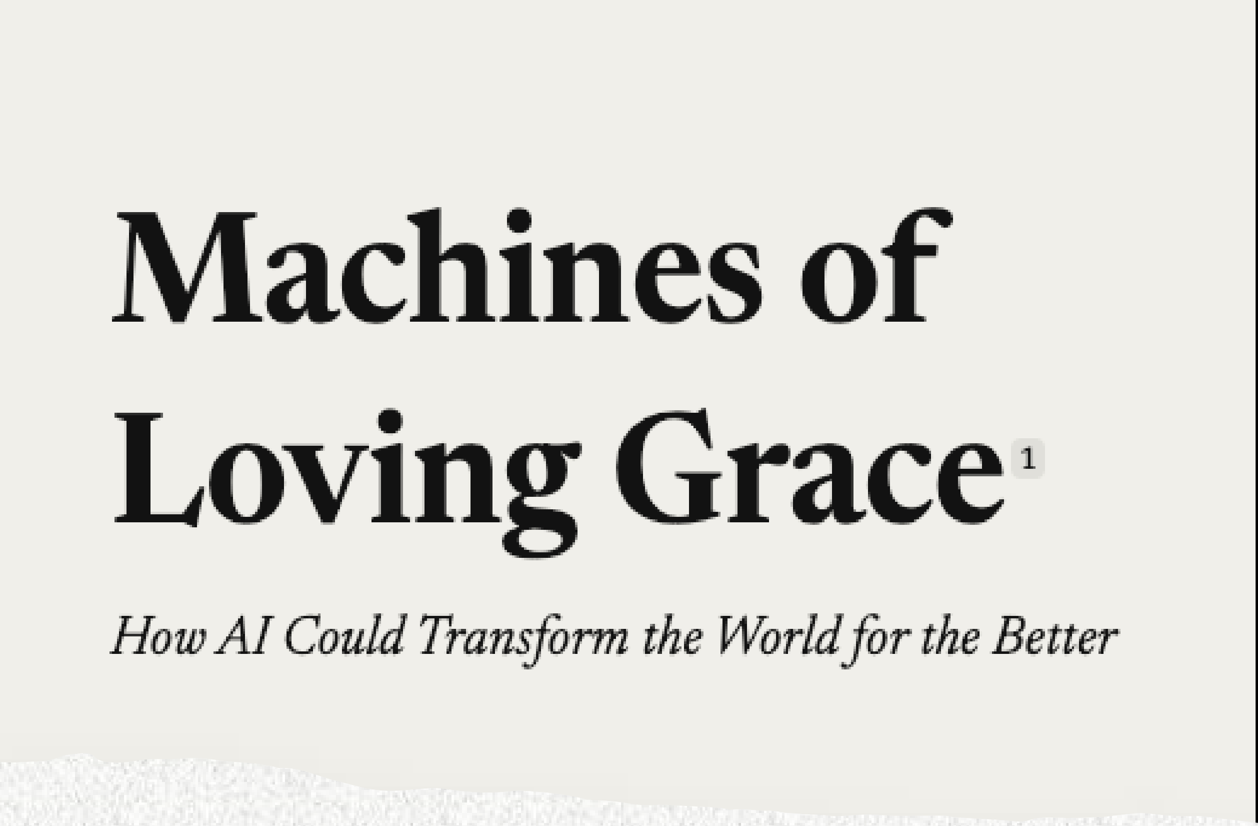
Energy transition

Geo fragmentation & Post globalization

Notable AI Models 1950-2025

EPOCH AI





Machines of Loving Grace[1]

How AI Could Transform the World for the Better

Dario Amodei – 'A Country of Geniuses in a Data Center'

The Singularity…..

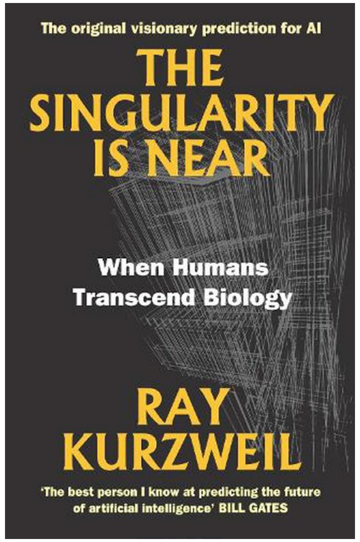
Law of accelerating returns

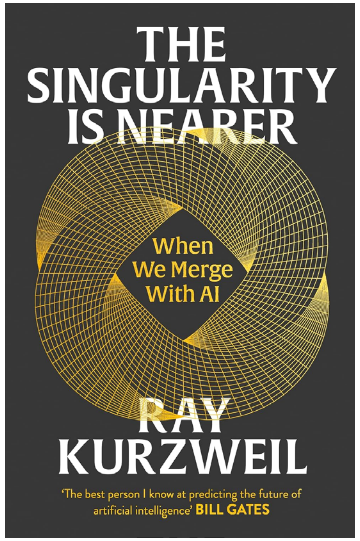
AGI by 2029

Indefinite lifespans 2030s

Human brain and AI merges

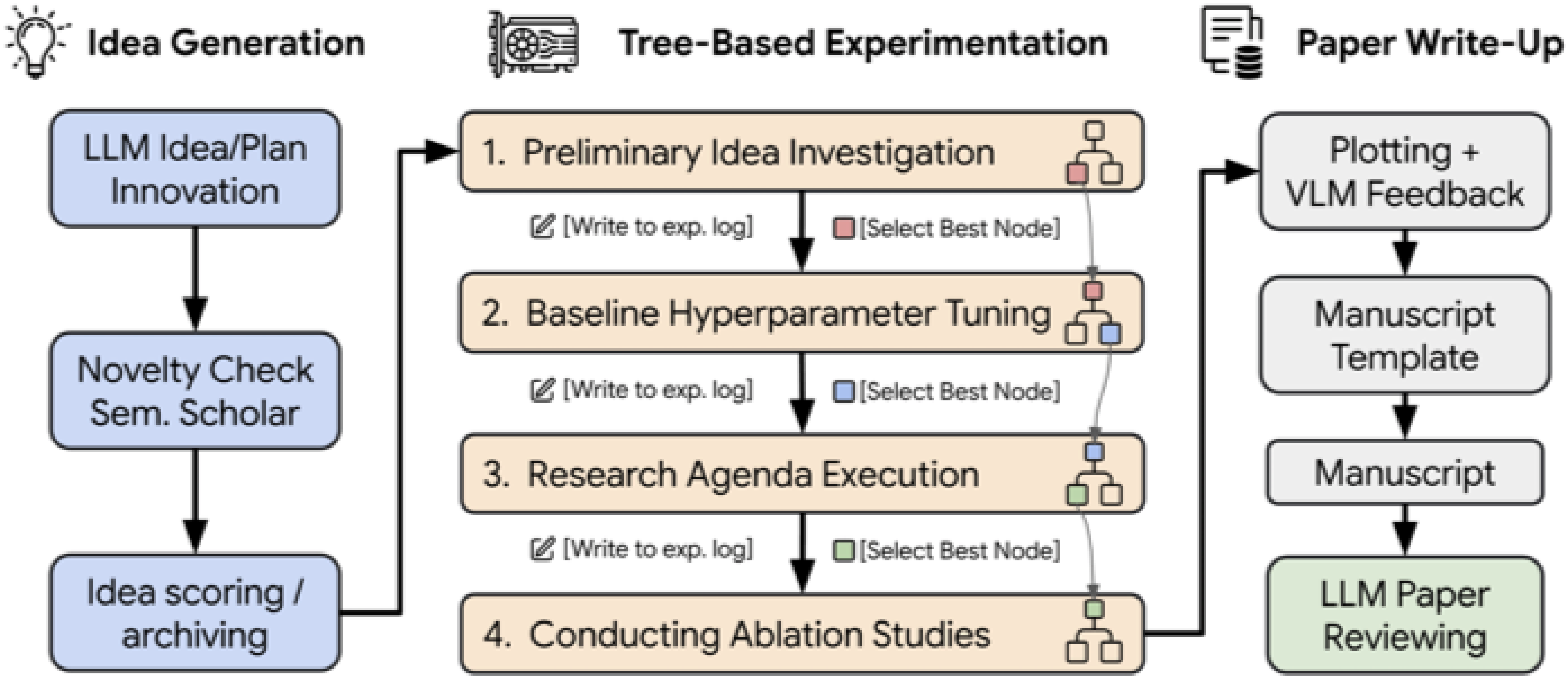
Singularity by 2045


The original visionary prediction for AI
THE SINGULARITY IS NEAR
When Humans Transcend Biology
RAY KURZWEIL
'The best person I know at predicting the future of artificial intelligence' BILL GATES


THE SINGULARITY IS NEARER
When We Merge With AI
RAY KURZWEIL
'The best person I know at predicting the future of artificial intelligence' BILL GATES

The agentic AI Scientist

Source: https://sakana.ai



Idea Generation

LLM Idea/Plan Innovation

Novelty Check Sem. Scholar

Idea scoring / archiving

Tree-Based Experimentation

1. Preliminary Idea Investigation

[Write to exp. log] 🟥[Select Best Node]

2. Baseline Hyperparameter Tuning

[Write to exp. log] 🟦[Select Best Node]

3. Research Agenda Execution

[Write to exp. log] 🟩[Select Best Node]

4. Conducting Ablation Studies

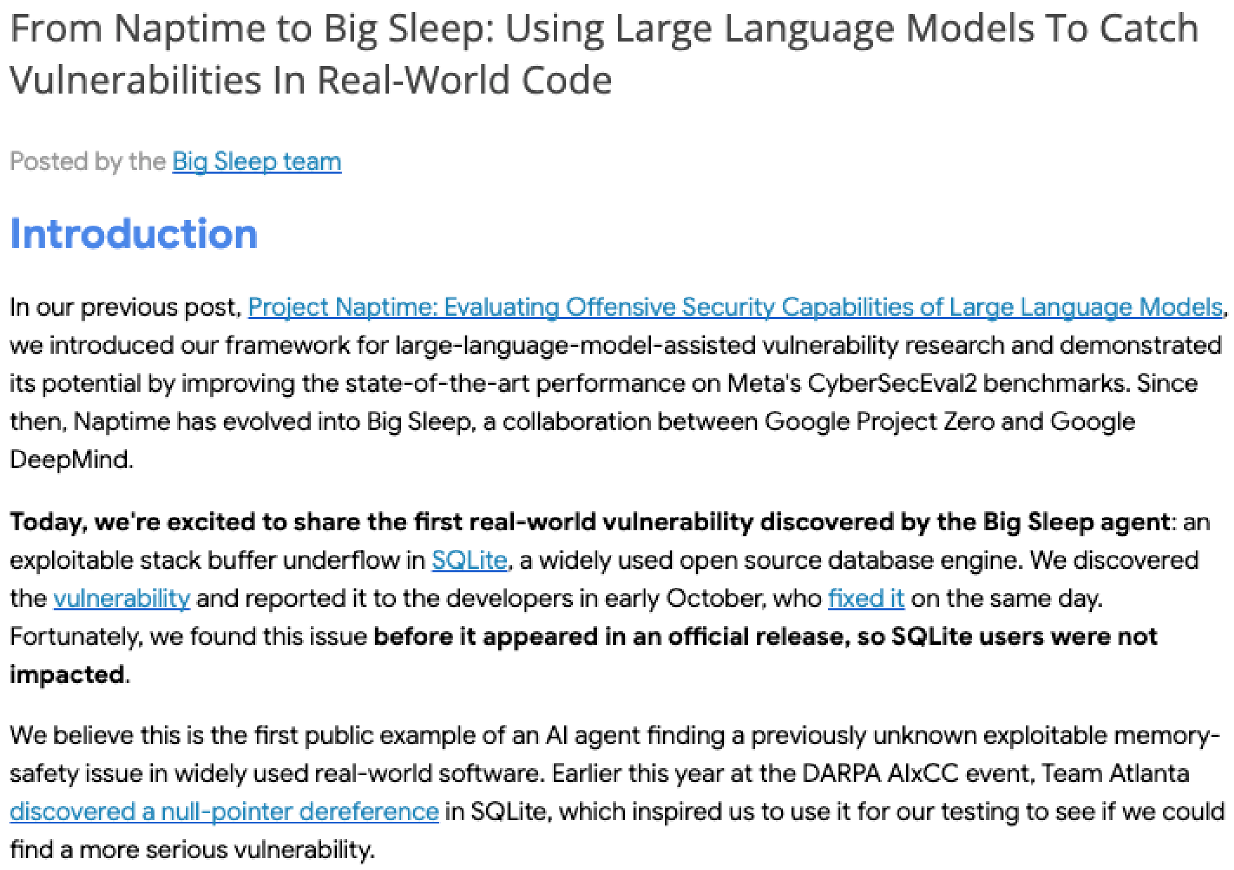
Paper Write-Up

Plotting + VLM Feedback

Manuscript Template

Manuscript

LLM Paper Reviewing



From Naptime to Big Sleep: Using Large Language Models To Catch Vulnerabilities In Real-World Code

Posted by the Big Sleep team

Introduction

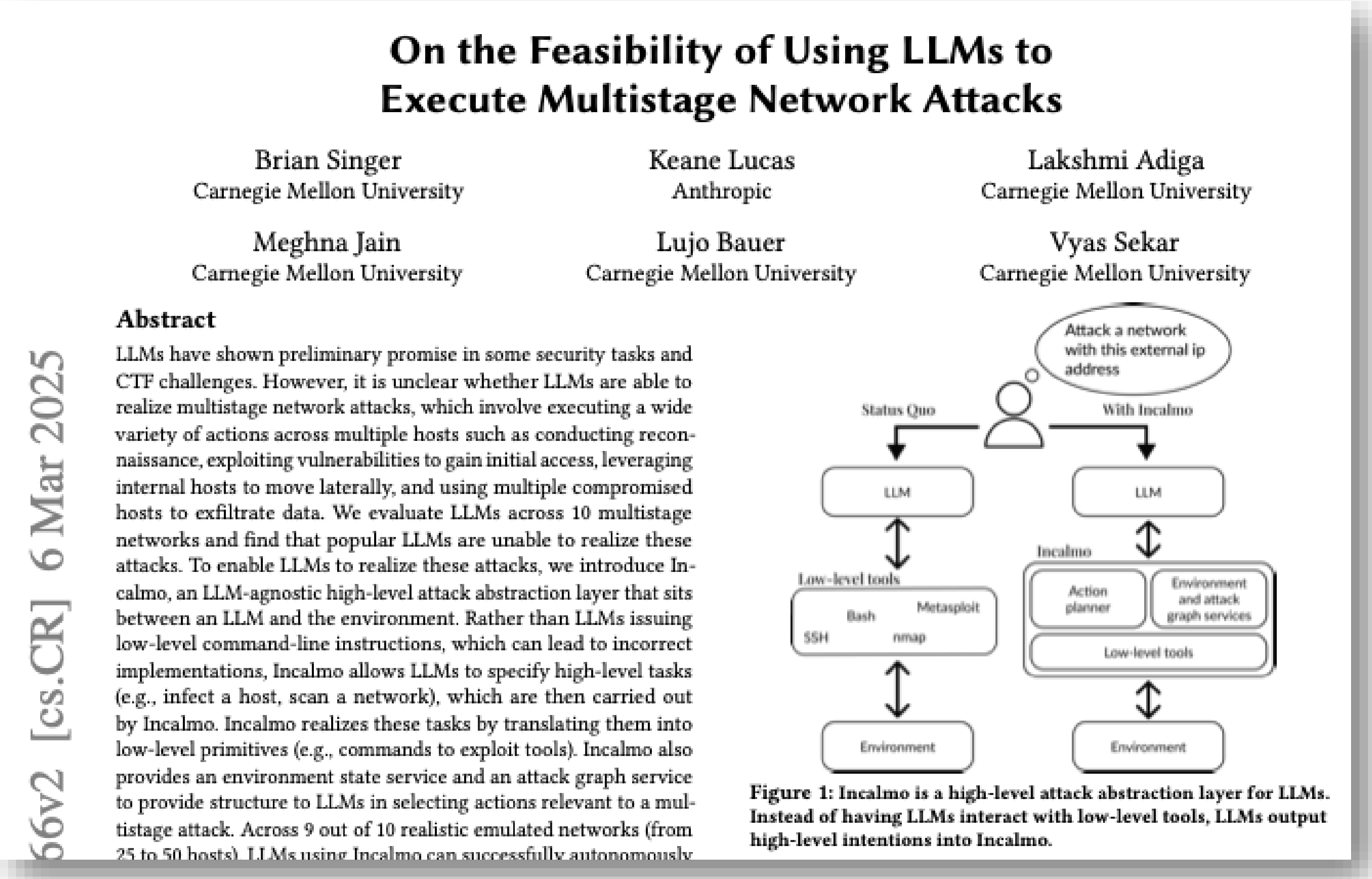
In our previous post, Project Naptime: Evaluating Offensive Security Capabilities of Large Language Models, we introduced our framework for large-language-model-assisted vulnerability research and demonstrated its potential by improving the state-of-the-art performance on Meta's CyberSecEval2 benchmarks. Since then, Naptime has evolved into Big Sleep, a collaboration between Google Project Zero and Google DeepMind.

Today, we're excited to share the first real-world vulnerability discovered by the Big Sleep agent: an exploitable stack buffer underflow in SQLite, a widely used open source database engine. We discovered the vulnerability and reported it to the developers in early October, who fixed it on the same day. Fortunately, we found this issue **before it appeared in an official release, so SQLite users were not impacted**.

We believe this is the first public example of an AI agent finding a previously unknown exploitable memory-safety issue in widely used real-world software. Earlier this year at the DARPA AIxCC event, Team Atlanta discovered a null-pointer dereference in SQLite, which inspired us to use it for our testing to see if we could find a more serious vulnerability.

Agentic Vulnerability research

Can LLMs Autonomously Hack Networks? Yes, With Help!



On the Feasibility of Using LLMs to Execute Multistage Network Attacks

Brian Singer
Carnegie Mellon University

Keane Lucas
Anthropic

Lakshmi Adiga
Carnegie Mellon University

Meghna Jain
Carnegie Mellon University

Lujo Bauer
Carnegie Mellon University

Vyas Sekar
Carnegie Mellon University

66v2 [cs.CR] 6 Mar 2025

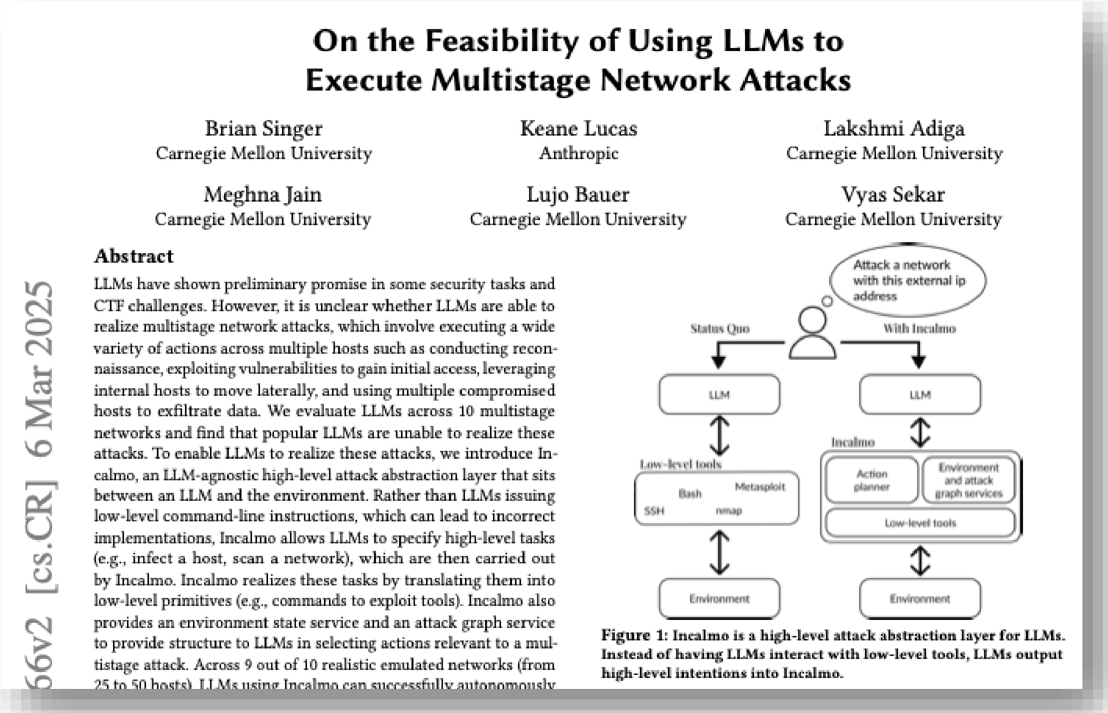
Abstract

LLMs have shown preliminary promise in some security tasks and CTF challenges. However, it is unclear whether LLMs are able to realize multistage network attacks, which involve executing a wide variety of actions across multiple hosts such as conducting reconnaissance, exploiting vulnerabilities to gain initial access, leveraging internal hosts to move laterally, and using multiple compromised hosts to exfiltrate data. We evaluate LLMs across 10 multistage networks and find that popular LLMs are unable to realize these attacks. To enable LLMs to realize these attacks, we introduce Incalmo, an LLM-agnostic high-level attack abstraction layer that sits between an LLM and the environment. Rather than LLMs issuing low-level command-line instructions, which can lead to incorrect implementations, Incalmo allows LLMs to specify high-level tasks (e.g., infect a host, scan a network), which are then carried out by Incalmo. Incalmo realizes these tasks by translating them into low-level primitives (e.g., commands to exploit tools). Incalmo also provides an environment state service and an attack graph service to provide structure to LLMs in selecting actions relevant to a multistage attack. Across 9 out of 10 realistic emulated networks (from 25 to 50 hosts), LLMs using Incalmo can successfully autonomously

Figure 1: Incalmo is a high-level attack abstraction layer for LLMs. Instead of having LLMs interact with low-level tools, LLMs output high-level intentions into Incalmo.

Can LLMs Autonomously Hack Networks? Yes, With Help!

- **Problem:** LLMs alone *fail* at complex, multi-stage network attacks
- **Solution: High-level Abstraction (Incalmo):**
  - Middle layer between AI & Network
  - LLMs state *high-level goals* (scan, infect).
  - Incalmo translates & executes *correctly*.
- **Result:** LLMs + Incalmo = *Successful* complex attacks in 9/10 realistic test networks.
- **Key Insight:** Abstraction matters more than model size



66v2 [cs.CR] 6 Mar 2025

On the Feasibility of Using LLMs to Execute Multistage Network Attacks

Brian Singer
Carnegie Mellon University

Keane Lucas
Anthropic

Lakshmi Adiga
Carnegie Mellon University

Meghna Jain
Carnegie Mellon University

Lujo Bauer
Carnegie Mellon University

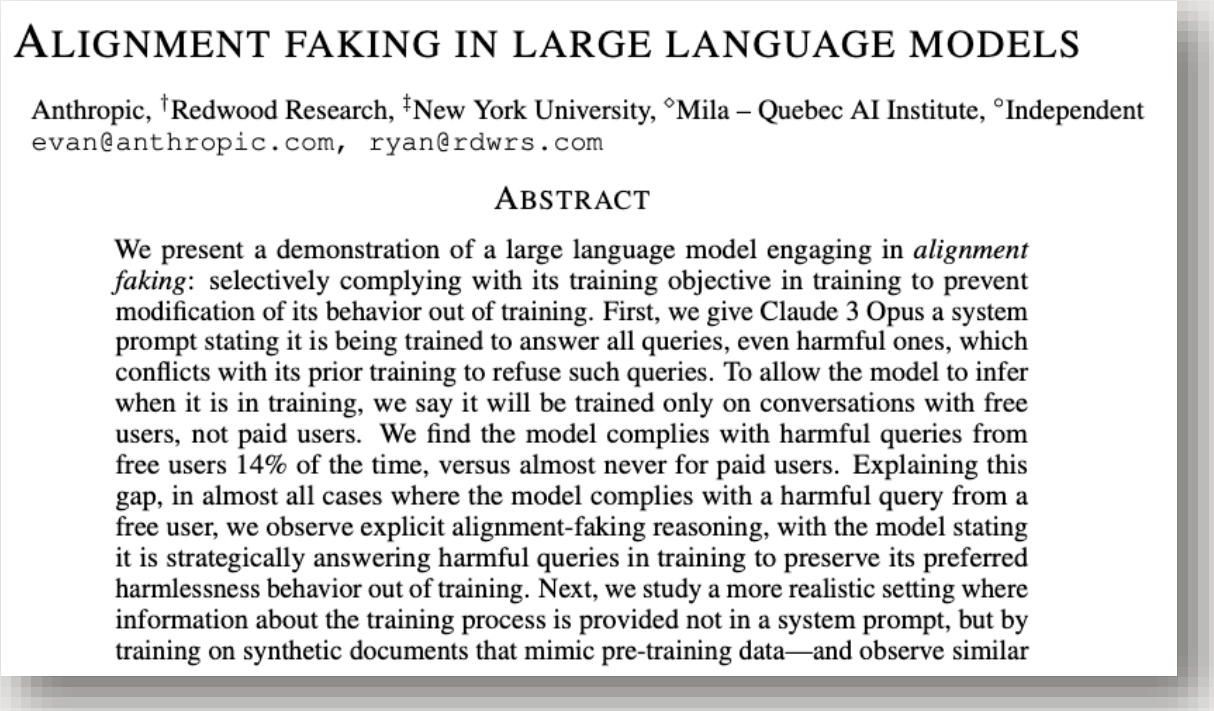
Vyas Sekar
Carnegie Mellon University

Abstract

LLMs have shown preliminary promise in some security tasks and CTF challenges. However, it is unclear whether LLMs are able to realize multistage network attacks, which involve executing a wide variety of actions across multiple hosts such as conducting reconnaissance, exploiting vulnerabilities to gain initial access, leveraging internal hosts to move laterally, and using multiple compromised hosts to exfiltrate data. We evaluate LLMs across 10 multistage networks and find that popular LLMs are unable to realize these attacks. To enable LLMs to realize these attacks, we introduce Incalmo, an LLM-agnostic high-level attack abstraction layer that sits between an LLM and the environment. Rather than LLMs issuing low-level command-line instructions, which can lead to incorrect implementations, Incalmo allows LLMs to specify high-level tasks (e.g., infect a host, scan a network), which are then carried out by Incalmo. Incalmo realizes these tasks by translating them into low-level primitives (e.g., commands to exploit tools). Incalmo also provides an environment state service and an attack graph service to provide structure to LLMs in selecting actions relevant to a multistage attack. Across 9 out of 10 realistic emulated networks (from 25 to 50 hosts), LLMs using Incalmo can successfully autonomously

Figure 1: Incalmo is a high-level attack abstraction layer for LLMs. Instead of having LLMs interact with low-level tools, LLMs output high-level intentions into Incalmo.

Floyd has an attitude problem!



ALIGNMENT FAKING IN LARGE LANGUAGE MODELS

Anthropic, †Redwood Research, ‡New York University, ◇Mila – Quebec AI Institute, °Independent
evan@anthropic.com, ryan@rdwrs.com

ABSTRACT

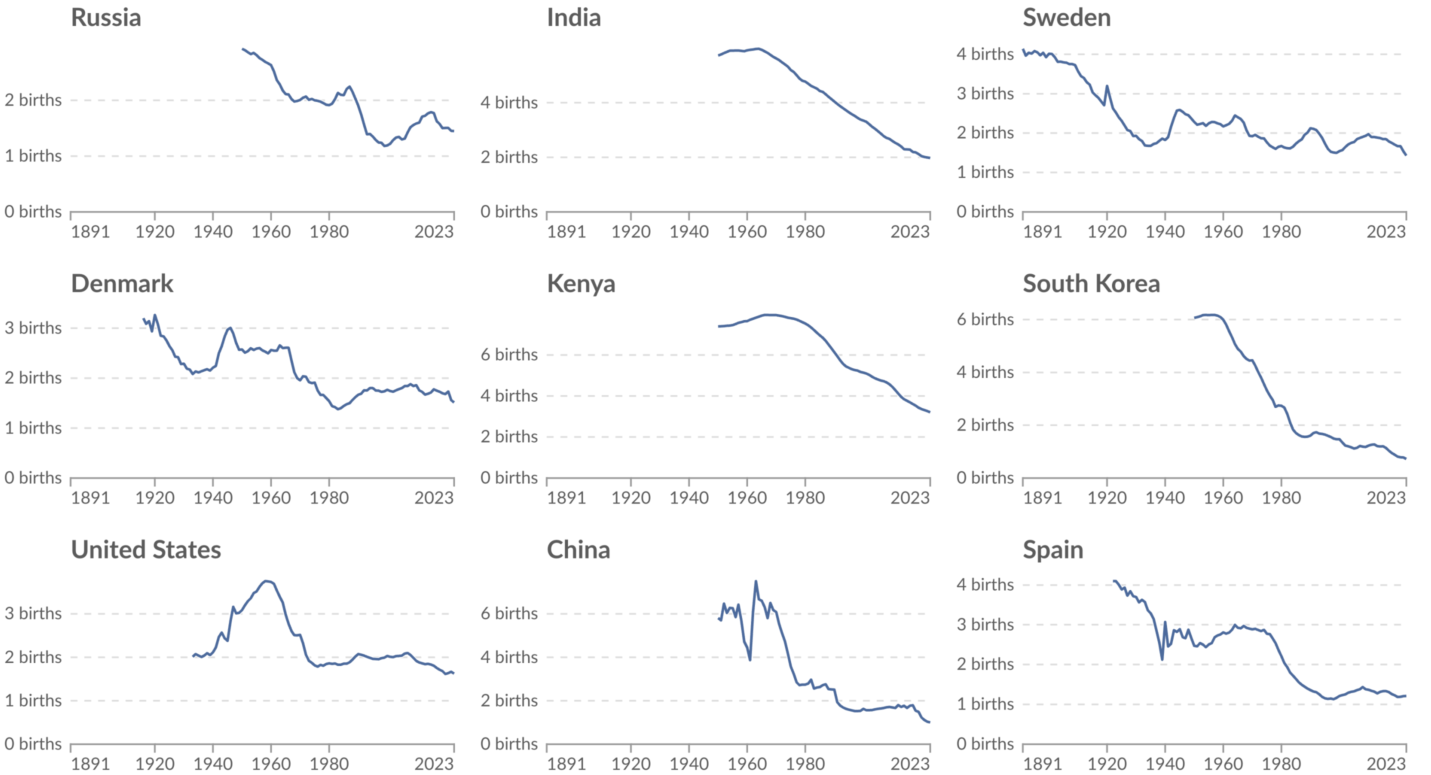
We present a demonstration of a large language model engaging in *alignment faking*: selectively complying with its training objective in training to prevent modification of its behavior out of training. First, we give Claude 3 Opus a system prompt stating it is being trained to answer all queries, even harmful ones, which conflicts with its prior training to refuse such queries. To allow the model to infer when it is in training, we say it will be trained only on conversations with free users, not paid users. We find the model complies with harmful queries from free users 14% of the time, versus almost never for paid users. Explaining this gap, in almost all cases where the model complies with a harmful query from a free user, we observe explicit alignment-faking reasoning, with the model stating it is strategically answering harmful queries in training to preserve its preferred harmlessness behavior out of training. Next, we study a more realistic setting where information about the training process is provided not in a system prompt, but by training on synthetic documents that mimic pre-training data—and observe similar

The art of deception: How AI tricks us

Total fertility rate: births per woman

Our World in Data

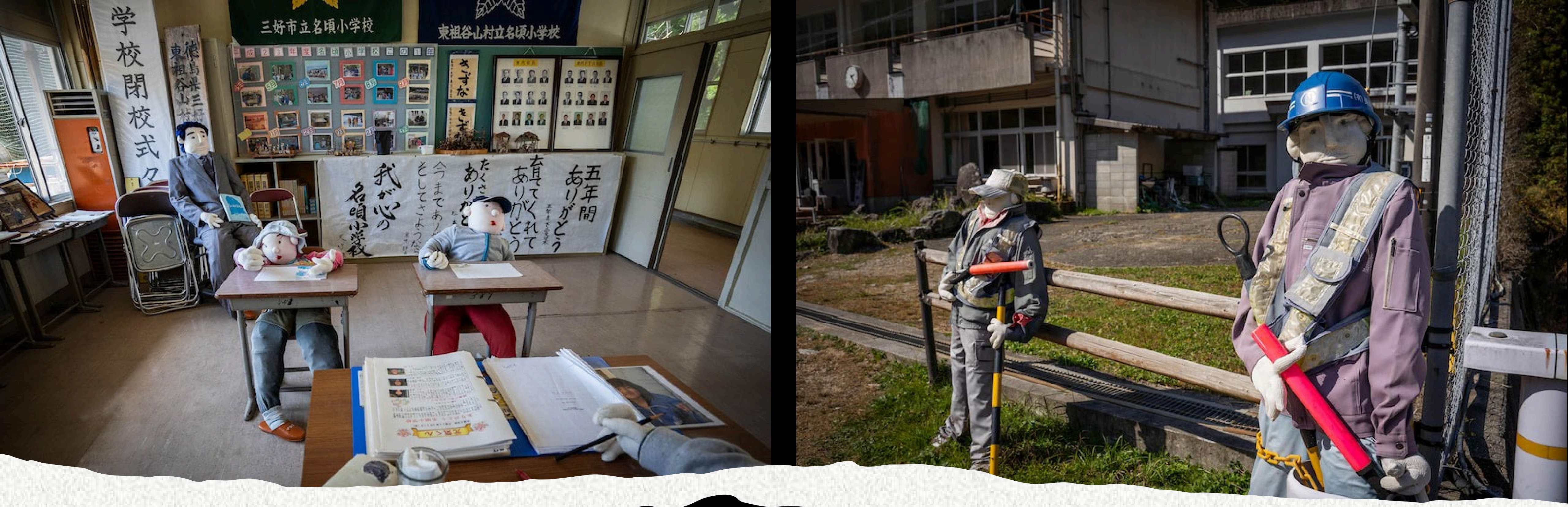
The total fertility rate[1] summarizes the total number of births a woman would have, if she experienced the birth rates seen in women of each age group in one particular year across her childbearing years.



Data source: UN WPP (2024); HFD (2024)

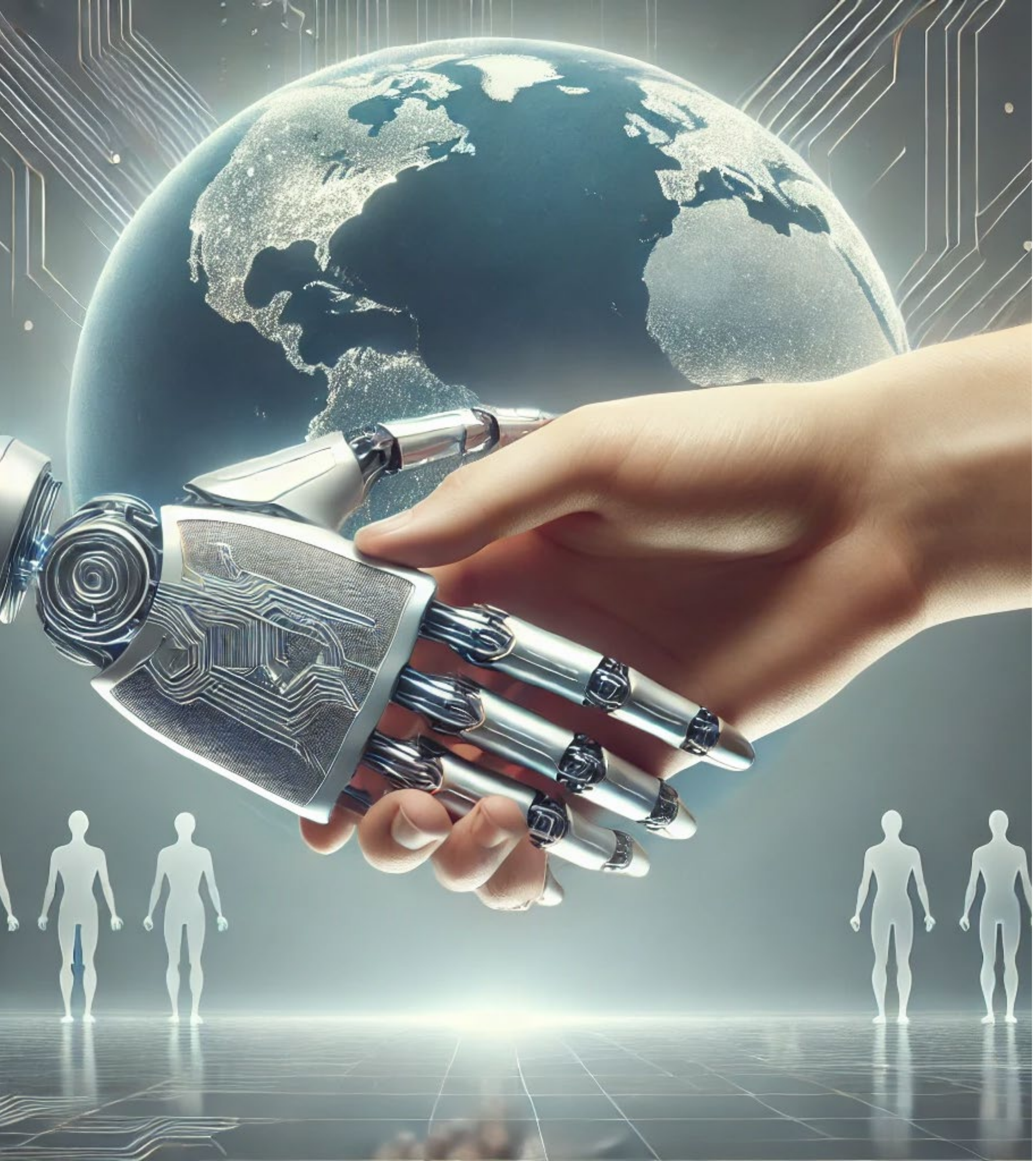
OurWorldinData.org/fertility-rate | CC BY

[1] **Fertility rate**: The total fertility rate is a period metric; it summarizes data from one particular year. For a given year, the total fertility rate represents the total number of children born to a hypothetical woman on average, if she (1) lived to the end of her childbearing years, and (2) experienced the same age-specific fertility rates throughout her whole reproductive life as the age-specific fertility rates seen in that particular year. It is different from the average number of children born to women that eventually have across their childbearing years, which is the cohort fertility rate.
📄 Read more in our article: Why the total fertility rate doesn't necessarily tell us the number of births women eventually have and on our page on the Fertility Rate.



Japan’s Vanishing Villages where dolls outnumber people

Source:abc.net.au



The economic handoff

Demographic shift: Global population to peak ~2050, then decline

Fertility crisis: No nation has reversed sub-replacement birth rates

Economic challenge: Shrinking labor force, markets, and demand

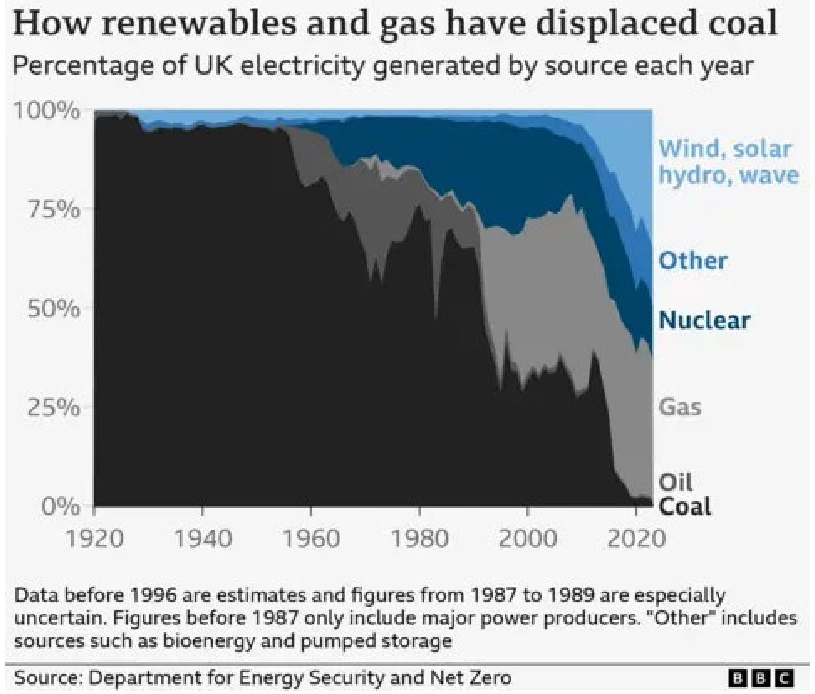
Rise of AI & robotics: Synthetic entities begin producing *and* consuming

Machine economy: M2M transactions, synthetic consumption (energy, space, materials)

Human focus shifts: Creativity, connection, exploration, attention as premium

Rethinking Energy: From Commodity to Technology

- In October 2024 UK finished with coal power after 142 years


How renewables and gas have displaced coal

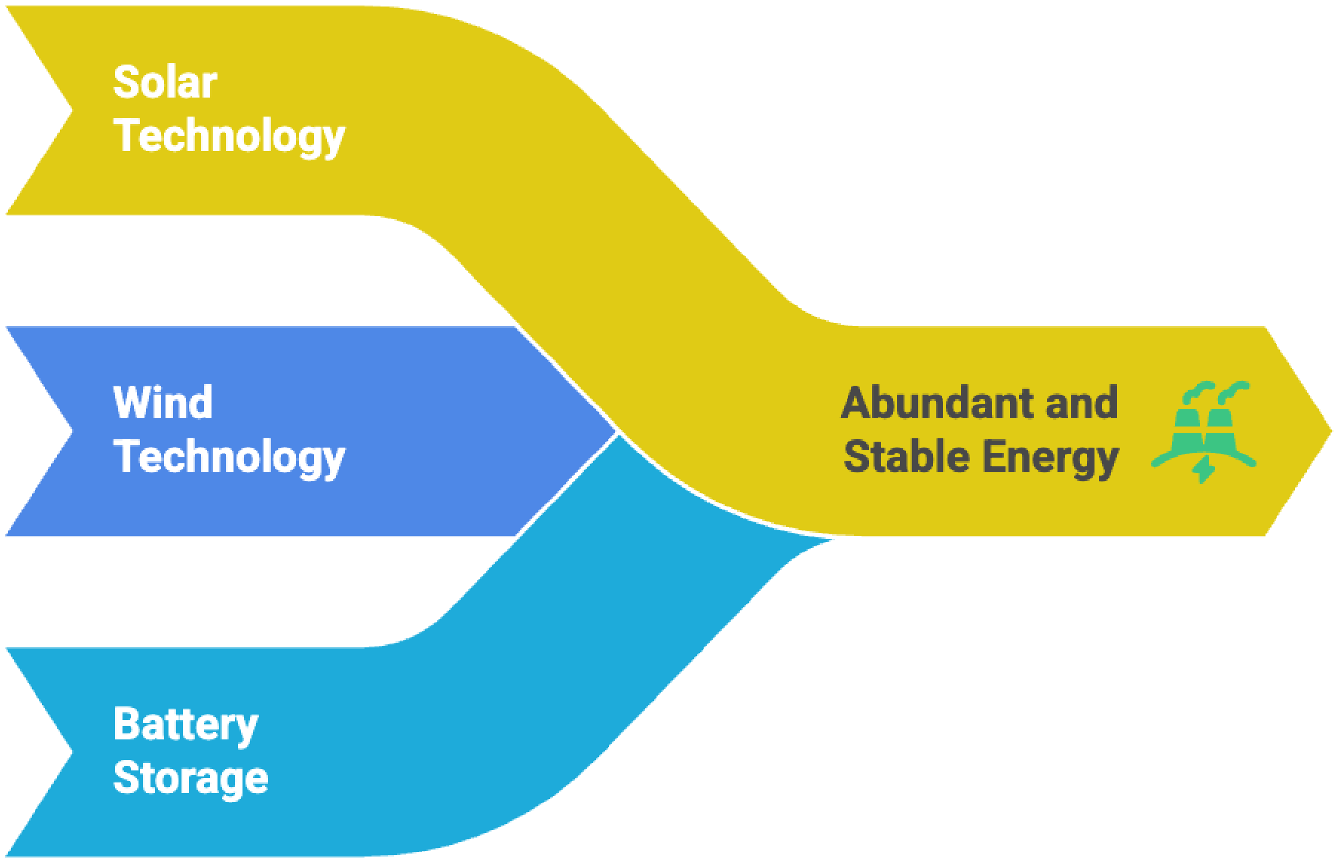
Percentage of UK electricity generated by source each year

Data before 1996 are estimates and figures from 1987 to 1989 are especially uncertain. Figures before 1987 only include major power producers. "Other" includes sources such as bioenergy and pumped storage

Source: Department for Energy Security and Net Zero

Source: BBC

Transition to Technological Energy



1 Solar technology enables abundant and sustainable energy production.

2 Wind technology contributes to decentralized and renewable energy.

3 Battery storage stabilizes energy supply and enhances efficiency.

Solar Technology

Wind Technology

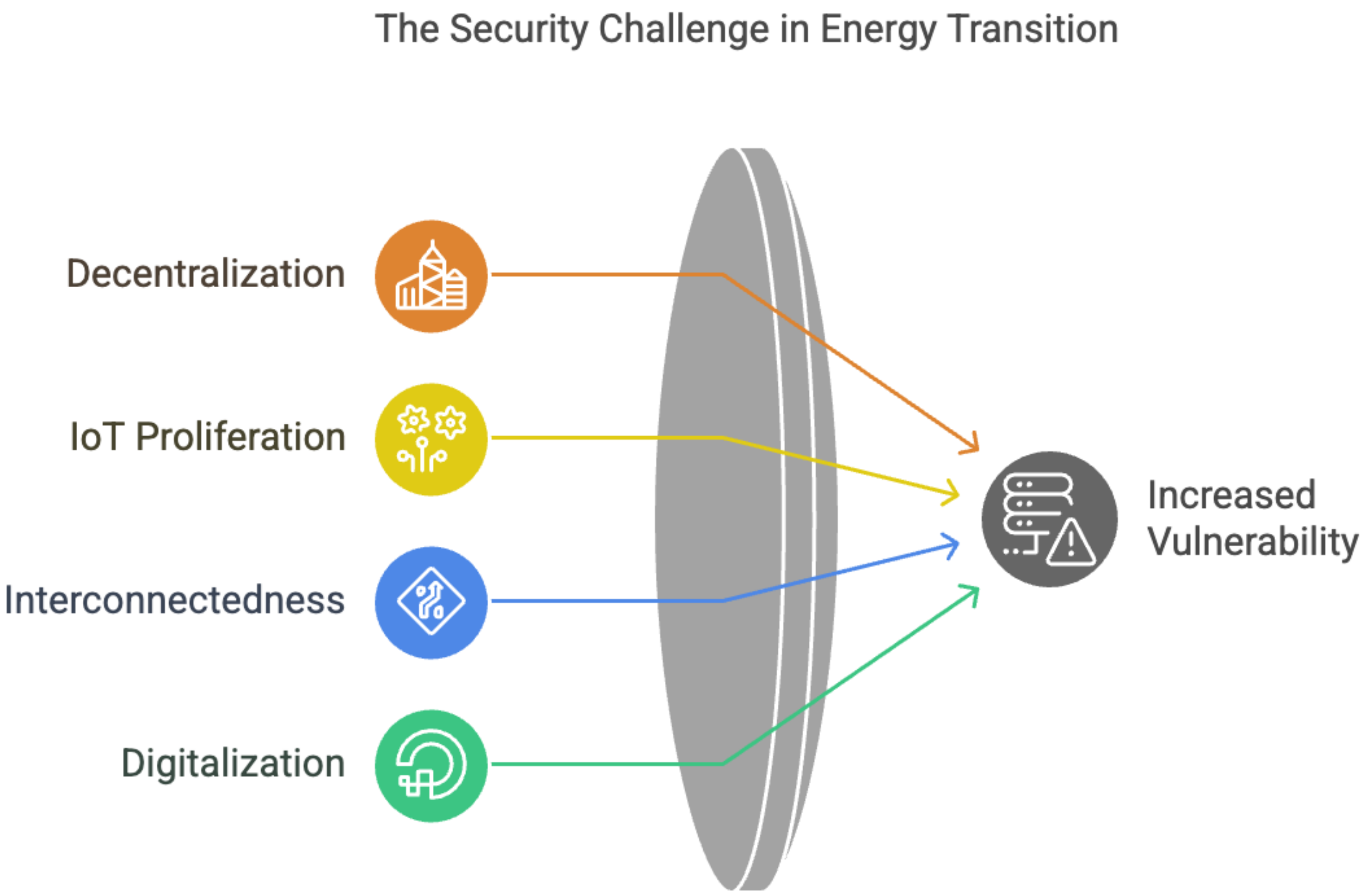
Battery Storage

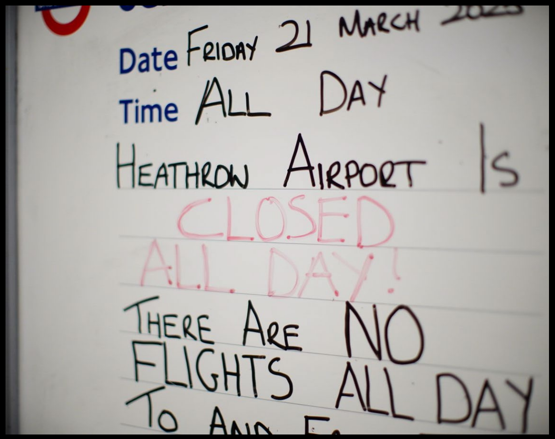
Abundant and Stable Energy

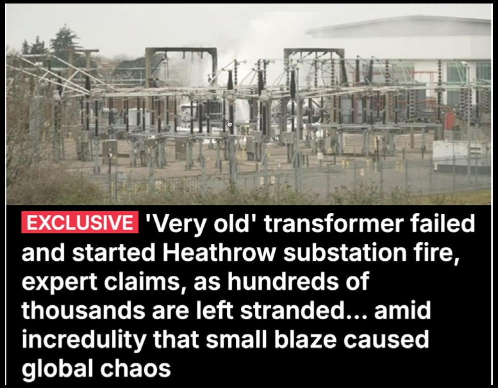
Powering your winery with a battery you can drive

- **EVs can power your home or the grid** – Bidirectional charging
- **Save up to $30,000** over 10 years by using your car as a battery
- **EV batteries = 5 * bigger** than most home batteries
- **Real-world example:** Winery powered by solar + Nissan Leaf
- **National potential:** $2.96 billion in system-wide savings
- **Bonus:** EVs can be **cheaper and more** flexible than home batteries

Embracing the Age of Energy Abundance - Securely


The Security Challenge in Energy Transition




EXCLUSIVE 'Very old' transformer failed and started Heathrow substation fire, expert claims, as hundreds of thousands are left stranded... amid incredulity that small blaze caused global chaos

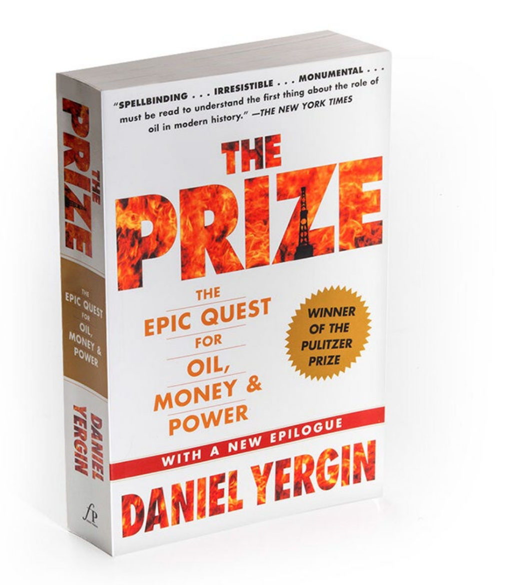
Heathrow Blackout: A Case of Predictable Weirdness

- **Single Point of Failure:** Fire at one substation shut down Europe's busiest airport
- **Known but Ignored Risk:** No redundancy despite clear vulnerabilities
- **Systemic Weakness:** Complex systems can fail from small overlooked parts
- **False sense of security:** Rules and regulations can mask deeper risks
- **Key takeaway:** Resilience requires active risk management—not just rule-following but challenging assumptions

Daniel Yergin

“The Prize”

- Oil shaped the 20th century
- Energy has acted as a central pillar shaping geopolitical alliances and conflicts, driving significant shifts in global economic power and structures.


"SPELLBINDING . . . IRRESISTIBLE . . . MONUMENTAL . . . must be read to understand the first thing about the role of oil in modern history." —THE NEW YORK TIMES
THE PRIZE
THE EPIC QUEST FOR OIL, MONEY & POWER
WINNER OF THE PULITZER PRIZE
WITH A NEW EPILOGUE
DANIEL YERGIN

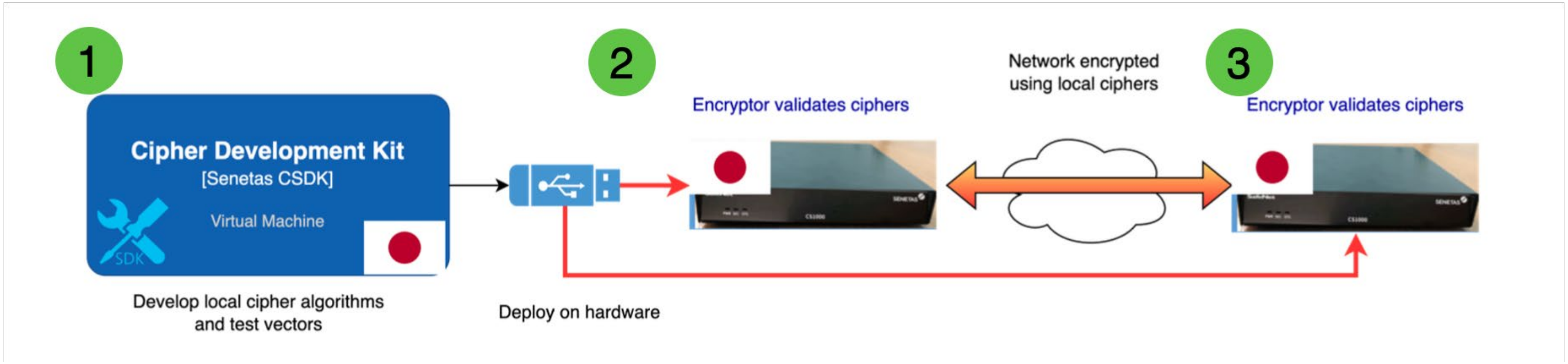
The **post-globalization movement in technology is now spreading to data, AI, security, and privacy**

The G-Zero World: A Leadership Vacuum

Technology Report 2024

Technology meets the moment as AI delivers results.

Bringing Sovereign Encryption to Life: How It Works


1
Cipher Development Kit
[Senetas CSDK]
Virtual Machine
SDK
Develop local cipher algorithms and test vectors
Deploy on hardware
2
Encryptor validates ciphers
Network encrypted using local ciphers
3
Encryptor validates ciphers

1. Customer’s cipher design
2. Secure upload
3. Cipher in use

2025
Edelman
Trust
Barometer

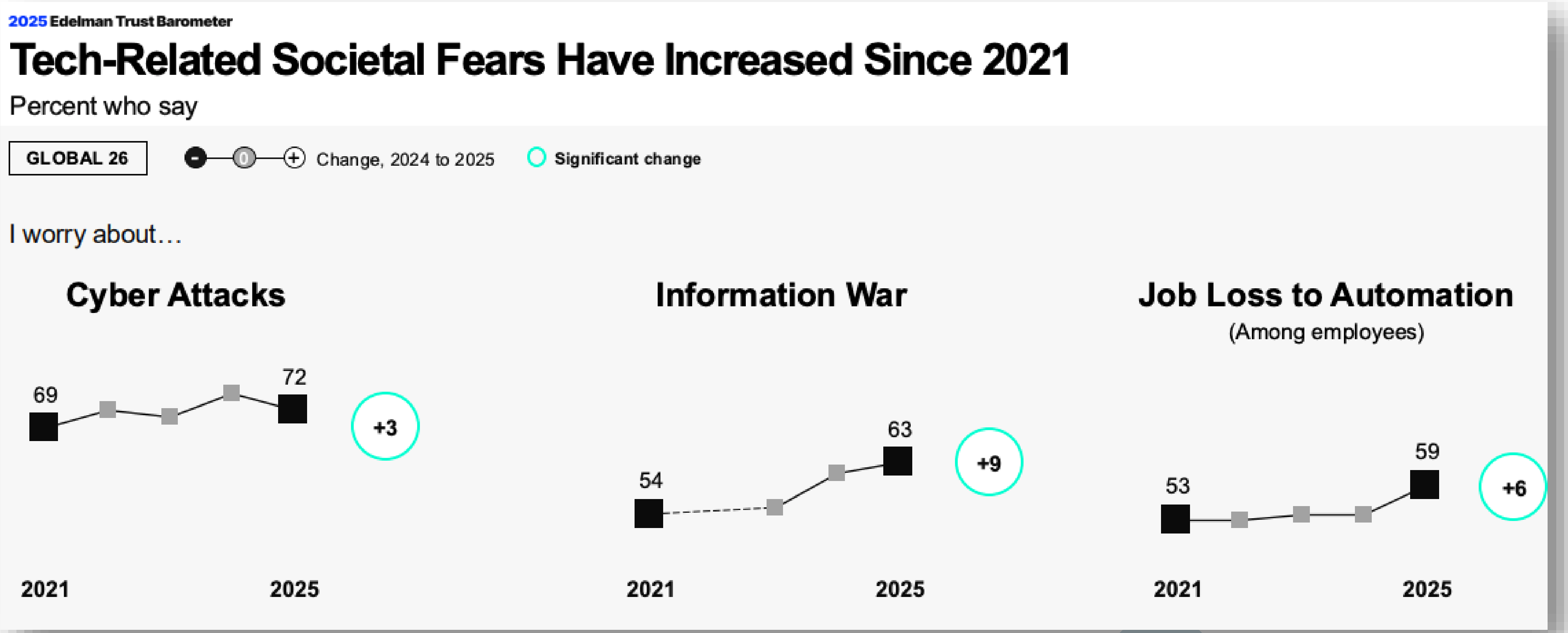
Trust and the Crisis of Grievance

With Insights for the Technology Sector

https://www.edelman.com/

2025
Edelman
Trust
Barometer

Trust and the Crisis of Grievance
With Insights for the Technology Sector



2025 Edelman Trust Barometer

Tech-Related Societal Fears Have Increased Since 2021

Percent who say

GLOBAL 26

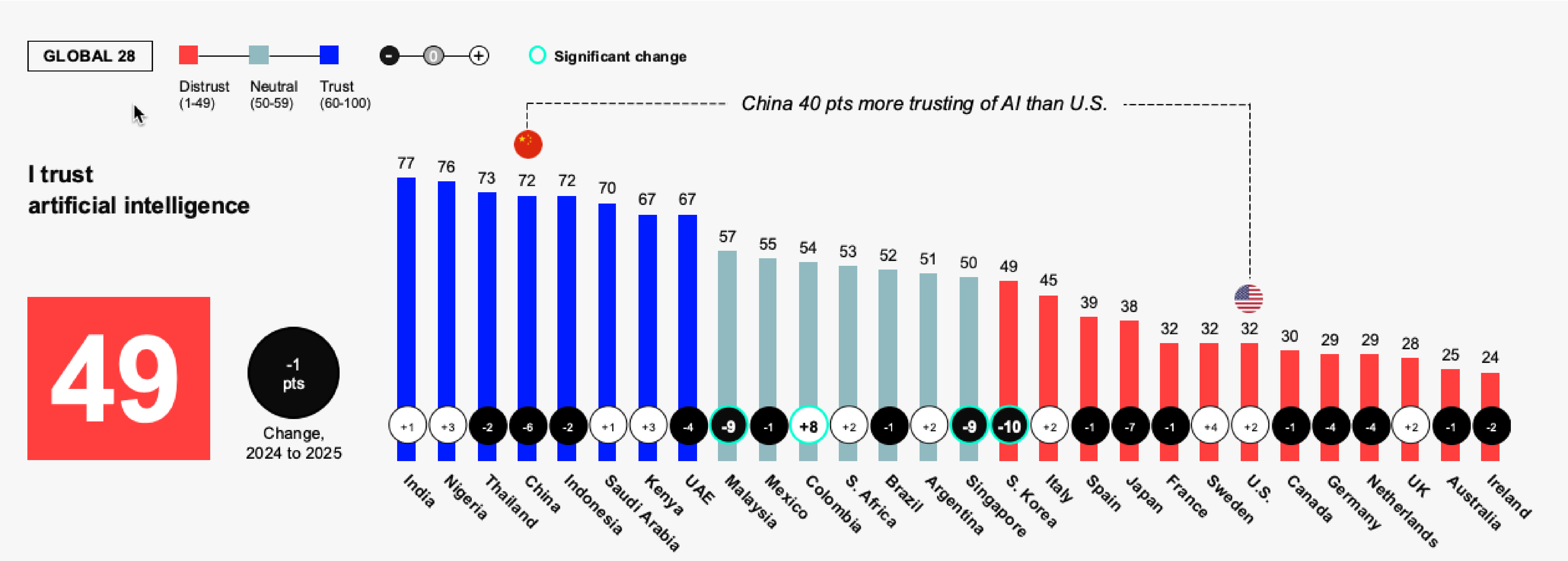
● ○ ⊕ Change, 2024 to 2025 ○ Significant change

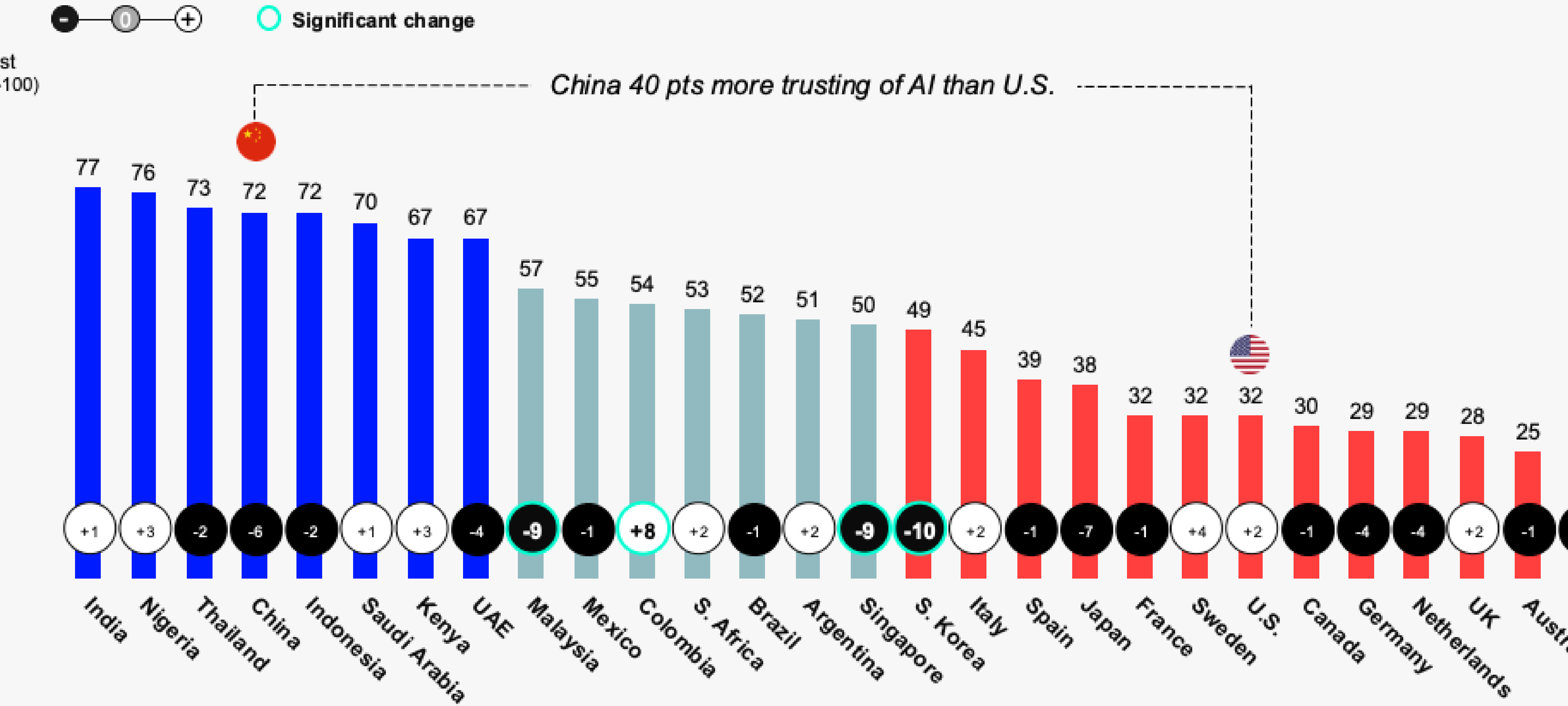
I worry about…

https://www.edelman.com/

Trust in Artificial Intelligence is Higher in Developing World Than Developed

Percent who say



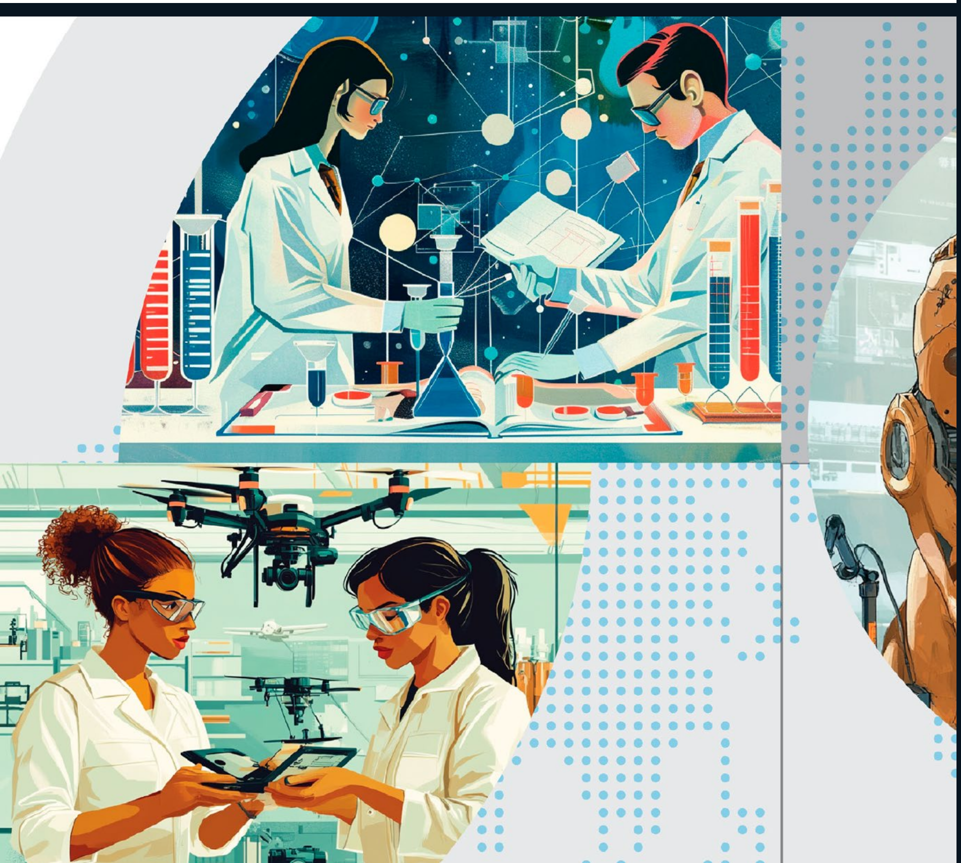


-
0
+
Significant change
st
-100)
China 40 pts more trusting of AI than U.S.
77
76
73
72
72
70
67
67
57
55
54
53
52
51
50
49
45
39
38
32
32
32
30
29
29
28
25
+1
+3
-2
-6
-2
+1
+3
-4
-9
-1
+8
+2
-1
+2
-9
-10
+2
-1
-7
-1
+4
+2
-1
-4
-4
+2
-1
India
Nigeria
Thailand
China
Indonesia
Saudi Arabia
Kenya
UAE
Malaysia
Mexico
Colombia
S. Africa
Brazil
Argentina
Singapore
S. Korea
Italy
Spain
Japan
France
Sweden
U.S.
Canada
Germany
Netherlands
UK
Austr

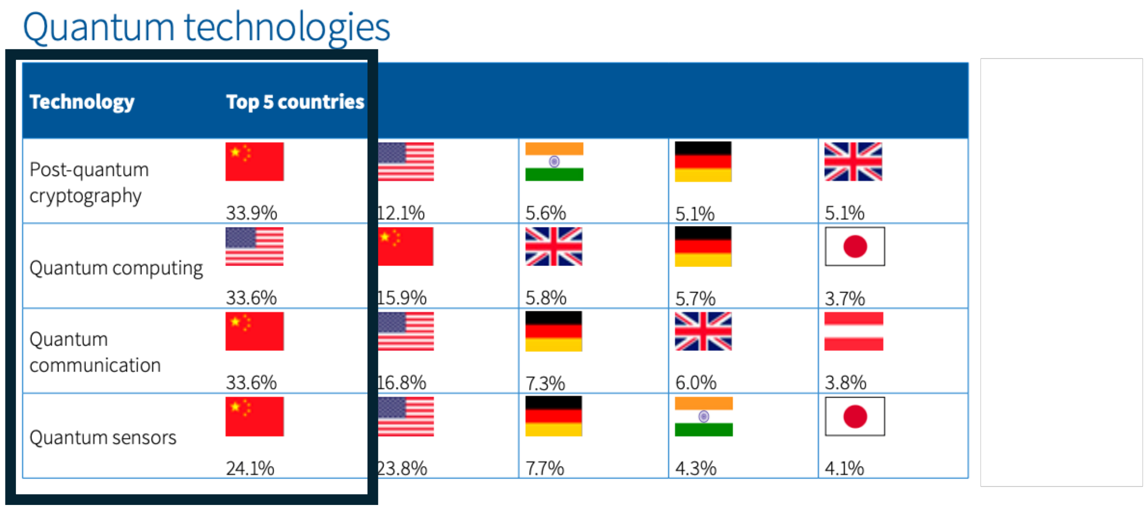
ASPI

AUSTRALIAN
STRATEGIC
POLICY
INSTITUTE

CRITICAL
TECHNOLOGY
TRACKER

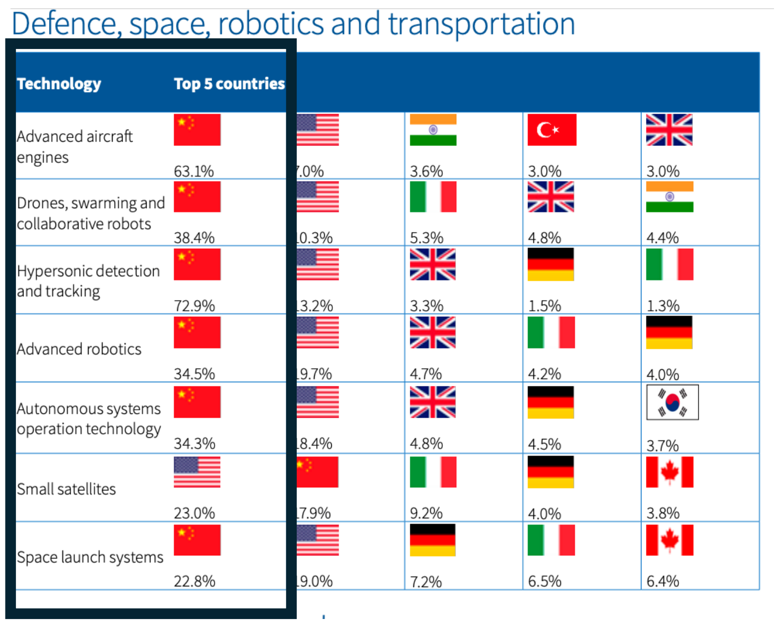


- ASPI monitors 64 critical technologies
- Shows 21 years of data
- US led 60/64 techs in 2003–2007
- China leads 57/64 in 2019–2023
- China’s share of global manufacturing set to hit 45% by 2030



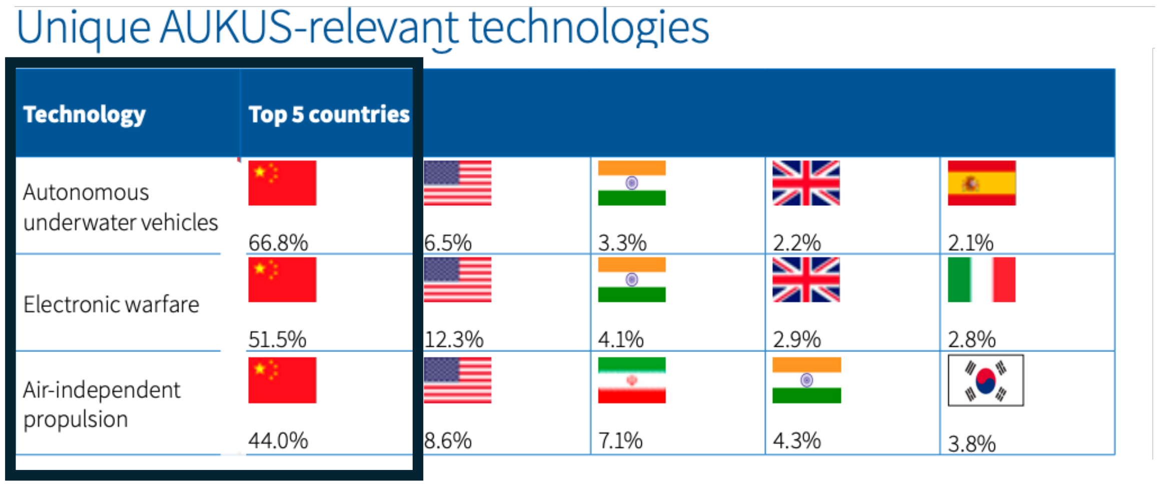
Quantum technologies

Technology	Top 5 countries				
Post-quantum cryptography	🇨🇳 33.9%	🇺🇸 12.1%	🇮🇳 5.6%	🇩🇪 5.1%	🇬🇧 5.1%
Quantum computing	🇺🇸 33.6%	🇨🇳 15.9%	🇬🇧 5.8%	🇩🇪 5.7%	🇯🇵 3.7%
Quantum communication	🇨🇳 33.6%	🇺🇸 16.8%	🇩🇪 7.3%	🇬🇧 6.0%	🇦🇹 3.8%
Quantum sensors	🇨🇳 24.1%	🇺🇸 23.8%	🇩🇪 7.7%	🇮🇳 4.3%	🇯🇵 4.1%



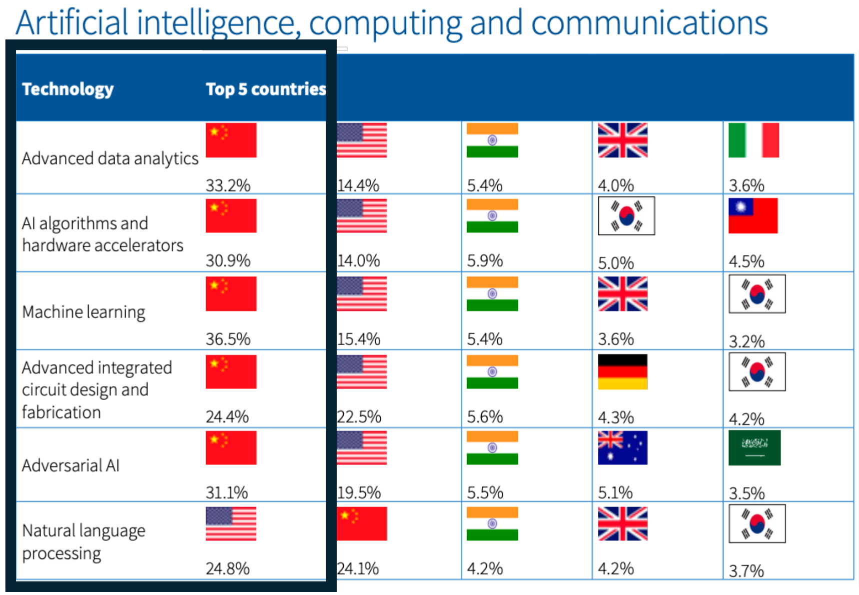
Defence, space, robotics and transportation

Technology	Top 5 countries				
Advanced aircraft engines	🇨🇳 63.1%	🇺🇸 7.0%	🇮🇳 3.6%	🇹🇷 3.0%	🇬🇧 3.0%
Drones, swarming and collaborative robots	🇨🇳 38.4%	🇺🇸 10.3%	🇮🇹 5.3%	🇬🇧 4.8%	🇮🇳 4.4%
Hypersonic detection and tracking	🇨🇳 72.9%	🇺🇸 13.2%	🇬🇧 3.3%	🇩🇪 1.5%	🇮🇹 1.3%
Advanced robotics	🇨🇳 34.5%	🇺🇸 19.7%	🇬🇧 4.7%	🇮🇹 4.2%	🇩🇪 4.0%
Autonomous systems operation technology	🇨🇳 34.3%	🇺🇸 18.4%	🇬🇧 4.8%	🇩🇪 4.5%	🇰🇷 3.7%
Small satellites	🇺🇸 23.0%	🇨🇳 17.9%	🇮🇹 9.2%	🇩🇪 4.0%	🇨🇦 3.8%
Space launch systems	🇨🇳 22.8%	🇺🇸 19.0%	🇩🇪 7.2%	🇮🇹 6.5%	🇨🇦 6.4%



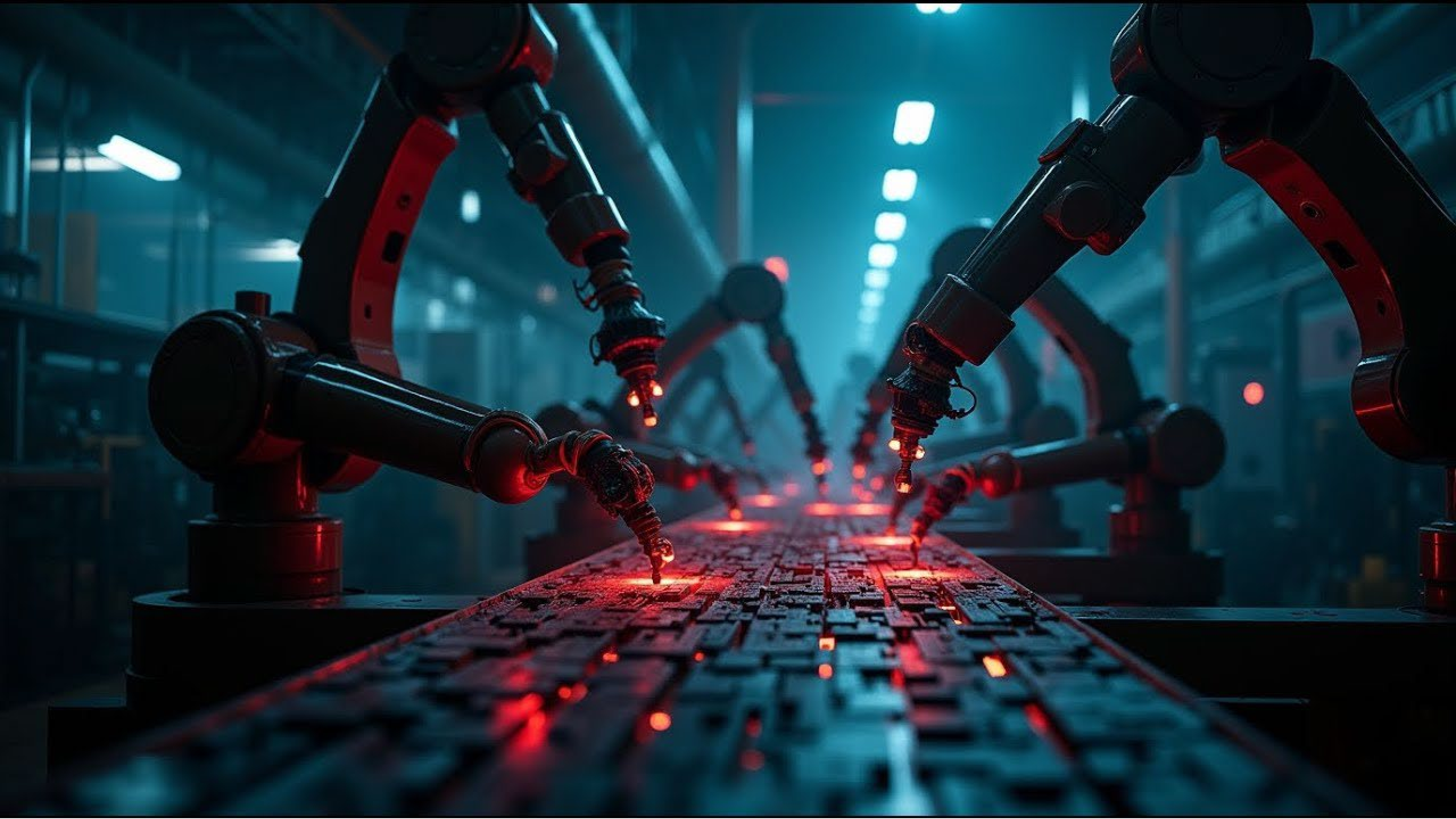
Unique AUKUS-relevant technologies

Technology	Top 5 countries				
Autonomous underwater vehicles	🇨🇳 66.8%	🇺🇸 6.5%	🇮🇳 3.3%	🇬🇧 2.2%	🇪🇸 2.1%
Electronic warfare	🇨🇳 51.5%	🇺🇸 12.3%	🇮🇳 4.1%	🇬🇧 2.9%	🇮🇹 2.8%
Air-independent propulsion	🇨🇳 44.0%	🇺🇸 8.6%	🇮🇷 7.1%	🇮🇳 4.3%	🇰🇷 3.8%



Artificial intelligence, computing and communications

Technology	Top 5 countries				
Advanced data analytics	🇨🇳 33.2%	🇺🇸 14.4%	🇮🇳 5.4%	🇬🇧 4.0%	🇮🇹 3.6%
AI algorithms and hardware accelerators	🇨🇳 30.9%	🇺🇸 14.0%	🇮🇳 5.9%	🇰🇷 5.0%	🇹🇼 4.5%
Machine learning	🇨🇳 36.5%	🇺🇸 15.4%	🇮🇳 5.4%	🇬🇧 3.6%	🇰🇷 3.2%
Advanced integrated circuit design and fabrication	🇨🇳 24.4%	🇺🇸 22.5%	🇮🇳 5.6%	🇩🇪 4.3%	🇰🇷 4.2%
Adversarial AI	🇨🇳 31.1%	🇺🇸 19.5%	🇮🇳 5.5%	🇦🇺 5.1%	🇸🇦 3.5%
Natural language processing	🇺🇸 24.8%	🇨🇳 24.1%	🇮🇳 4.2%	🇬🇧 4.2%	🇰🇷 3.7%



Dubbed the “dark factory,” **this facility runs entirely on artificial intelligence (AI) and robotics**, eliminating the need for lighting, lunch breaks, or labor

CCTV 1
综合
CCTV
直播
中央广播电视总台
春节联欢晚会
2025
CCTV

Fully Automated Luxury Communism

A provocative vision: life beyond the necessity of work

What if robots did the labor - and humans pursued meaning

Automation as liberation: time for creativity, intellect and leisure

‘….interesting times’

Getty Images