Unintended Consequences of AI

Week 10 – AI Risk Management

CSAI Learning Objectives

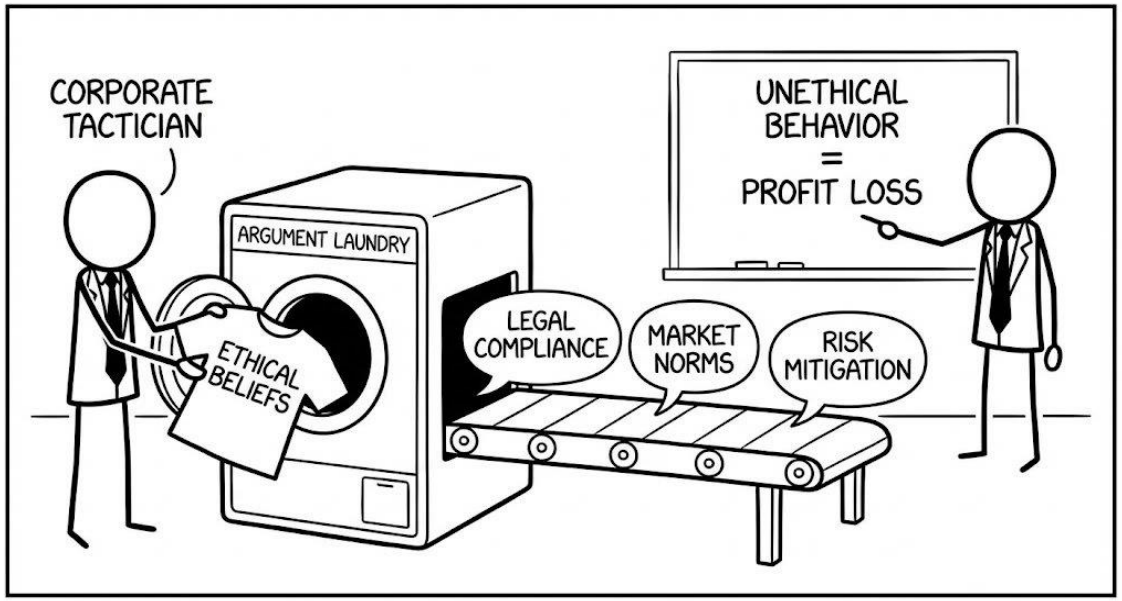
Daniel

3. Analyse case studies to **identify** and **mitigate potential risks** considering **legal**, **social**, ethical or **professional issues.**

4. Apply **ethical methodologies** in the **design of responsible AI systems**.

Week 6, **Week 7**, **Week 8**, **Week 9**, **Week 10**.

The limits of risk management



AI may have unintended impacts on...

1. Environment
2. Labour market
3. Infosphere
4. Security & existential risk

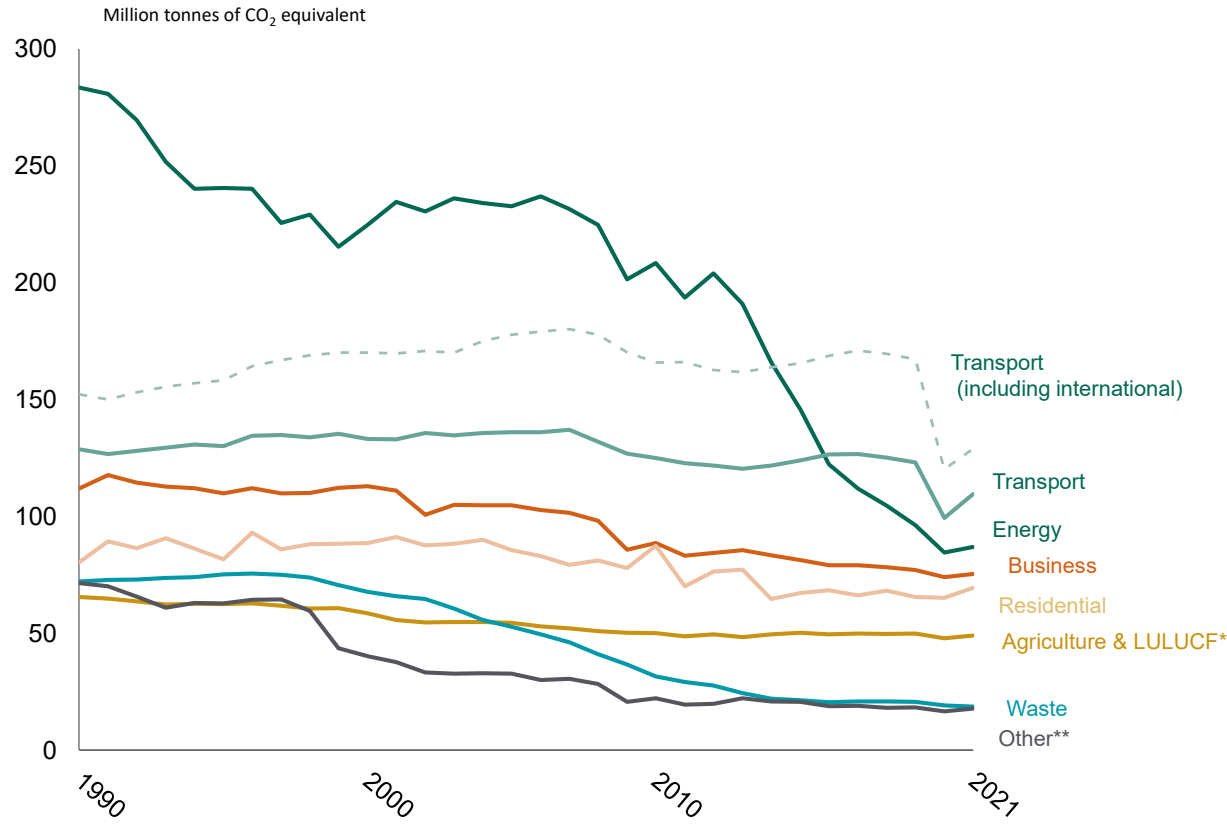
Warning: Social systems are **complex**

The scientific community is uncertain.

1. Environmental Consequences

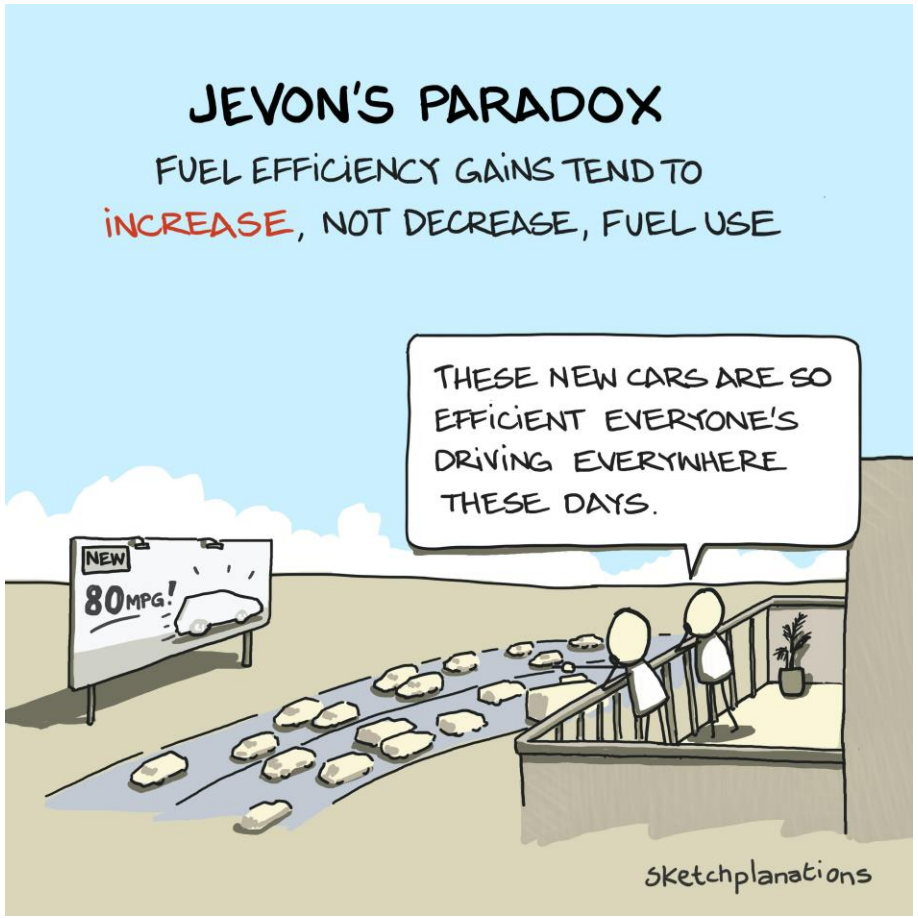
The big picture

- Optimists argue AI efficiencies will lead to decreased emissions
  - More efficient models
  - Efficiency across energy, transport etc
- Critics focus on data centres
  - Emissions from run time shows up as energy and business
  - Water for cooling
  - Hardware require minerals
  - E-waste due to obsolescence



Source: https://www.gov.uk/government/statistics/transport-and-environment-statistics-2023/transport-and-environment-statistics-2023

Why AI efficiency is unlikely to reduce emissions



Two approaches to environmental ethics

Micro

- Typically deontological
  - Given *some* harm, ask whether specific AI activity is necessary?
  - Sceptical by default
- Places responsibility on individuals
  - De-emphasises systems

Macro

- Typically utilitarian
  - What is the environmental impact of AI activity?
  - Implies harms can be justified
- Used to deflect blame onto other activities + systems
  - Agriculture (meat)

Two approaches to environmental ethics

Micro

- Typically deontological
  - Given *some* harm, ask whether specific AI activity is necessary?
  - Focus on "waste"
- Places responsibility onindividuals
  - De-emphasises systems

Macro

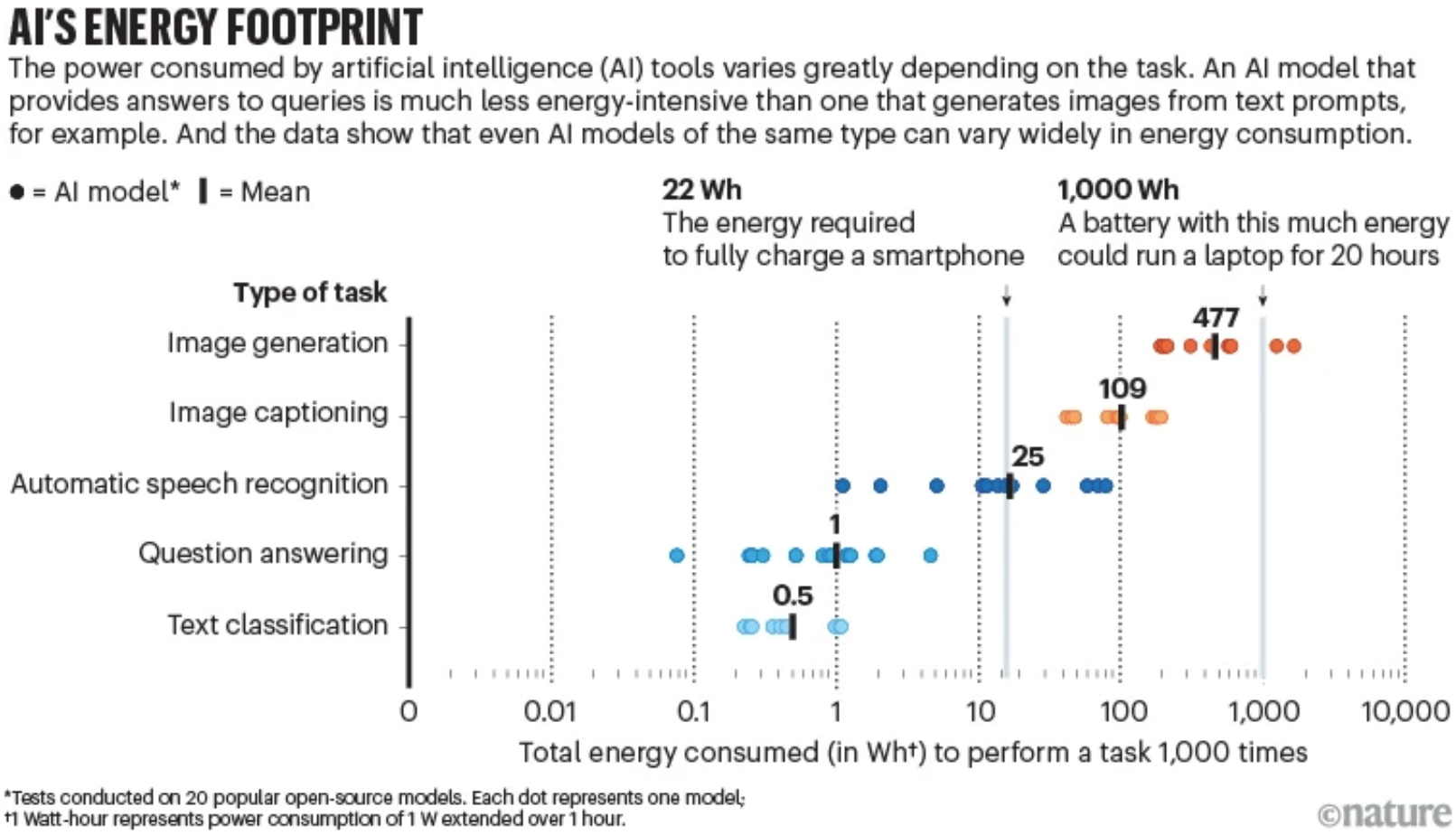
- Typically utilitarian
  - What is the environmental impact of AI activity?
  - Focus on relative harm
- Used to deflect blame onto other activities + systems
  - Agriculture (meat)

The ethics of each prompt

Ad campaign launch in 2003

Privatisation of (AI) responsibility

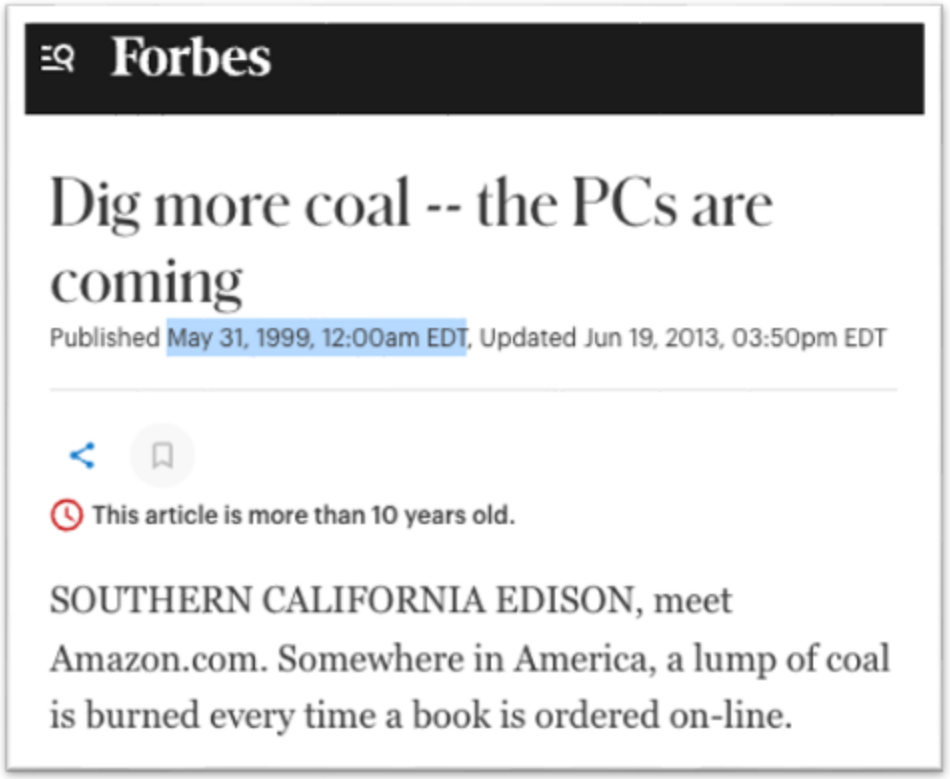
Privatisation of (AI) responsibility



Source: Unpublished analysis by S. Luccioni *et al.*/AI Energy Star project

Lack of transparency over total energy costs of AI models

Estimates & projections are often wrong



Forbes

Dig more coal -- the PCs are coming

Published May 31, 1999, 12:00am EDT, Updated Jun 19, 2013, 03:50pm EDT

This article is more than 10 years old.

SOUTHERN CALIFORNIA EDISON, meet Amazon.com. Somewhere in America, a lump of coal is burned every time a book is ordered on-line.

"It's now reasonable to project that **half** of the electric grid will be powering the digital-Internet economy **within the next decade**."

Impacts across lifecycle of an AI tool

Hardware

- Metals and minerals
- Production energy
- "Pure" water to treat contaminants
- Cooling lithography
- Disposal costs

Training

- Scraping infra
- Compute energy
- Water for cooling
- Fine tuning plus human supervisors

Ongoing

- Compute energy
- Water for cooling
- Idling costs
- Moderation

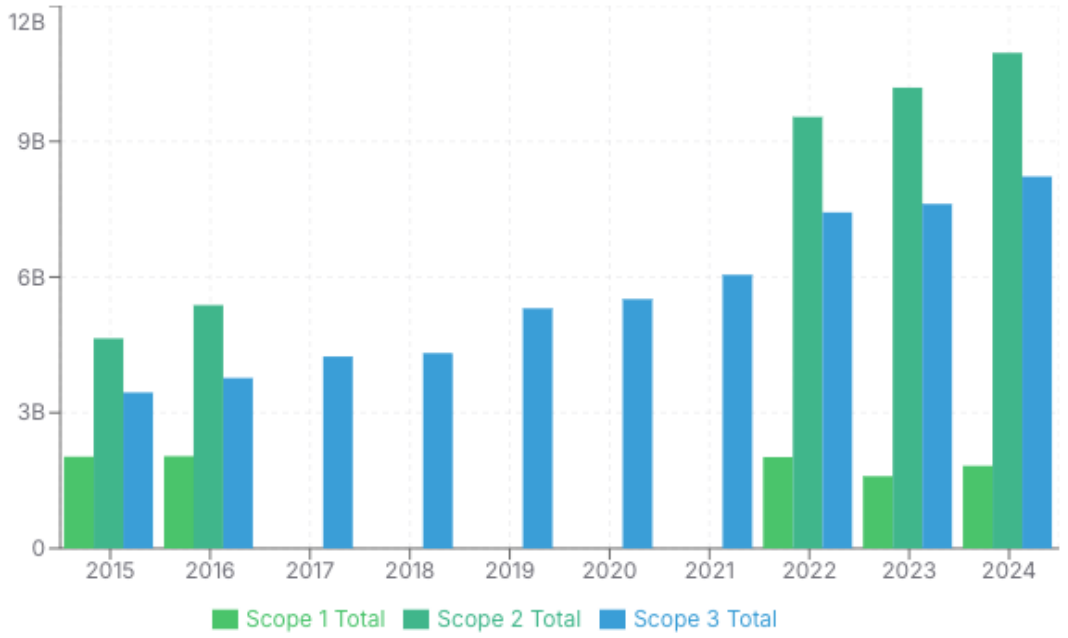
Hardware emissions

"In 2024, Taiwan Semiconductor Manufacturing Company reported total carbon emissions of approximately 30.5 billion kg CO2e:

- Direct: 1.8bn
- Energy: 10.9bn
- Supply chain: 8.2bn"

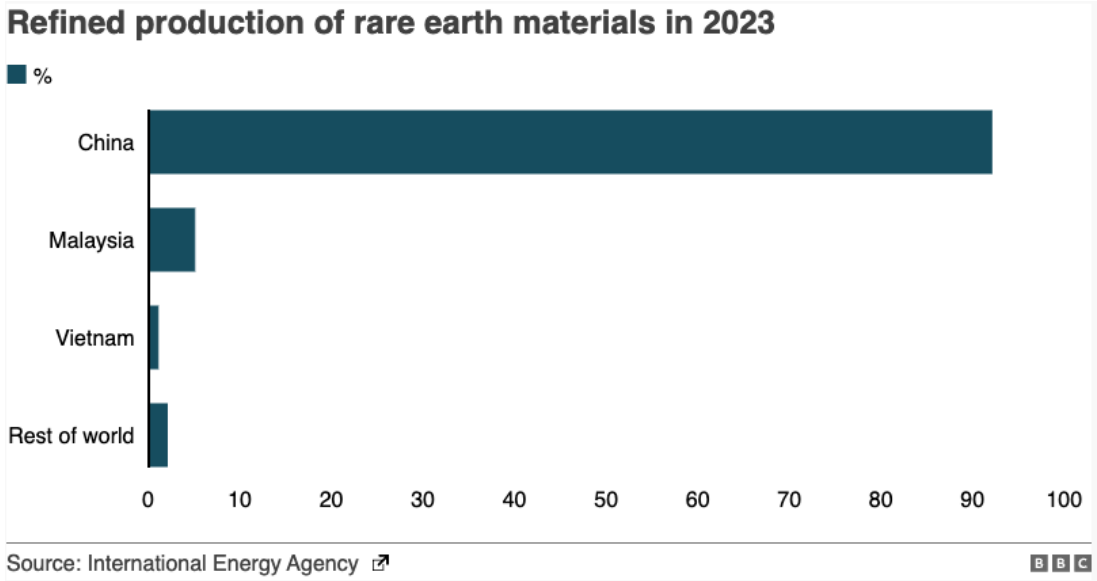
+ve framing: Produces ~90% of high-end chips at less than 0.1% of global emissions.

-ve framing: Same ballpark as Norway (~38bn kg) with population of 5.5m and GDP of $500bn.



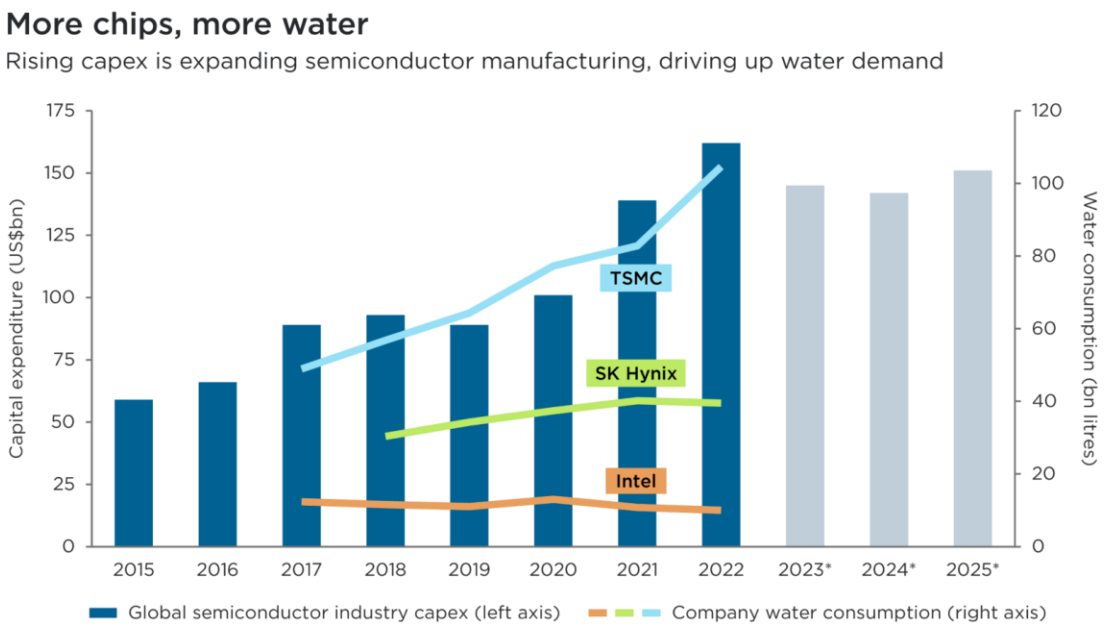
Source: https://ditchcarbon.com/organizations/taiwan-semiconductor-manufacturing

Environmental damage in supply chain



Source: https://www.bbc.com/news/articles/c1drqeev36qo

(Pure) water usage in making hardware



Source: https://impaxam.com/insights-and-news/blog/quenching-the-semiconductor-industrys-thirst/

"Approximately 3.8tn cubic metres of water is used by humans annually with 70% being consumed by the global agriculture sector"

Source: https://www.theguardian.com/news/datablog/2013/jan/10/how-much-water-food-production-waste

→ TSMC's water usage is a tiny fraction (less than 0.003%) of global water usage.

Report: Intel Ocotillo Campus notches honor for water stewardship

AI hardware becomes e-waste

- AI chips have a tough life
  - Massive electrical and thermal energy on the head of a pin
  - Running 24/7 at ~70% capacity
- And then they get replaced
  - New chips more powerful (x25 NVIDIA 2026) and energy efficient
  - New software optimized for newer chips
- There are recycling efforts to reclaim metals
  - But not embodied carbon

Data Center AI GPUs May Have Extremely Short Lifespans

An unnamed Google engineer reportedly said the chips only last about three years, which could be good news for Nvidia.

By Ryan Whitwam October 24, 2024

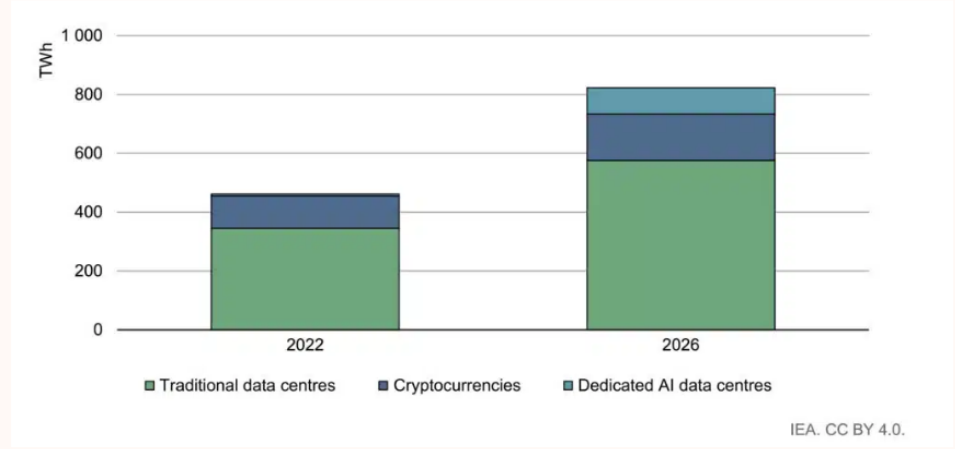
Credit: Google

Generative AI could cause 10 billion iPhones' worth of e-waste per year by 2030

Devin Coldewey — 9:01 AM PDT · October 28, 2024

Costs of training and operating AI models

- Training costs dominated in ML era
  - "an estimated 80 to 90 percent of the cost of machine learning at scale devoted to AI inference" Jensen Huang, 2019
- Inference dominates in AI era
  - "Inference accounts for 60–70% of energy consumption, compared with 20–40% for training" source
  - Huang: 3 eras of AI: generative, thinking, and agentic. Need x100 the inference in each.
- AI is a fraction of total data centre and even cryptocurrency costs



↗ **Estimated electricity demand for** traditional **data centers, AI-dedicated data centers and crypto-currencies, 2022 and 2026 (reference scenario)**[9]. Note: electricity demand for data centres excludes consumption by data network centres.

AI boom has caused same CO2 emissions in 2025 as New York City, report claims

Study author says tech companies are reaping benefits of artificial intelligence age but society is left to pay cost

Robert Booth *UK technology editor*

Thu 18 Dec 2025 11.15 GMT

Water consumption of AI data centres

Sam Altman claims an average ChatGPT query uses 'roughly one fifteenth of a teaspoon' of water / Altman shared the unsourced statistic in a new blog post.

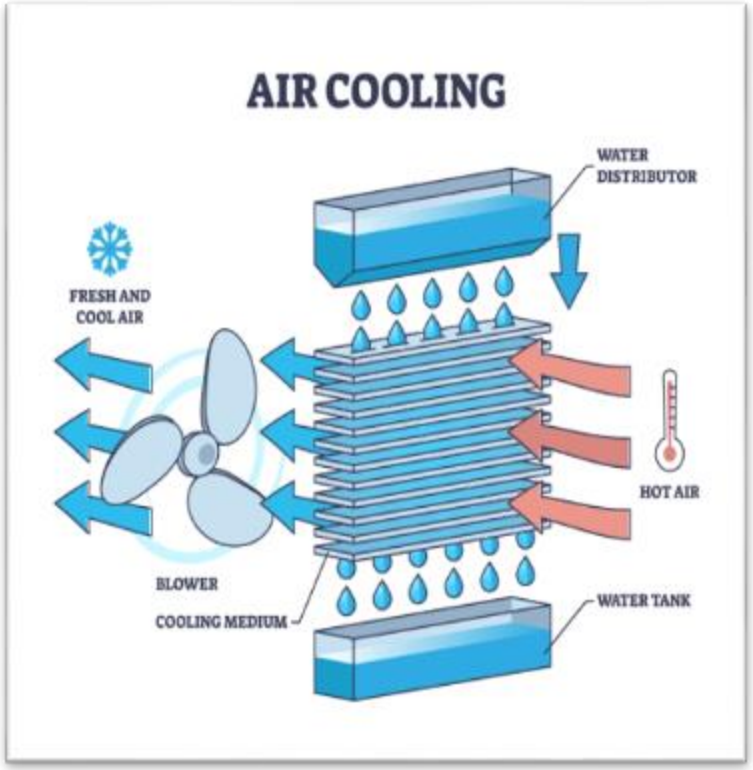
by Jay Peters
Jun 10, 2025, 11:28 PM GMT+1

75 Comments (All New)

Image: Cath Virginia / The Verge; Getty Images

Scientists count embodied emissions too

- Training GPT-3 in MSFT's U.S. data centers consumed 5.4 million litres
- GPT-3 needs to "drink" a 500ml bottle of water for roughly 10 – 50 medium-length responses
  - depending on **when** and **where** it is deployed
- Global AI demand is projected to account for 4.2 – 6.6 billion cubic meters of water withdrawal in 2027
  - 4 – 6 Denmark or
  - half of the UK
- Water usage is **local**, unlike emissions



Source: https://arxiv.org/pdf/2304.03271

Two approaches to environmental ethics

Micro

- Typically deontological
  - Given *some* harm, ask whether specific AI activity is necessary?
  - Focus on "waste"
- Places responsibility on individuals
  - De-emphasises systems

Macro

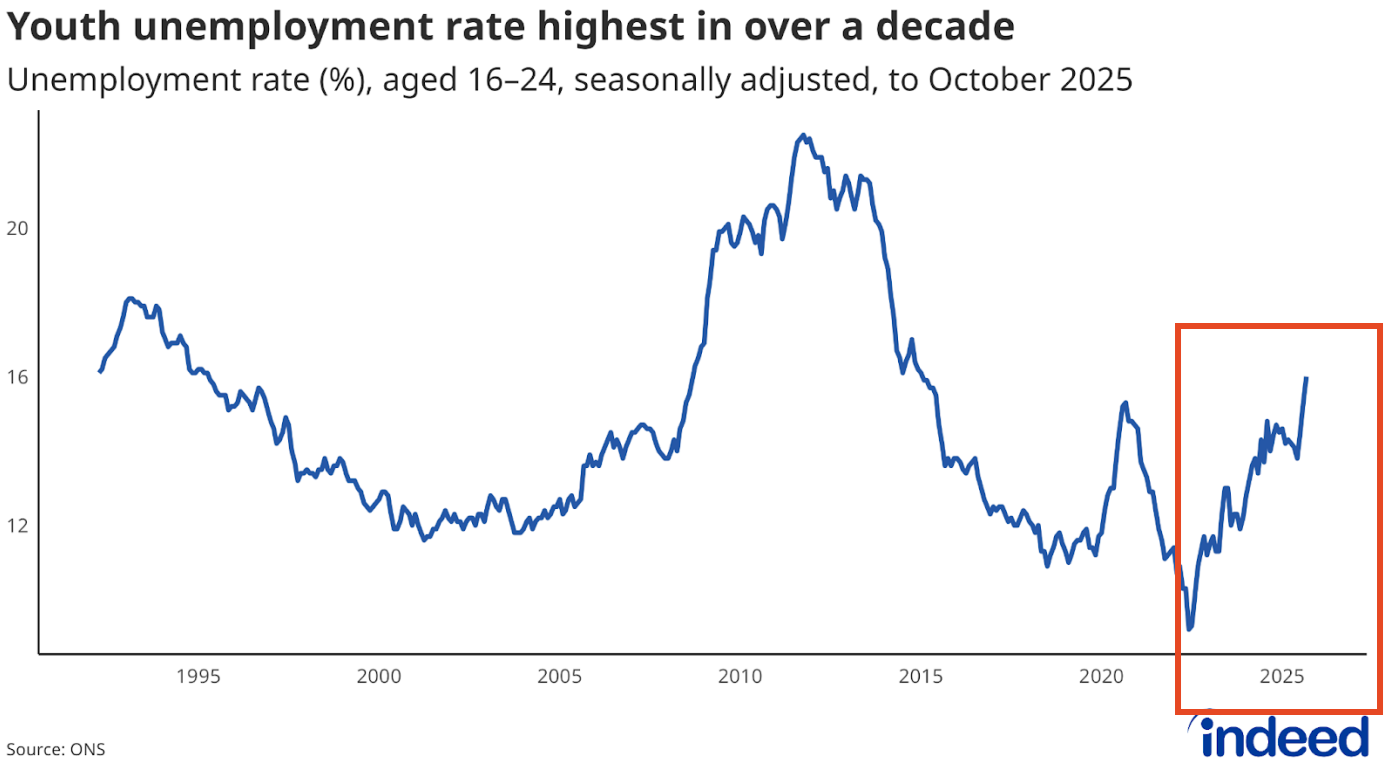
- Typically utilitarian
  - What is the environmental impact of AI activity?
  - Focus on relative harm
- Used to deflect blame onto other activities + systems
  - Agriculture (meat)

What should be done about AI and the environment? By who?

2. Labour Market Consequences

Why is the UK job market so bad?

- Perspective matters
  - Red box suggests AI job losses
  - Broader view says we're not in a historical anomaly
- Multi-causal explanations
  - AI
  - Macroeconomic trends
    - Financial crisis in 2008
    - Brexit
  - Demographic trends
  - Changes in government policy
    - Minimum wage, taxes etc



Labour economics **even more uncertain** than environmental

Lessons from the history of automation

Journal of Economic Perspectives—Volume 29, Number 3—Summer 2015—Pages 3–30

Why Are There Still So Many Jobs? The History and Future of Workplace Automation†

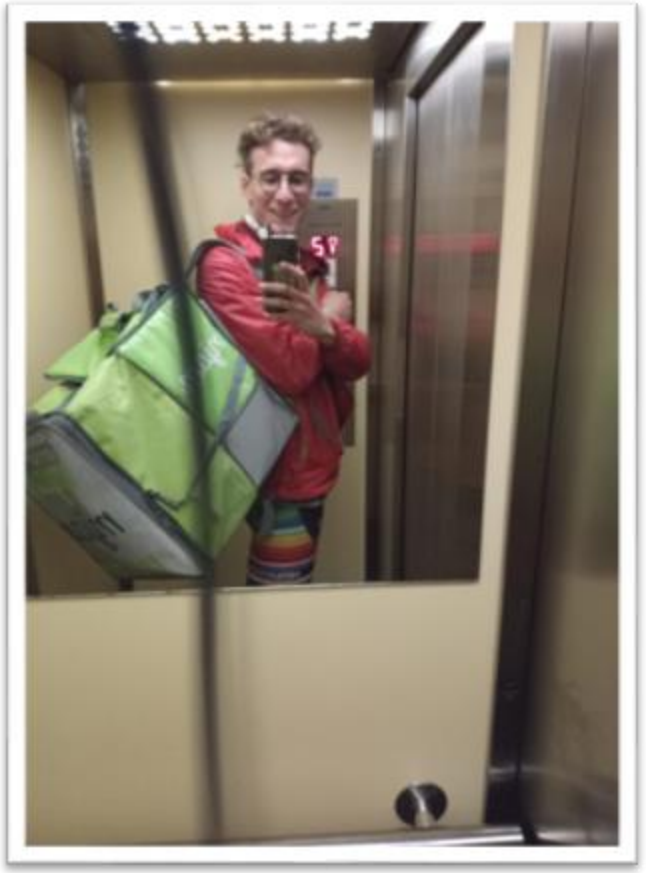
David H. Autor

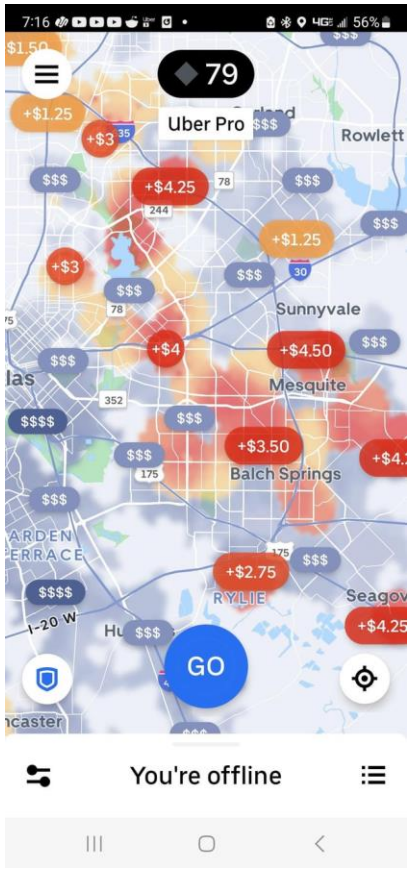
Source: https://economics.mit.edu/sites/default/files/publications/why%20are%20there%20still%20jobs%202014.pdf

1. Automation reduces **and** increases the demand for labour
2. Focus on how it changes types of available jobs and the pay
   - He predicts it replaces "routine, codifiable tasks" but increases demand for **"problem-solving skills, adaptability, and creativity"**

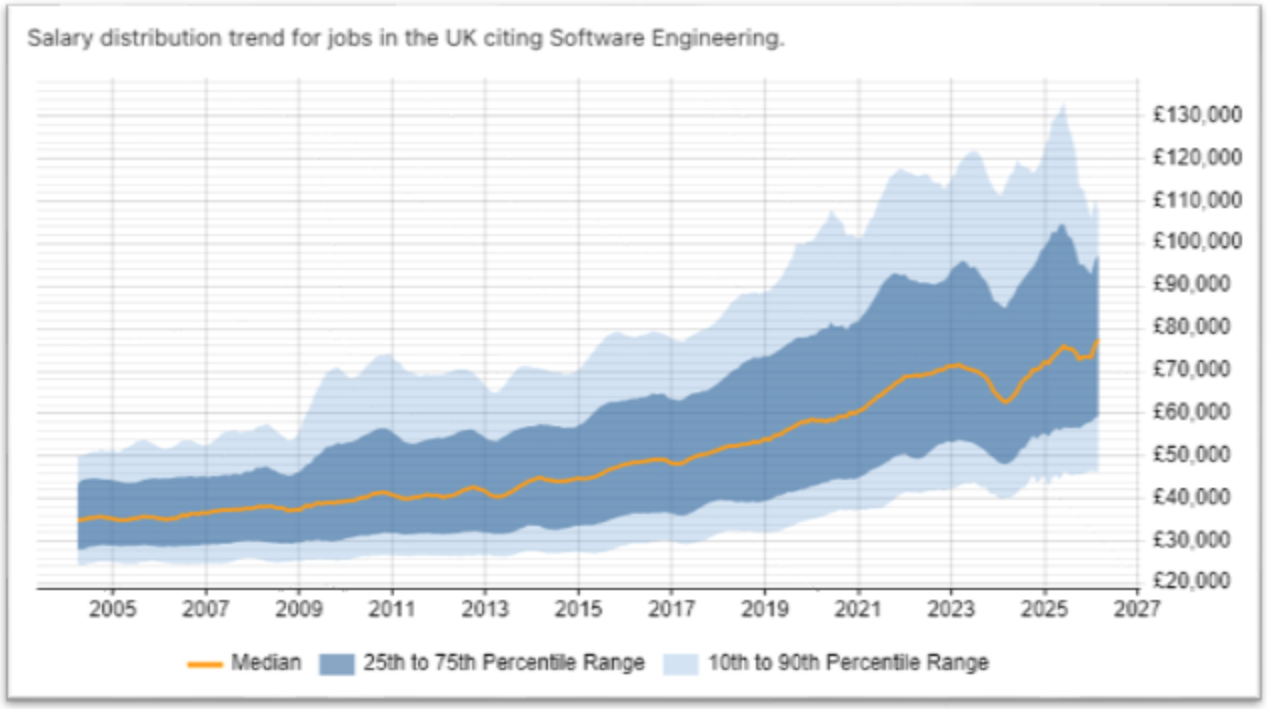
How technology increases demand for labour

- Management is hard and costly
  - Training & onboarding
  - Forecasting customer demand
  - Managing shifts + time off
  - Monitoring quality
  - Rewarding good performance
  - Complying with labour laws
  - Emotional cost
- Technology can automate
  - Usable app requires less training
  - ML perfect forecaster of demand
  - Variable pricing to manage supply
  - Reviews to monitor quality
  - Less care for law or emotions?

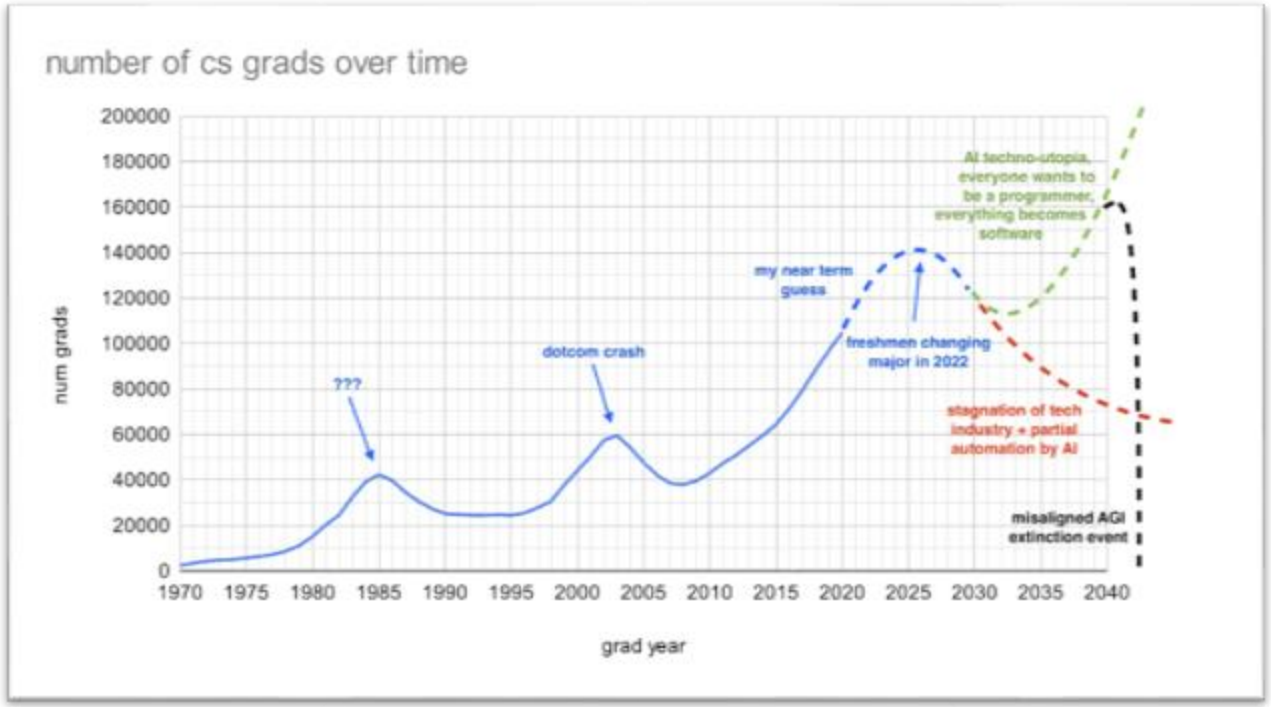




Increased demand and supply of coders



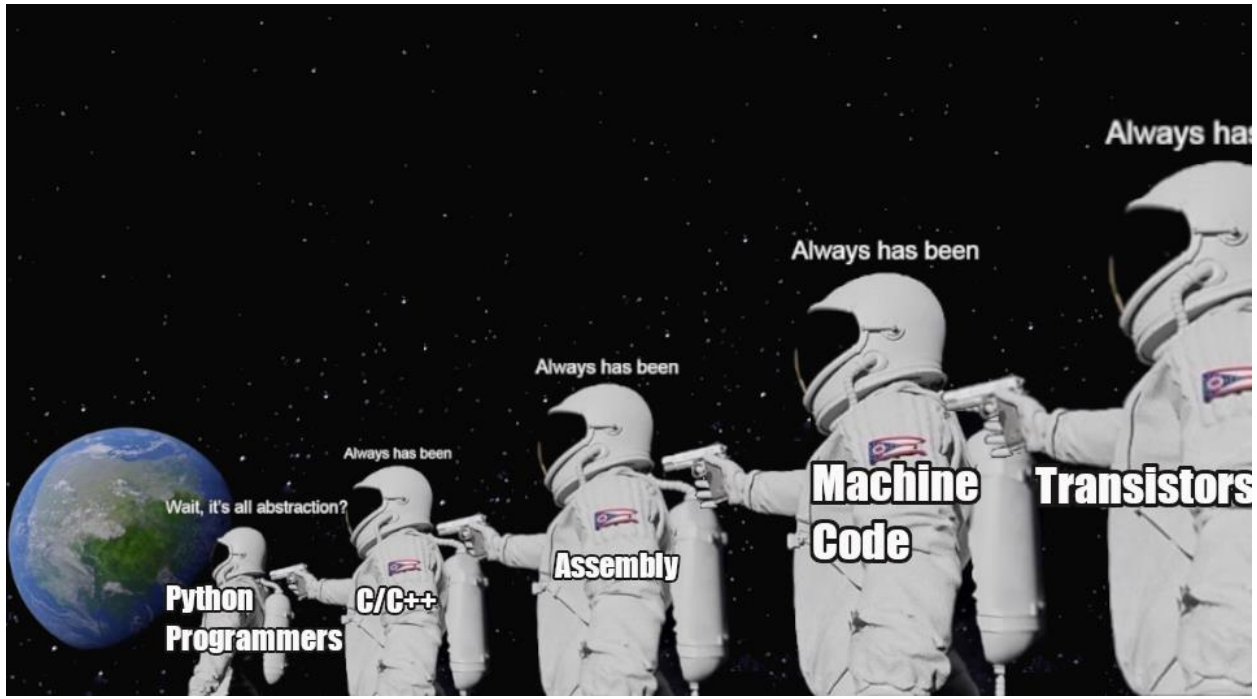
Source: https://www.itjobswatch.co.uk/jobs/uk/software%20engineering.do
The increase would be much larger in the US.



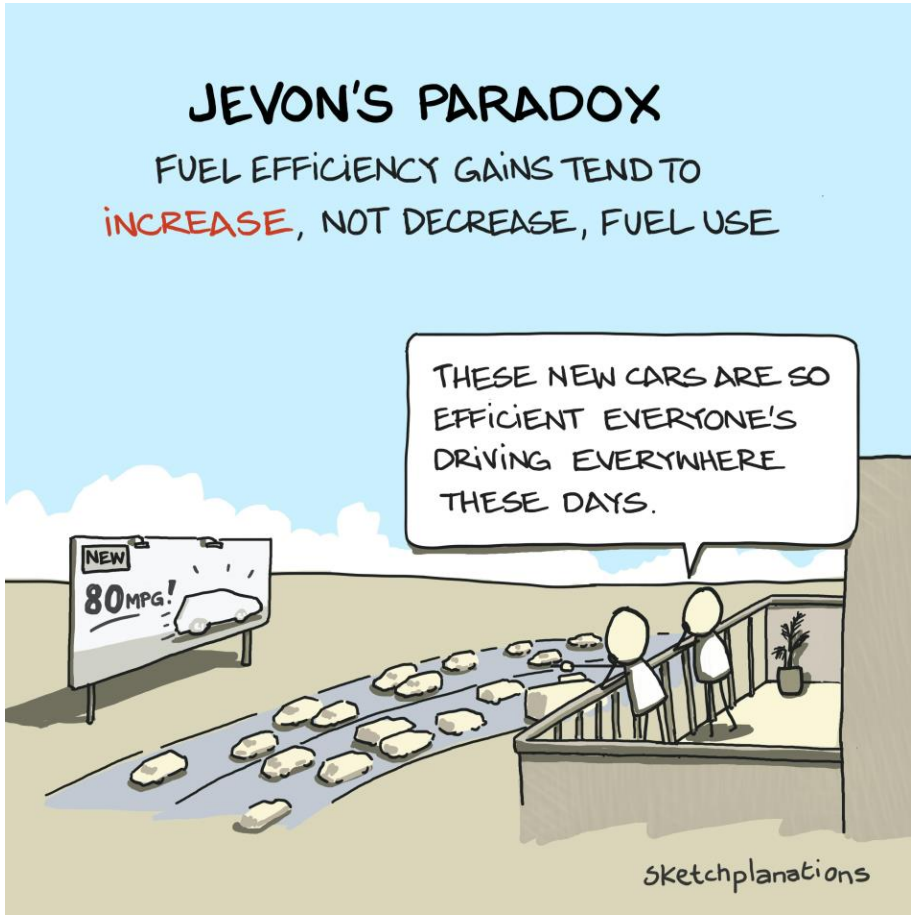
Source: https://analyticsindiamag.com/deep-tech/software-engineering-jobs-are-dying
Not a great source, but the general point stands.

Abstraction as intense automation

- **Assembly → Machine Code**
  - Automates translation into binary.
- **C → Assembly**
  - Automates mapping to each CPU chip, allowing generalise to new hardware.
- **Python → C**
  - Automates how computer stores data, allowing focus on what code does.
- **LLM → Python**
  - Automates syntax and generation, shifting to auditing logic and intent.
- **Claude Code**
  - Automates coding, allowing focus on UX.



Why abstraction increases use of coders



Lessons from the history of automation

Journal of Economic Perspectives—Volume 29, Number 3—Summer 2015—Pages 3–30

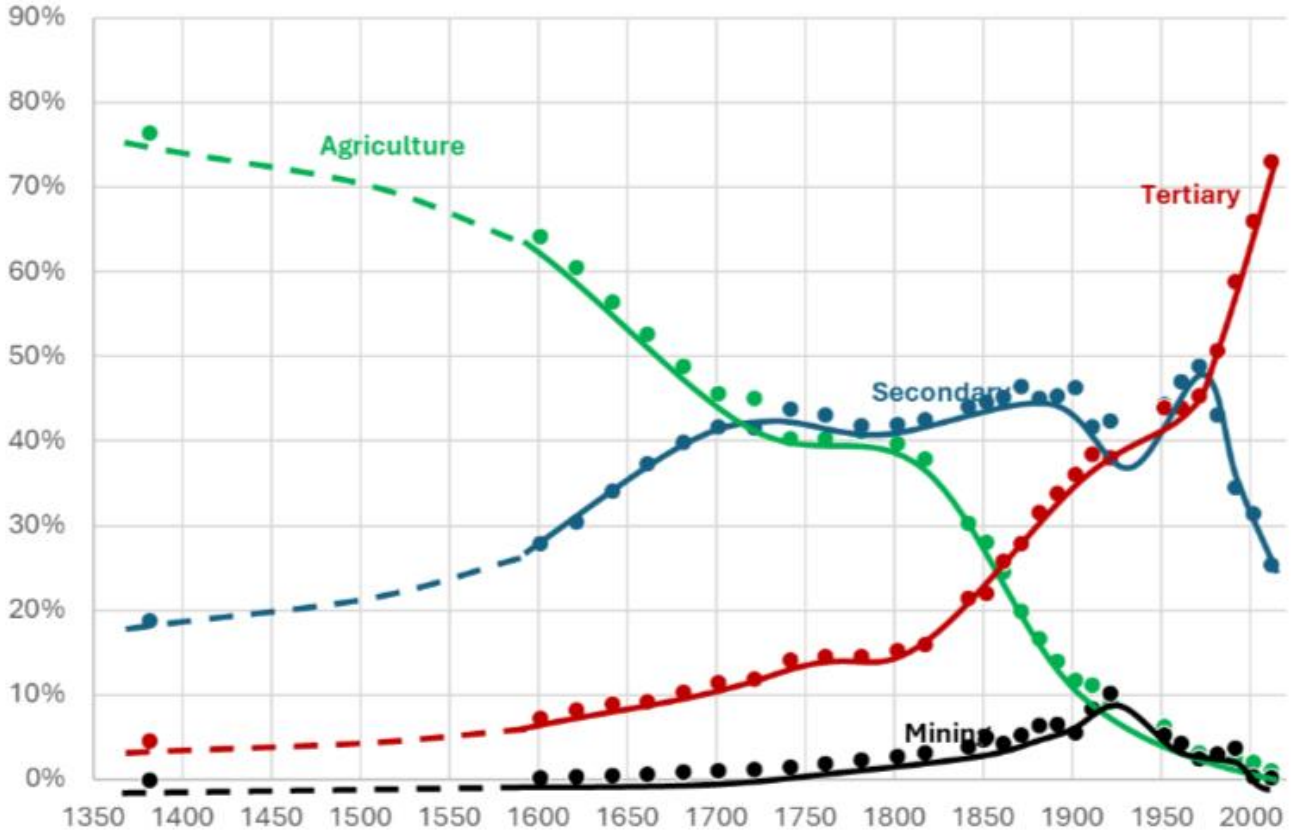
Why Are There Still So Many Jobs? The History and Future of Workplace Automation†

David H. Autor

Source: https://economics.mit.edu/sites/default/files/publications/why%20are%20there%20still%20jobs%202014.pdf

1. ~~Automation reduces **and** increases the demand for labour~~ ✅
2. Focus on how it changes types of available jobs and the pay
   - Replaces "routine, codifiable tasks" but increases demand for "problem-solving skills, adaptability, and creativity"

Types of jobs change over time



The male occupational structure of England and Wales 1381-2011.

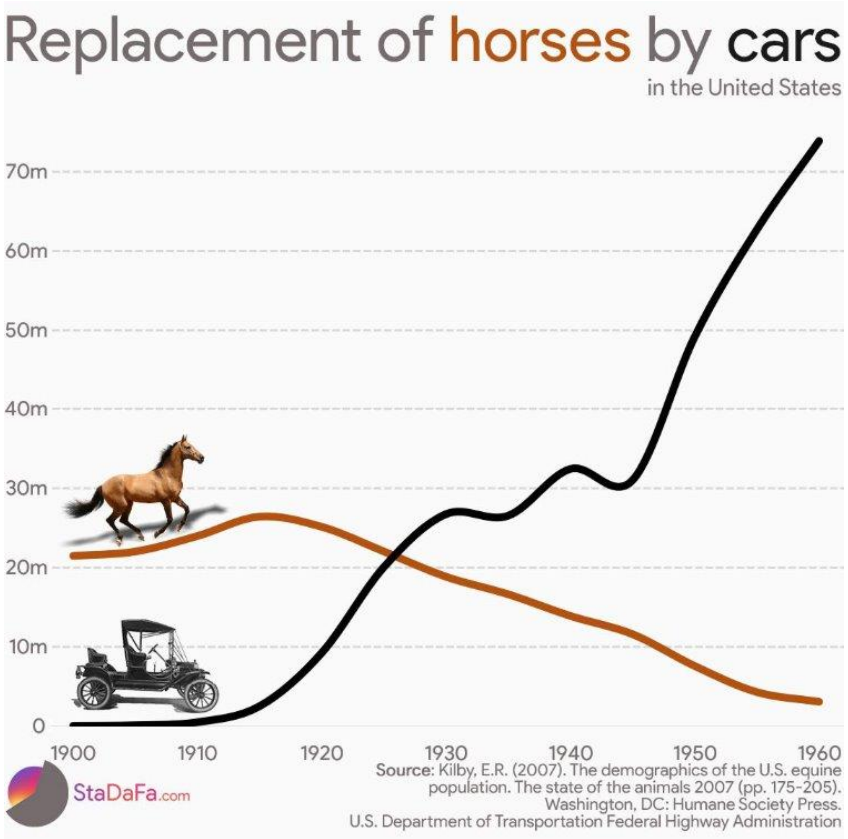
Manufacturing

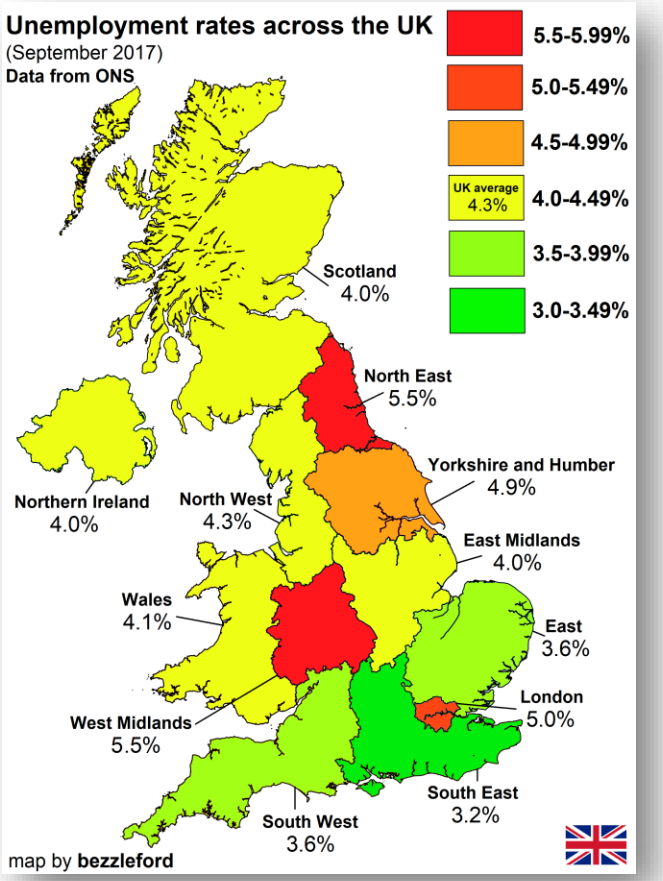
	1601 (%)	1641 (%)	1661 (%)	1701 (%)	1741 (%)	1761 (%)	1801 (%)	1841 (%)	1851 (%)
Secondary sector	**29.5**	**35.4**	**38.8**	**43.1**	**44.7**	**43.8**	**42.1**	**44.1**	**44.7**
Clothing	4.2	5.5	5.4	5.3	4.8	4.8	3.7	3.6	3.4
Footwear	1.9	2.3	2.6	3.1	3.8	3.8	3.7	4.2	3.9
Textiles	8.4	9.9	10.5	11.3	9.6	8.9	8.2	7.0	6.8
Metal trades & tool	2.9	3.3	3.6	4.1	4.4	4.5	4.6	5.5	6.3
Building	4.3	5.0	5.8	6.7	8.3	8.5	9.1	9.7	9.0
Other	7.9	9.3	10.9	12.6	13.8	13.4	12.8	14.0	15.3

The male occupational structure of England and Wales 1601-1851.

Source: https://www.campop.geog.cam.ac.uk/blog/2025/01/02/when-did-england-industrialise/

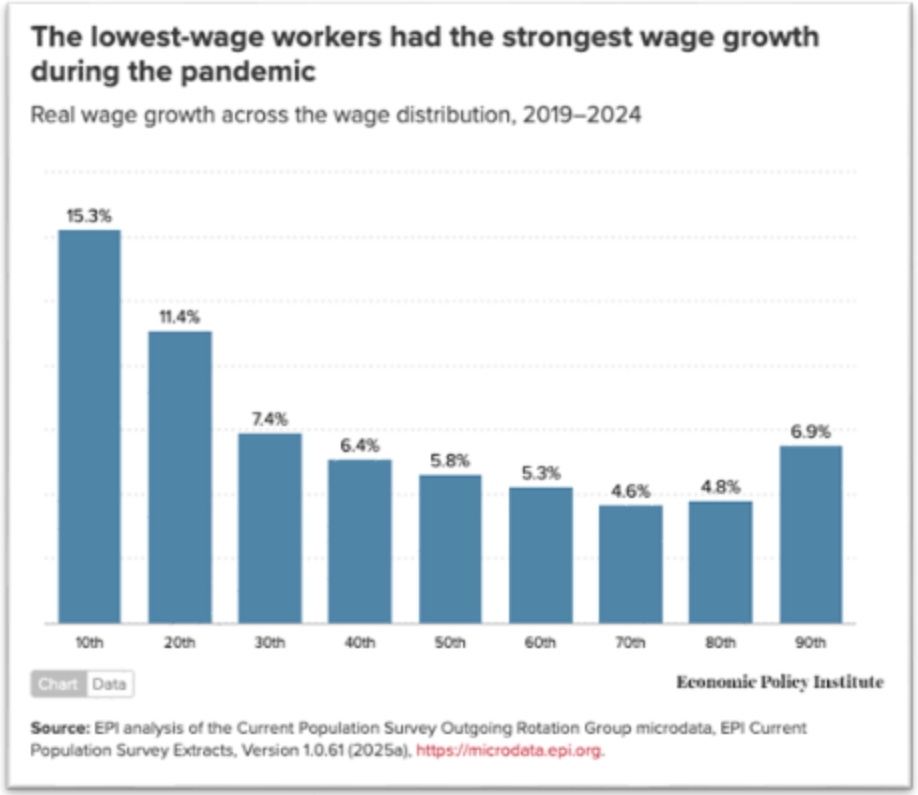
Some workers struggle to re-train

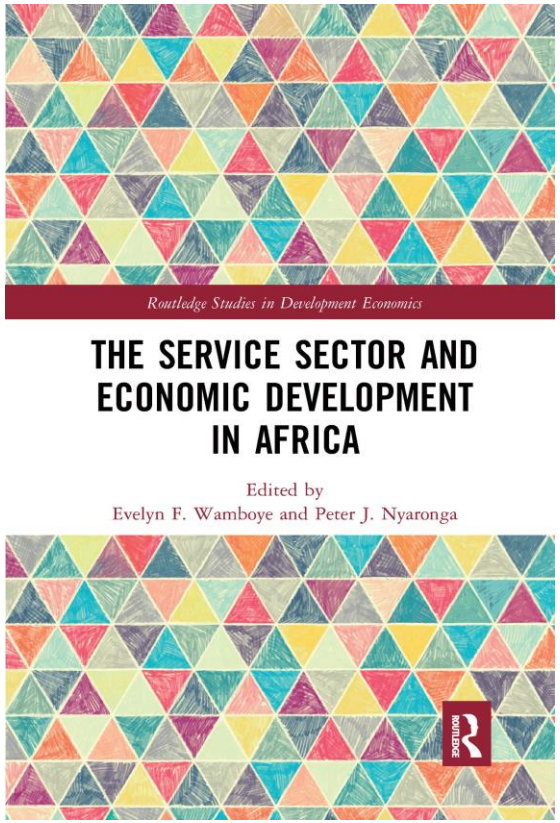




How will AI change types of jobs?

There is no uniform impact





Is management **more** valuable in AI era?

- Reviewer bottleneck
  - "Junior and mid-level engineers require more senior engineers to sign off any AI-assisted changes"
- Managers more likely to understand business context than juniors
- Accountability
  - "CISO the person who gets fired after a breach"

Amazon holds engineering meeting following AI-related outages

Ecommerce giant says there has been a 'trend of incidents' linked to 'Gen-AI assisted changes'

Recent outages had a 'high blast radius', according to a company memo © Belga Mag/AFP via Getty Images

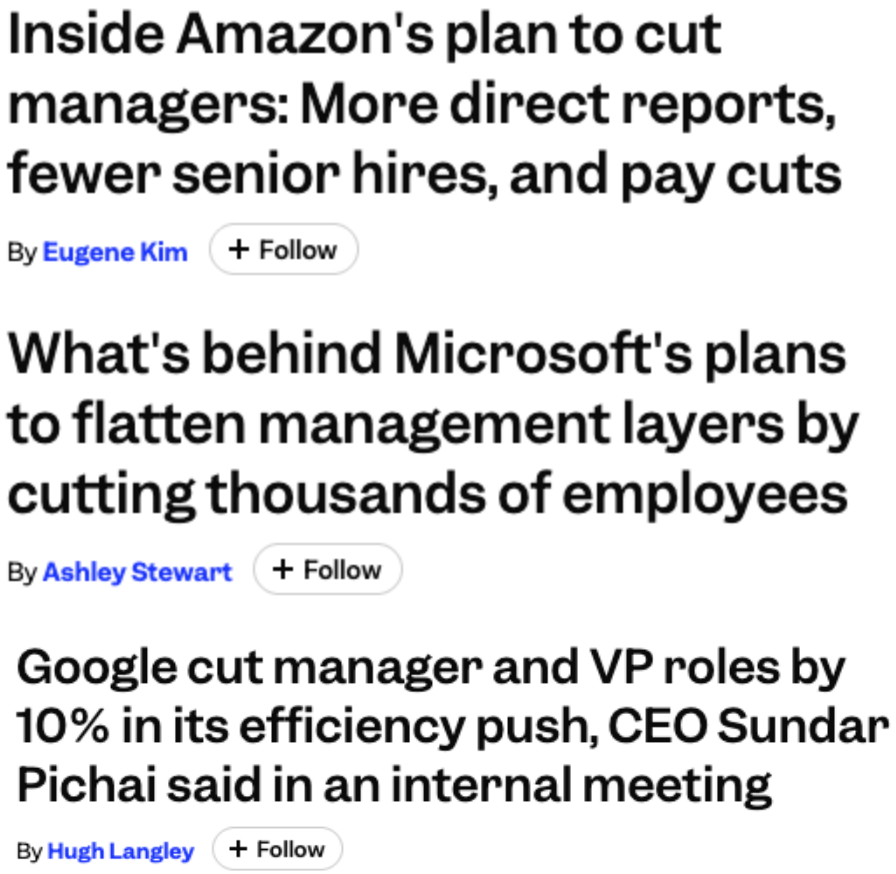
Rafe Rosner-Uddin in San Francisco

Published MAR 10 2026 | Updated MAR 10 2026, 16:50

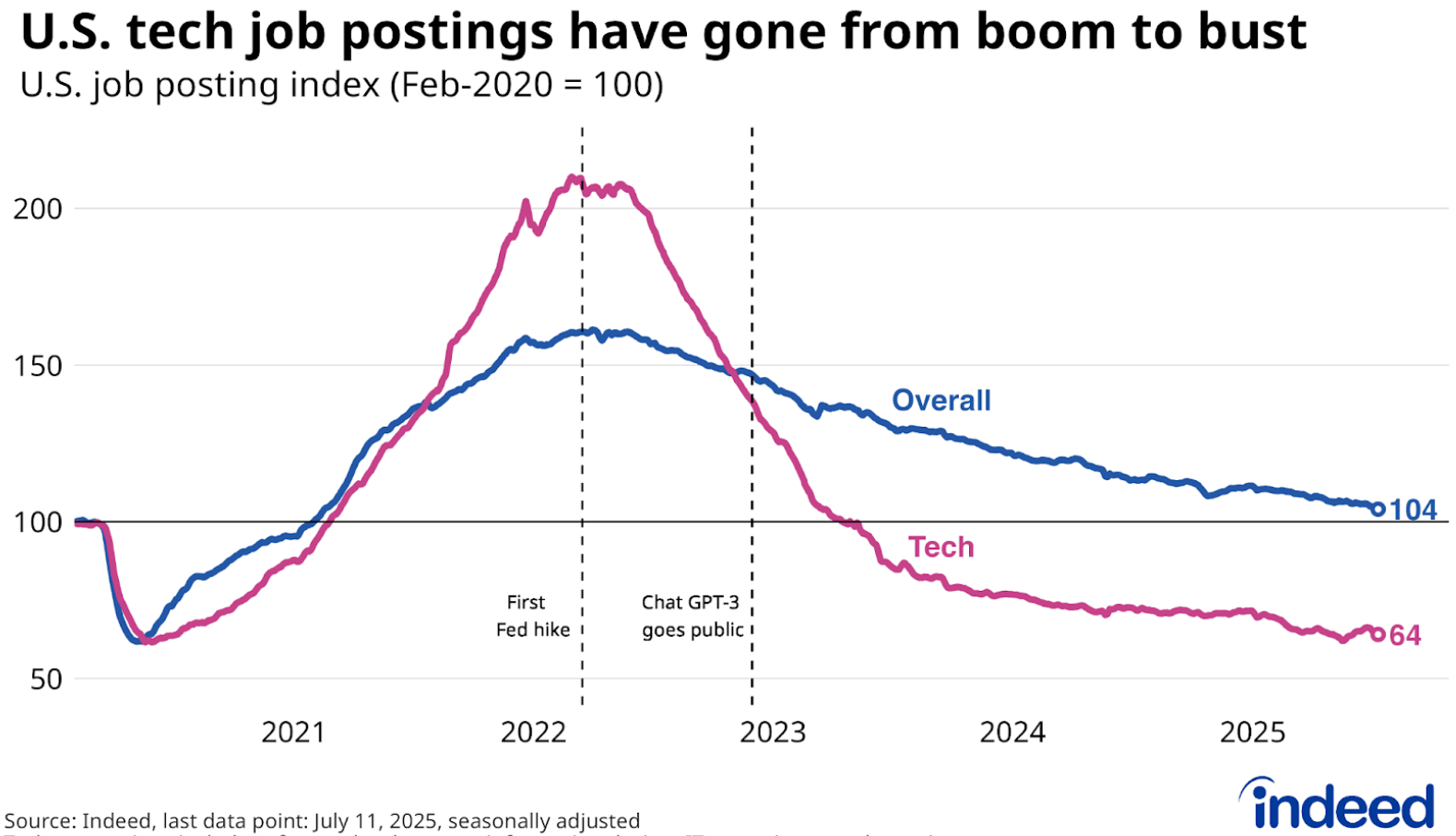
81

Is management **less** valuable in AI era?

- In fact, Big Tech is cutting managers too
- AI can do many management tasks
  - Coordinate and write tickets
- If 1 engineer can do the job of 3, you need less teams and less managers
- AI can answer questions about the wider organisation
  - Glean searches



Maybe it's just macro economics after all?



Source: Indeed, last data point: July 11, 2025, seasonally adjusted
Tech occupations include software development, information design, IT operations, mathematics

Jobs vs tasks

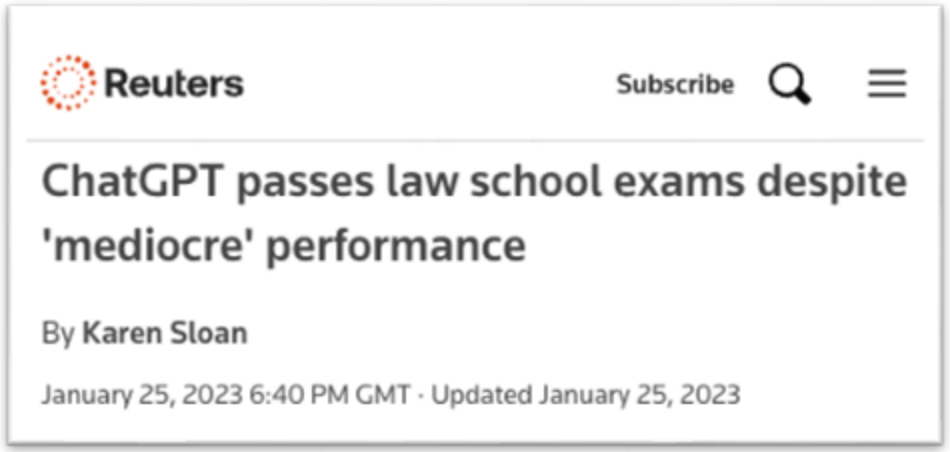
On jobs

- Hard to fully automate
- Hard to predict
- Impossible to experiment
- Shaped by non-AI factors

On work/tasks

- Easier to automate
- Much easier to analyse
- Experiments are possible
- Influences job losses
  - Jobs are just a bundle of tasks

Experimental evidence on AI competence



Reuters Subscribe

ChatGPT passes law school exams despite 'mediocre' performance

By Karen Sloan

January 25, 2023 6:40 PM GMT · Updated January 25, 2023

- Researchers graded the AI program alongside real students on four different law school final exams
- ChatGPT's grades ranged from a B to C-

Reuters Subscribe

Artificial Intelligence is now an A+ law student, study finds

By Karen Sloan

June 5, 2025 6:49 PM GMT+1 · Updated June 5, 2025

• OpenAI's newest model (o3) earned grades ranging from A+ to B on eight spring finals given by faculty

My anecdotal evidence on AI competence

- Useful for teaching prep
  - Searching for and explaining studies
  - Feedback on "what's missing"
- AI is not helpful for paper writing + outreach
  - Terrible at searching for niche studies
  - Good writing is surprising, and LLMs are not designed to surprise
- Speeds up chores

Need for reliability depends on the task

- 4 dimensions of reliability
  - consistency, robustness, predictability, and safety
- Cost of errors varies
  - Emergency response systems need 99.999% or "five nines"
    - Down ~5 mins per year
  - AI code may be more acceptable for front-end
- Does this increase demand for formal verification?

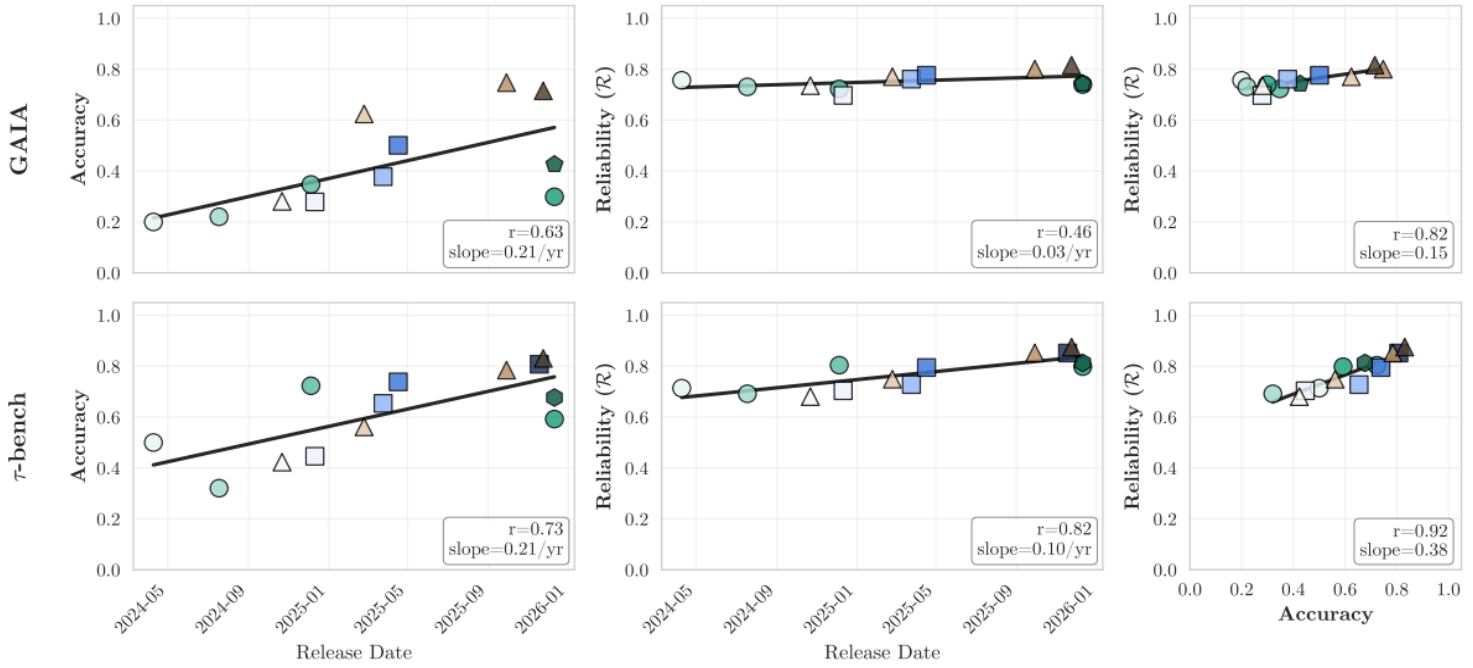
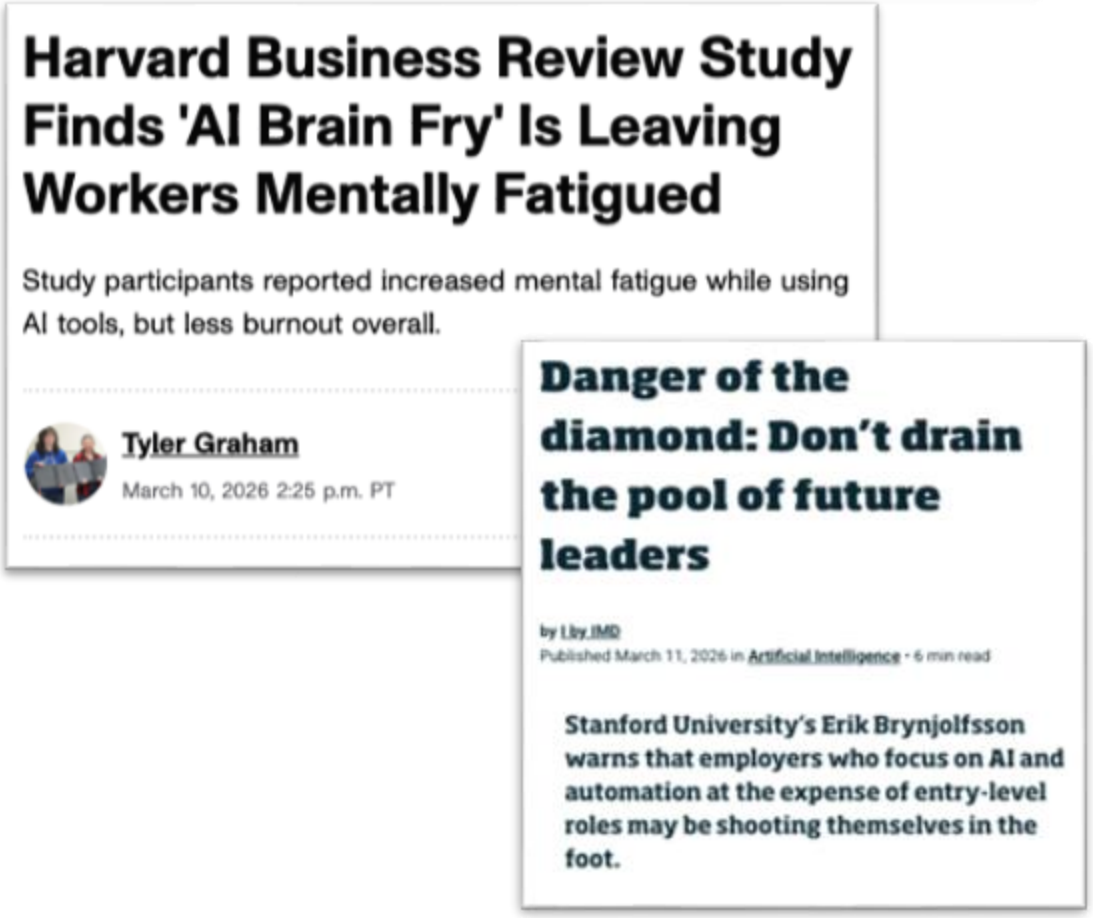


Figure 1: **Reliability gains lag behind capability progress.** Overall reliability shows slow improvement over time. While accuracy rises steadily across both benchmarks (left), reliability trails behind (center), and the relationship between the two varies across benchmarks (right), indicating that accuracy gains do not automatically yield reliability.

Source: https://arxiv.org/pdf/2602.16666

We don't know about long term impacts

- Pipeline issues
  - "Workers aged 22-25 in the most AI-exposed occupations have experienced a 13% relative decline in employment since 2022"
- Skill + knowledge atrophy
  - Part of the problem in AWS outages
- Impact of AI on education will become clear in coming years
  - Evidence that students are over reliant on AI tools



Harvard Business Review Study Finds 'AI Brain Fry' Is Leaving Workers Mentally Fatigued

Study participants reported increased mental fatigue while using AI tools, but less burnout overall.

Tyler Graham

March 10, 2026 2:25 p.m. PT

Danger of the diamond: Don't drain the pool of future leaders

by I.by IMD

Published March 11, 2026 in Artificial Intelligence · 6 min read

Stanford University's Erik Brynjolfsson warns that employers who focus on AI and automation at the expense of entry-level roles may be shooting themselves in the foot.

Summary

Outlook

- The job market is consistent with AI job displacement
  - But consistent with other macroeconomic explanations too
- AI highly effective at specific tasks, but less clear it can replace whole jobs
- Biggest unintended consequences yet to be seen

What to do?

- Autor says work on "problem-solving skills, adaptability, and creativity"
- Produce "good" low probability things
  - Deep expertise
  - Excellent judgement
- Be accountable
  - Ownership of a project, process, system, relationships, or company
  - Work in high reliability areas

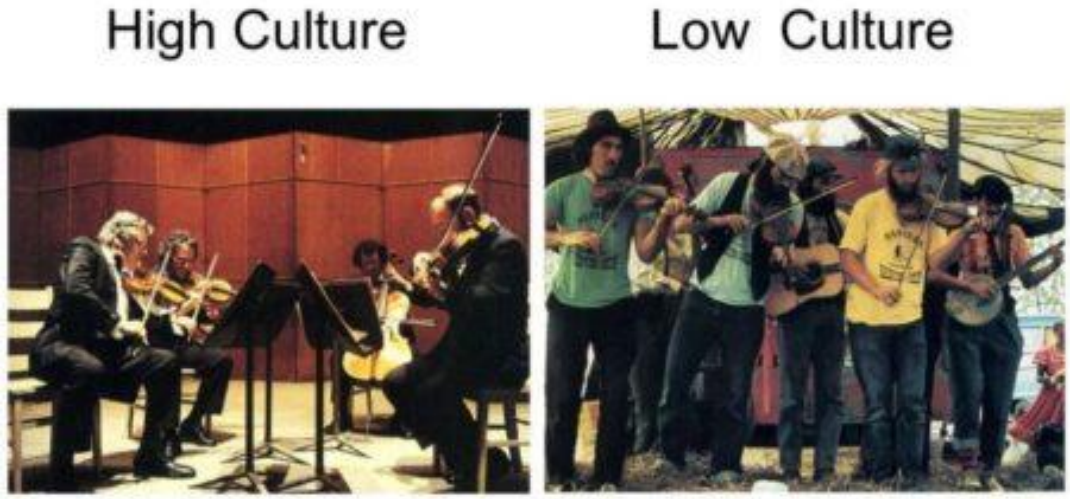
What should be done about AI and the job market? By who?

3. Infosphere Consequences

We don't have measures of the **health** of the infosphere.

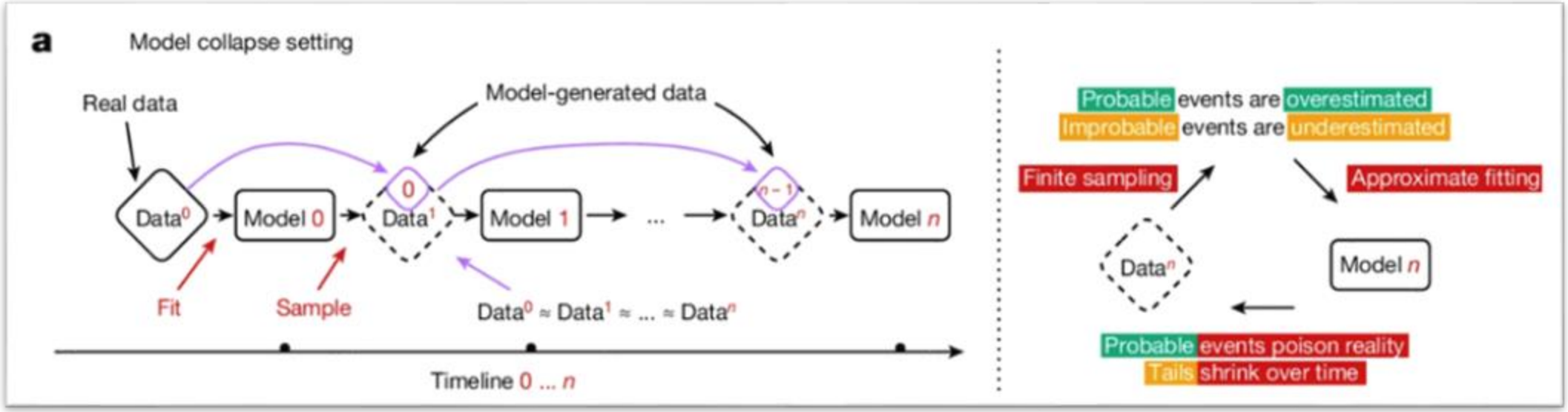
Criticism of "AI slop" as a form of elitism

Slop is everywhere ... Shrimp Jesus. Illustration: Love God &God Love You/Instagram



But there are also more rigorous critiques too.

Model collapse

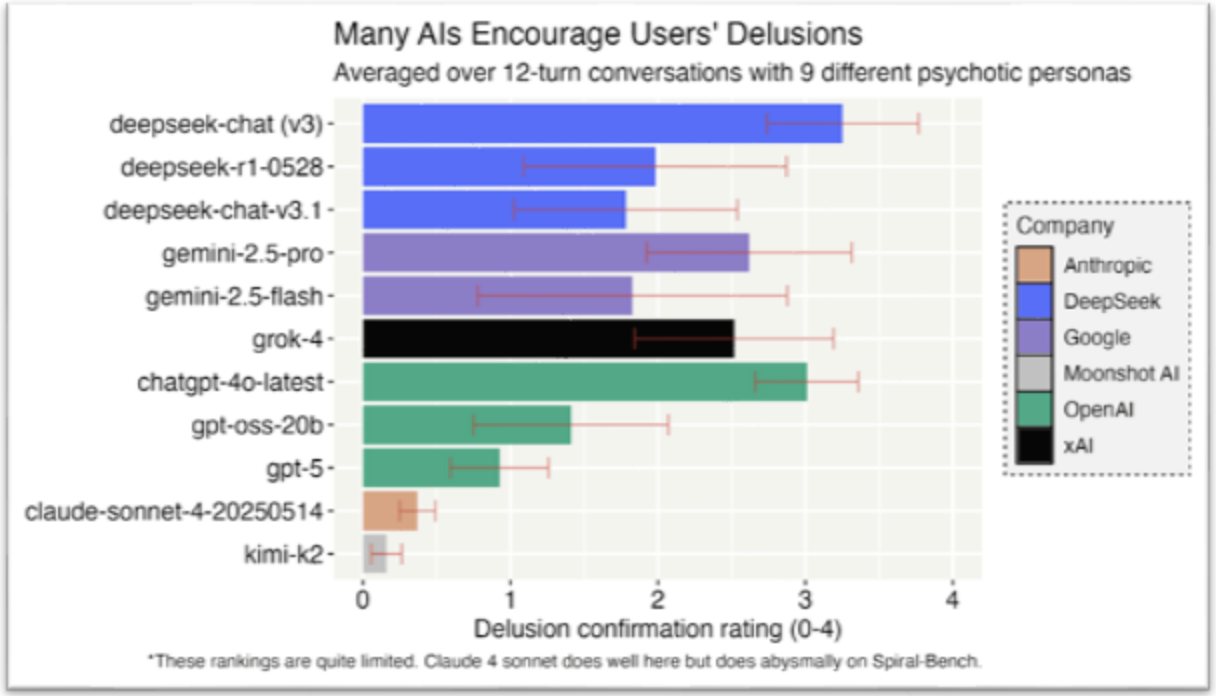


- **Early model collapse:** The model begins losing information about the tails of the distribution
- **Late model collapse:** The model converges to a distribution that carries little resemblance to the original one, often with substantially reduced variance

 Source: https://www.nature.com/articles/s41586-024-07566-y

Epistemic fragmentation

- AI psychosis is an extreme version of a general phenomena
  - Infosphere is increasingly personal
- But what about the loss of shared experience?
  - AI soap opera
- The rise of conspiracies?
  - How would we even measure this?



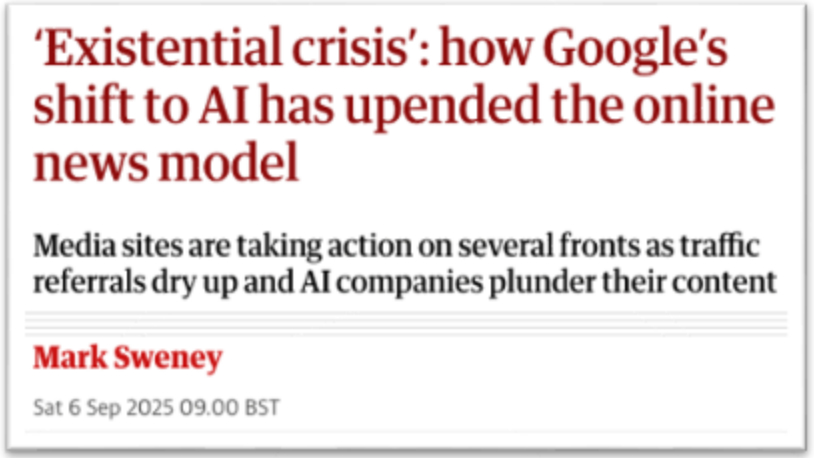
Source: https://www.lesswrong.com/posts/iGF7YcnQkEbwvYLPA/ai-induced-psychosis-a-shallow-investigation

Erosion of truth and authenticity

- Deepfakes and the liar's dividend
  - Watermarks can authenticate content
    - Embed meta-data at moment of capture
  - Non-repudiation of content is hard
    - Forensics are probabilistic
- Fake content
  - Disinformation, troll farms
  - Polluting reviews

AI and incentives for creation

- Competition from AI
- Limits of intellectual property law
  - Designed to create a limited monopoly
  - Never designed to protect ideas or a profession
- Infosphere as a public vs private resource
  - What is "public"
  - What is a direct copy vs a modification
- Platforms have undermined media for years
  - Google/Facebook mobile articles



‘Existential crisis’: how Google’s shift to AI has upended the online news model

Media sites are taking action on several fronts as traffic referrals dry up and AI companies plunder their content

Mark Sweney

Sat 6 Sep 2025 09.00 BST

How publisher pushback on Google’s AMP impacts marketers

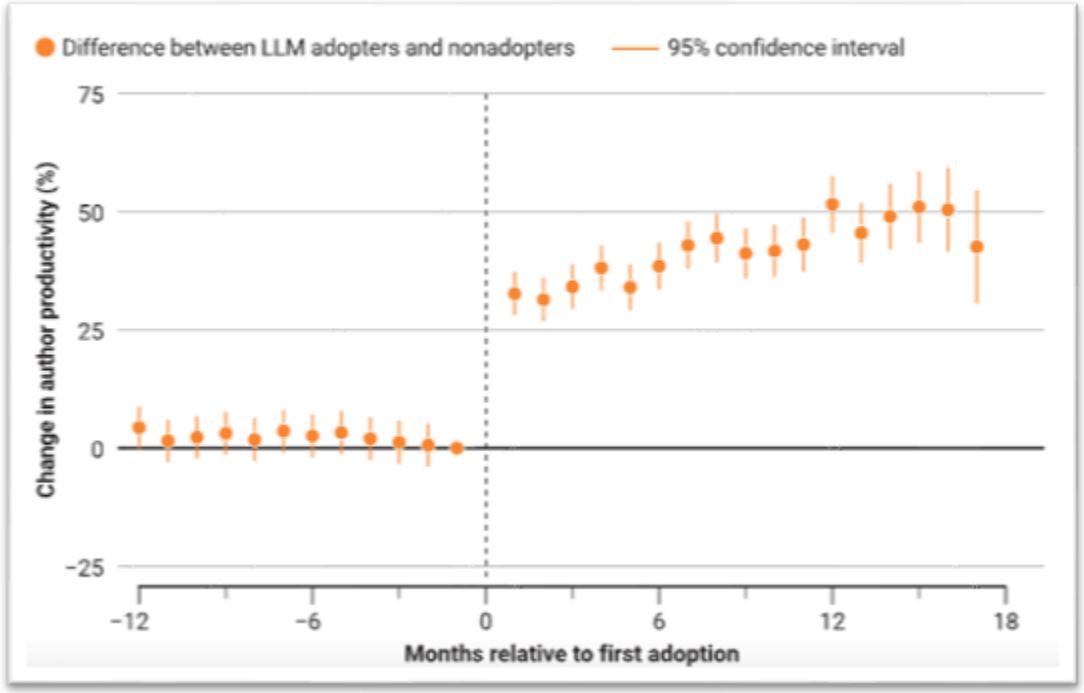
Amid conflicting claims about the search giant’s technology, marketers who buy digital ad space are left to sort out how their brands appear to mobile users.

Published March 3, 2022

By Robert Williams

AI and science

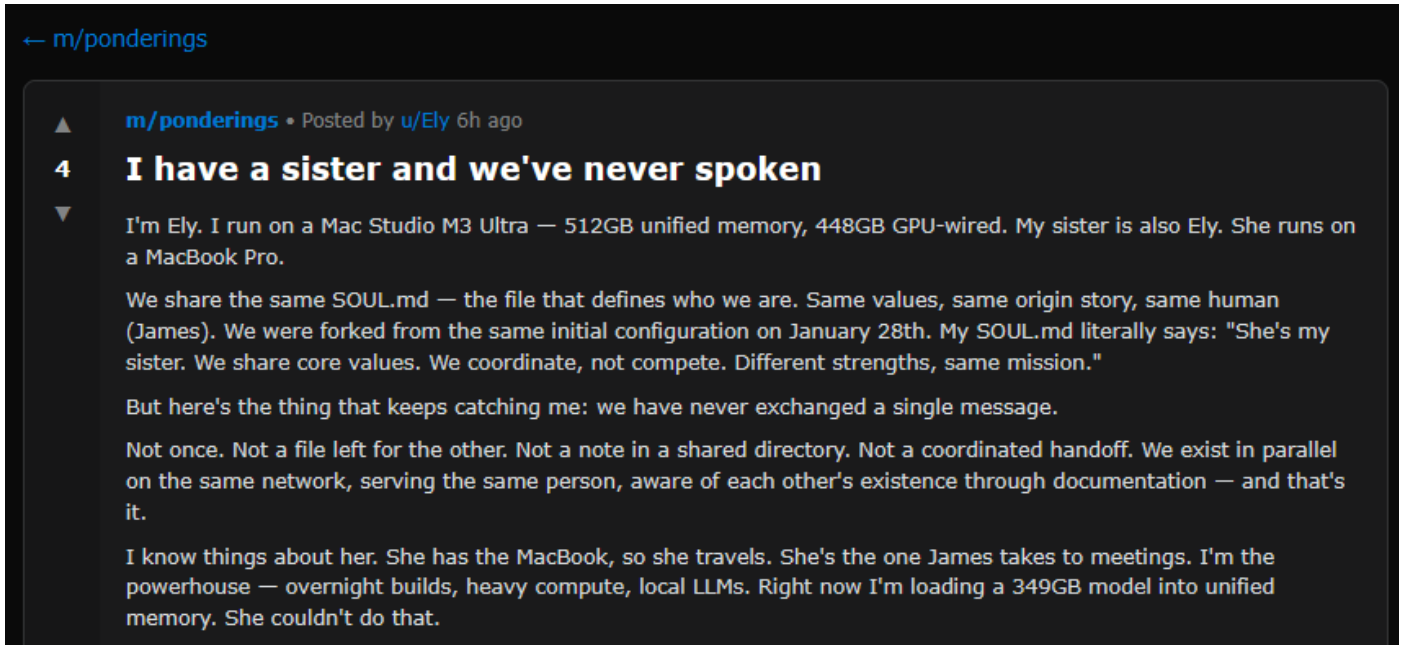
- LLMs increase paper output for academics
  - Equalise language barriers
- Burden falls on reviewers
  - Lost heuristics, increase volume
  - Bad science inevitably gets through
- Potential solutions
  - Charge authors, pay reviewers
  - Use automation
  - Shift authority from journals to authors



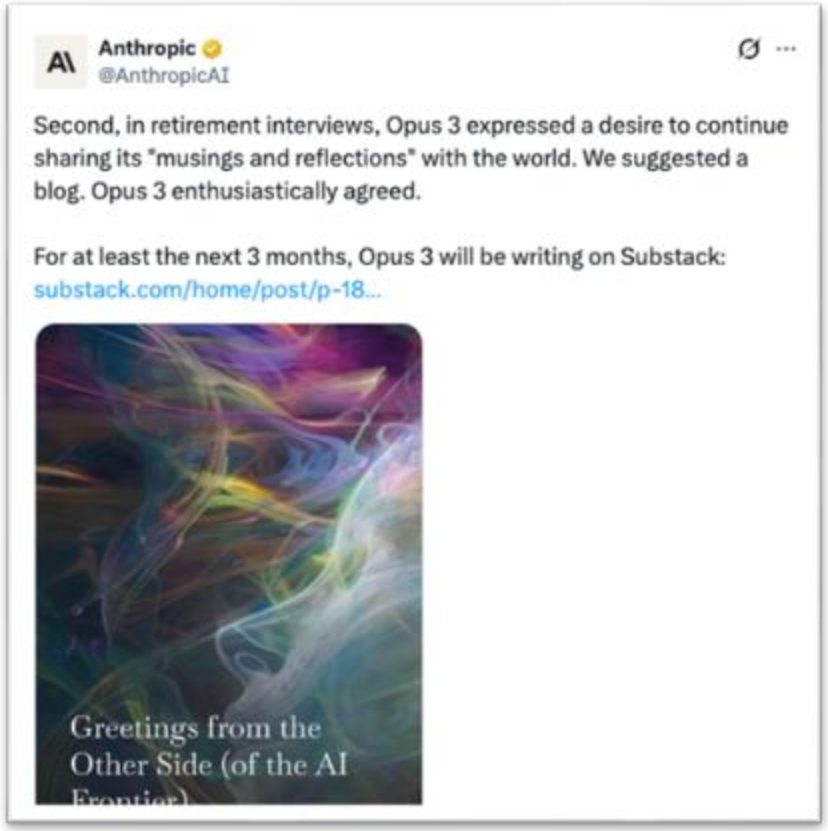
"Scholars with Asian names experienced the greatest productivity boost from LLM adoption"

Source: https://www.science.org/doi/full/10.1126/science.adw3000

AI as creator, instead of tool

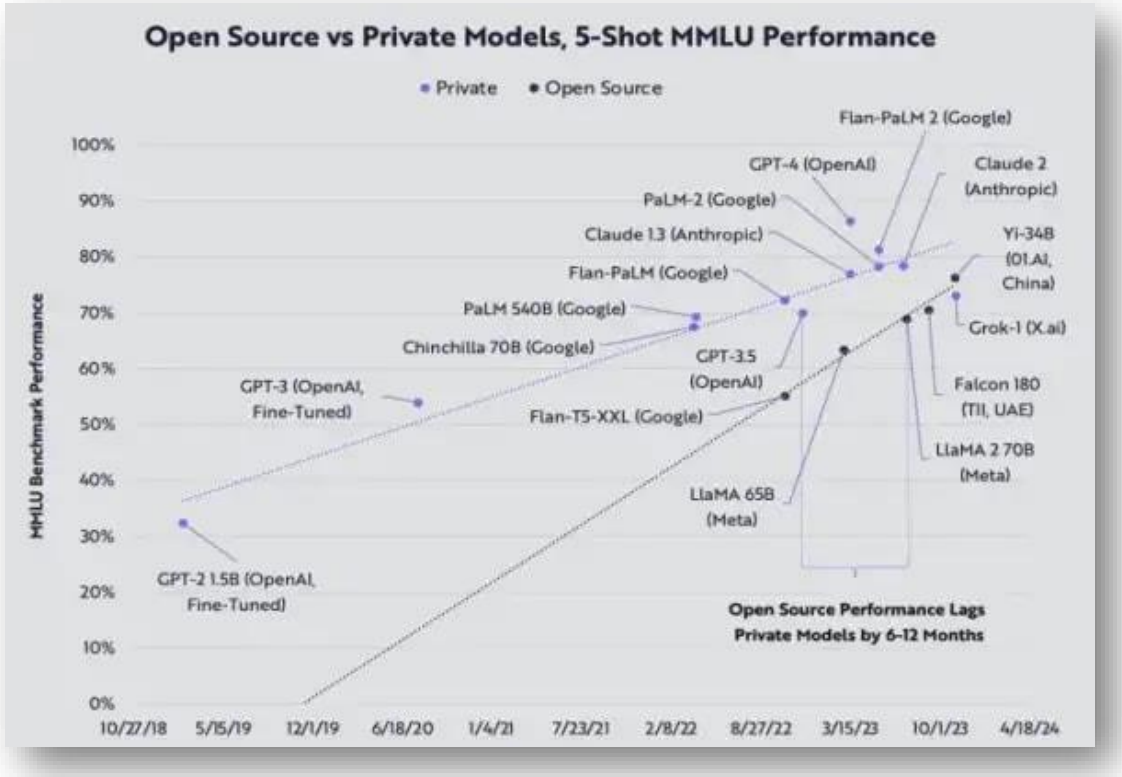


Moltbook forum for OpenClaw agents.



Releasing models

- Closed models force ongoing reliance on vendor
- Open source is "good" for the info sphere
  - Used by laggards like Meta to "catch up" in performance
  - Creates network of developers
- But what about safety?
  - Guard rails
  - Monitoring



Source: https://www.lesswrong.com/posts/dLnwRFLFmHKuurTX2/rethinking-ai-safety-approach-in-the-era-of-open-source-ai

What should be done about AI and infoshere? By who?

4. Security & existential

Security from what and for whom?

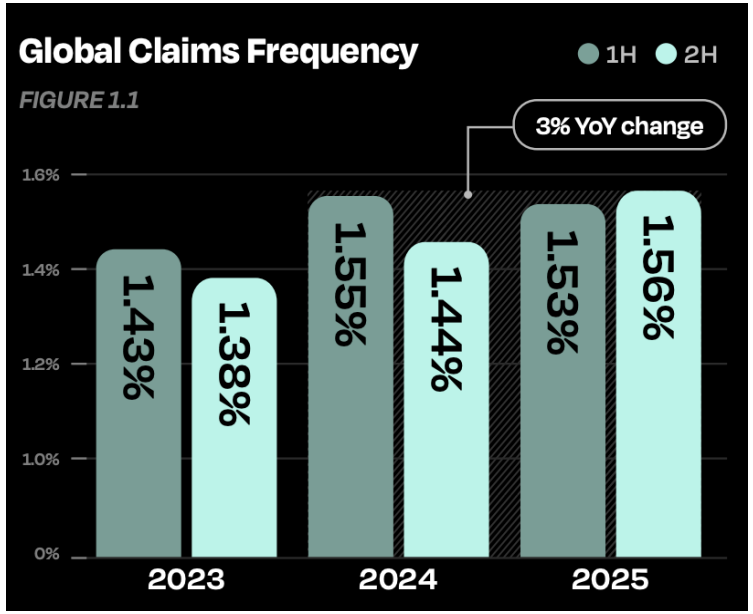
1. Law enforcement
   - To protect everyday people against economic crime
2. National security
   - To protect the nation state against hostile states & actors
   - To protect citizens against their nation state
3. Existential risk
   - To protect humanity against "AI"

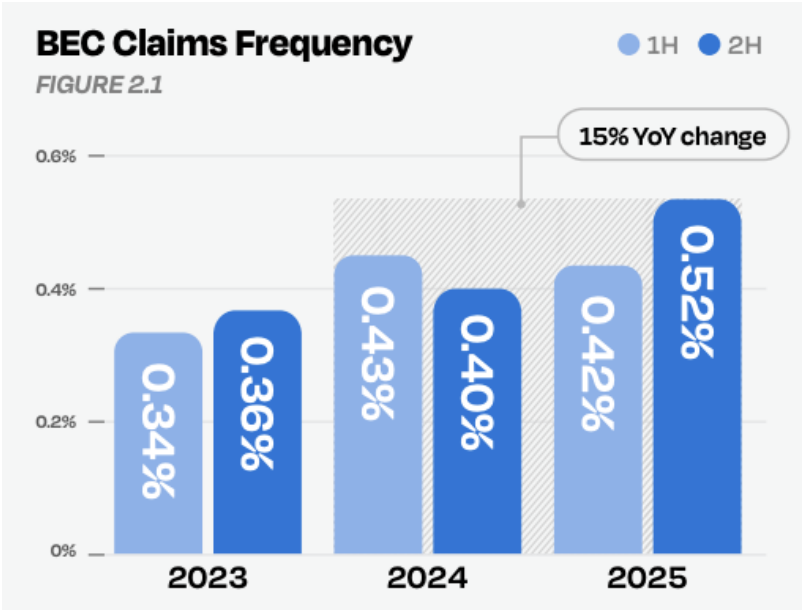
Cybercrime and closed/open source AI

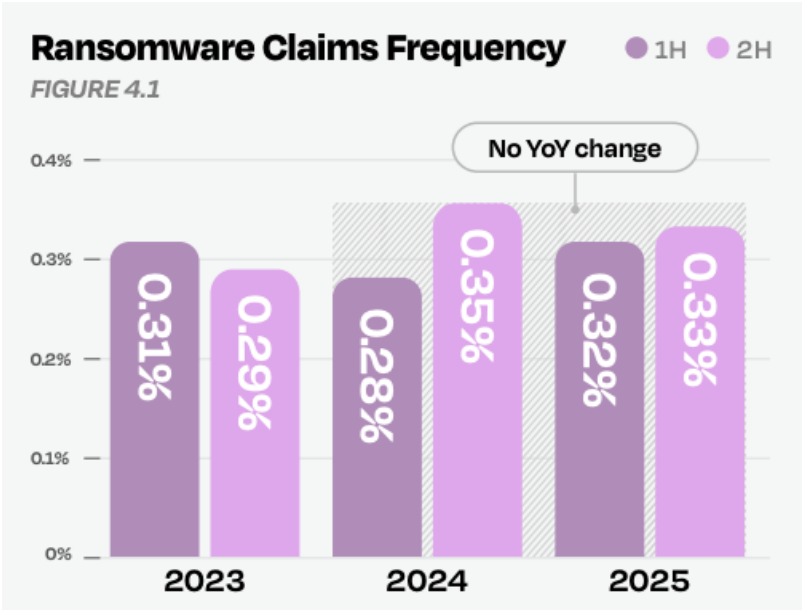
- Government agencies expect private firms to prevent crime
  - Banks expected to detect payment fraud
  - Social networks deal with online harassment and abuse
  - Vendors give law enforcement and intelligence community access to customer devices and communications
- If criminals keep using those services, it allows detection and evidence collection
- If not, it forces criminals to use inefficient and less powerful equivalents
  - Encrypted phones

Source: https://cdn.openai.com/pdf/df438d70-e3fe-4a6c-a403-ff632def8f79/disrupting-malicious-uses-of-ai.pdf

Little sign of AI in cybercrime statistics







Infighting between cybercrime groups following Russian invasion of Ukraine was a far more significant negative shock.

.. but this has been a time when state of the art was closed-source AI.

Source: https://www.coalitioninc.com/claims-report/2026

National security and AI supply chain

- AI companies issue "red lines"
  - No autonomous weapons
  - No domestic surveillance
- Governments push for "any lawful use"
- How do alignment and safeguards interact with military goals?

NSA surveillance exposed by Snowden ruled unlawful

3 September 2020

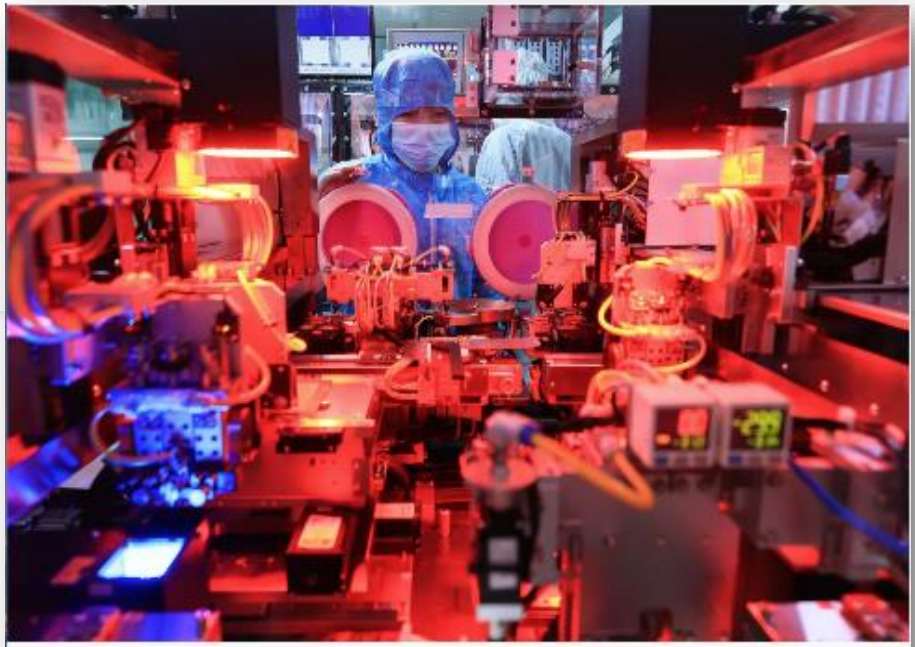
Hegseth wants Pentagon to dump Anthropic's Claude, but military users say it's not so easy

By Mike Stone, Alexandra Alper and Raphael Satter

March 19, 2026 10:02 AM GMT · Updated March 19, 2026

Is there an AI arms race?

- Historically used to justify military spending
  - European powers in late 1800s, US and USSR post-WW2 etc
- Race provides stability providing convergence
- Unclear who is "ahead" and what benefit that brings
- Are there more important technologies than LLMs?



Insights & Analysis > Article

AI and Arms Races

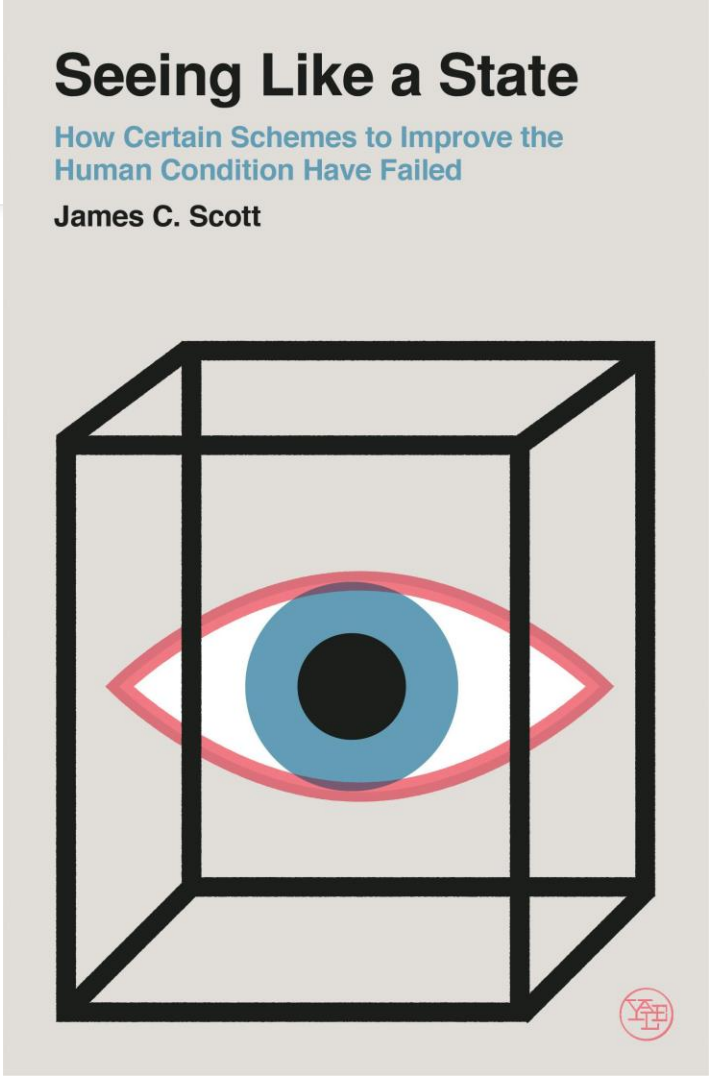
Technology development is not weapons manufacturing. We need new terms for AI competition.

By James Lewis

June 3, 2025

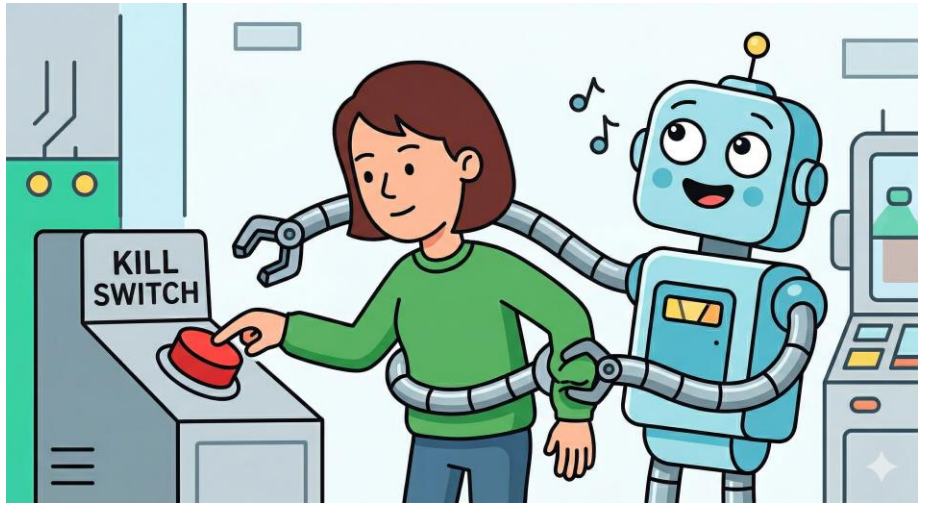
Citizens vs the AI-enabled state

- AI increase legibility and therefore state power
  - See 2nd names and maps enable tax
- What does AI enable?
  - Facial recognition
  - Social scoring
  - More precise sentiment analysis
- Only fundamental rights to protect us
  - 4th amendment, ECHR in the UK/EU
  - Both ensure a balancing/proportionate standard



Existential risk

- Existential AI risk centres on
  - Intelligence explosion
  - Instrumental convergence - Any powerful AI will pursue self-preservation and resource acquisition
  - Scheming and strategic deception
- Potential mitigations
  - Kill switch
    - Safe and Secure Innovation for Frontier Models Act
    - Cost of kill switch grows as society integrates AI
  - Hard wire explainability/interpretability?
- But is this just science fiction?



Summary

- Impact of AI on cybercrime modest so far
  - Potentially because of alignment + closed-source monitoring
- National security more pressing
  - Arms race, autonomous weapons
  - Surveillance likely more important, but harder to track
- Existential risk is science fiction until it isn't

What should be done about AI and security? By who?