VNU-HUS MAT1206E/3508: Introduction to AI

Introduction
In-class Discussion

Hoàng Anh Đức

Bộ môn Tin học, Khoa Toán-Cơ-Tin học
Đại học KHTN, ĐHQG Hà Nội
hoanganhduc@hus.edu.vn

Contents

What is Artificial Intelligence (AI)?

What this section covers

There is no single agreed definition of AI, so we approach it from several angles:

- *Classic definitions* (McCarthy, Encyclopedia Britannica, Elaine Rich, Turing) — and why none is fully satisfactory.
- *A framework* that organises them: Russell & Norvig's *four approaches* and the *rational agent*.
- *The Turing test* — and whether today's large language models pass it (and what that really shows).
- *Searle's Chinese Room*: behaving intelligently vs. genuinely understanding.
- *Weak/strong* and *narrow/general (AGI)/super (ASI)* AI.
- *AI today*: foundation models, generative AI, and agentic AI.

Goal

By the end you should be able to *place any system* on these maps and argue *how you would define AI*.

What is Artificial Intelligence (AI)?

At a very first glance,

- Humans have *real* intelligence
- Machines have *artificial* intelligence

Basic Idea

Artificial intelligence is based on the assumption that the process of human thought can be *mechanized*

- Human thought can be described as a mechanical sequence of steps that machines can execute.
- This implies that machines can be designed to simulate human cognitive processes, such as *adaptation, learning, and decision making*.

Note

Indeed, *there is no officially agreed definition of AI*. It is thus not easy to categorize what is AI and what is not

What is Artificial Intelligence (AI)?

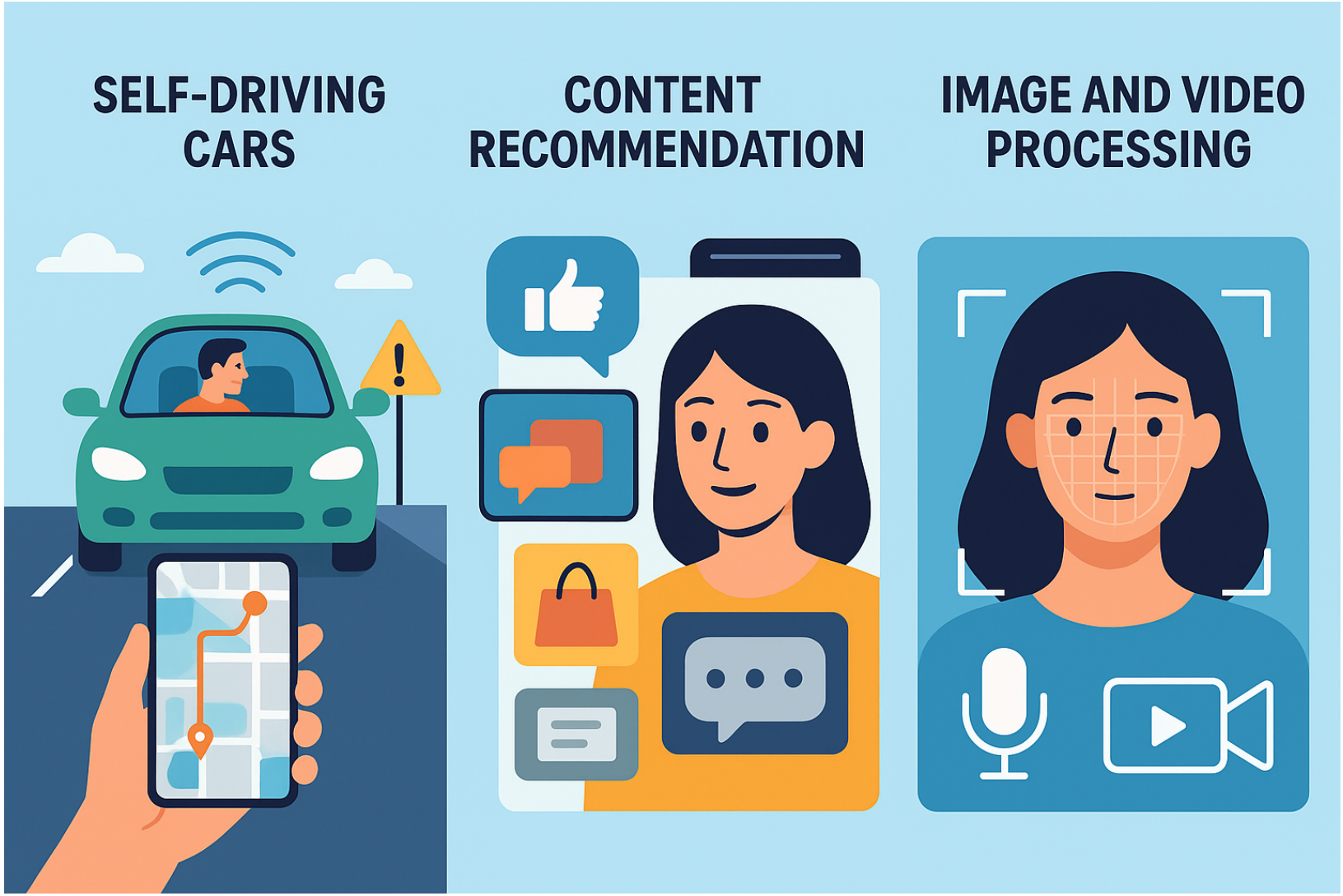


Figure: Some real-world applications which combine several "AI-techniques". Image generated by ChatGPT

What is Artificial Intelligence (AI)?

What do you think?

John McCarthy (1955) [one of the pioneers of AI]

The goal of AI is to develop machines that behave as though they were intelligent

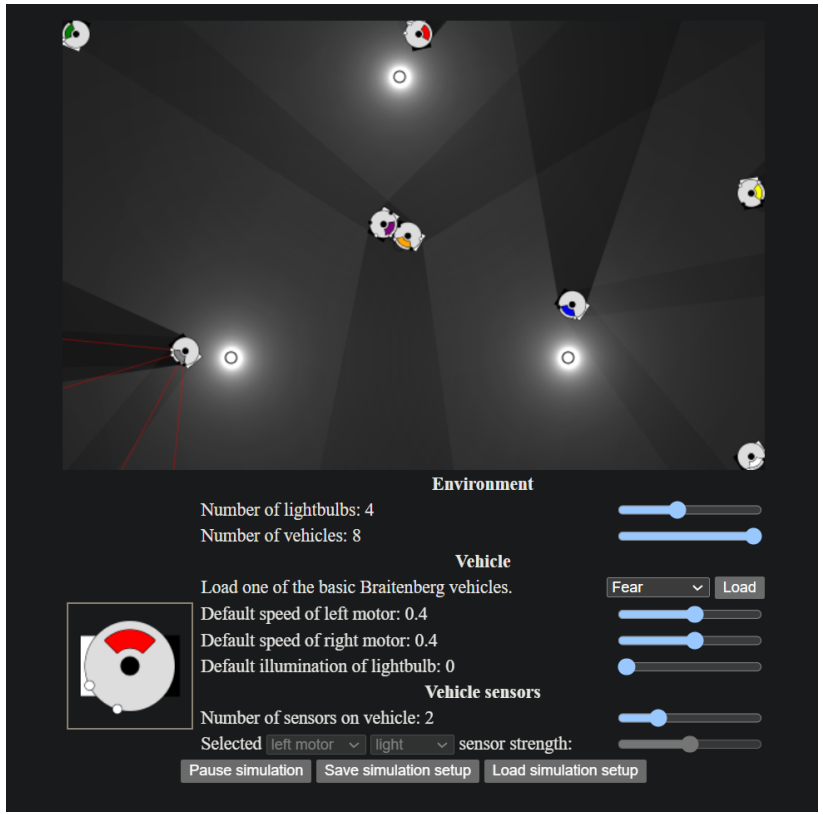


Figure: Example of Braitenberg vehicles simulating simple behaviors. (See http://www.harmendeweerd.nl/braitenberg-vehicles/)

What is Artificial Intelligence (AI)?

What do you think?

Encyclopedia Britannica (1991)

AI is the ability of digital computers or computer controlled robots to solve problems that are normally associated with the higher intellectual processing capabilities of humans ...

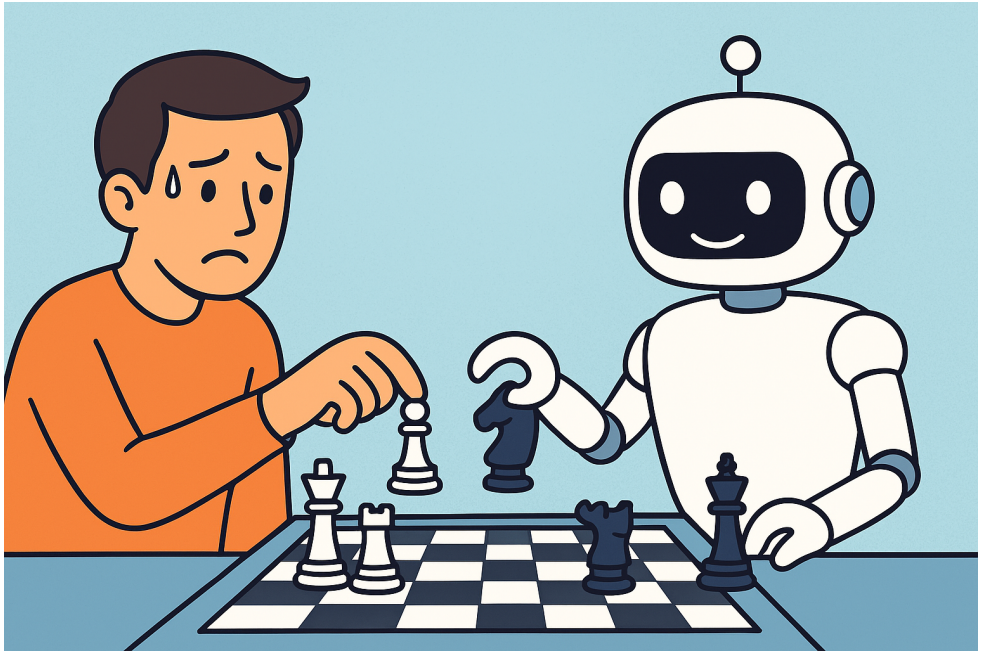


Figure: Computers can now beat humans in complex games (e.g., chess, Go), yet robots still struggle with dexterous tasks such as grasping and precise object placement — an instance of *Moravec's paradox* (1988). Image generated by ChatGPT

What is Artificial Intelligence (AI)?

What do you think?

Elaine Rich [Rich 1983]

Artificial Intelligence is the study of how to make computers do things at which, at the moment, people are better

Figure: AI systems can now outperform humans in tasks such as playing chess, yet they still have limitations in true understanding and general reasoning. (Image generated by Grok)

What is Artificial Intelligence (AI)?
What do you think?

Alan Turing [Turing 1950]

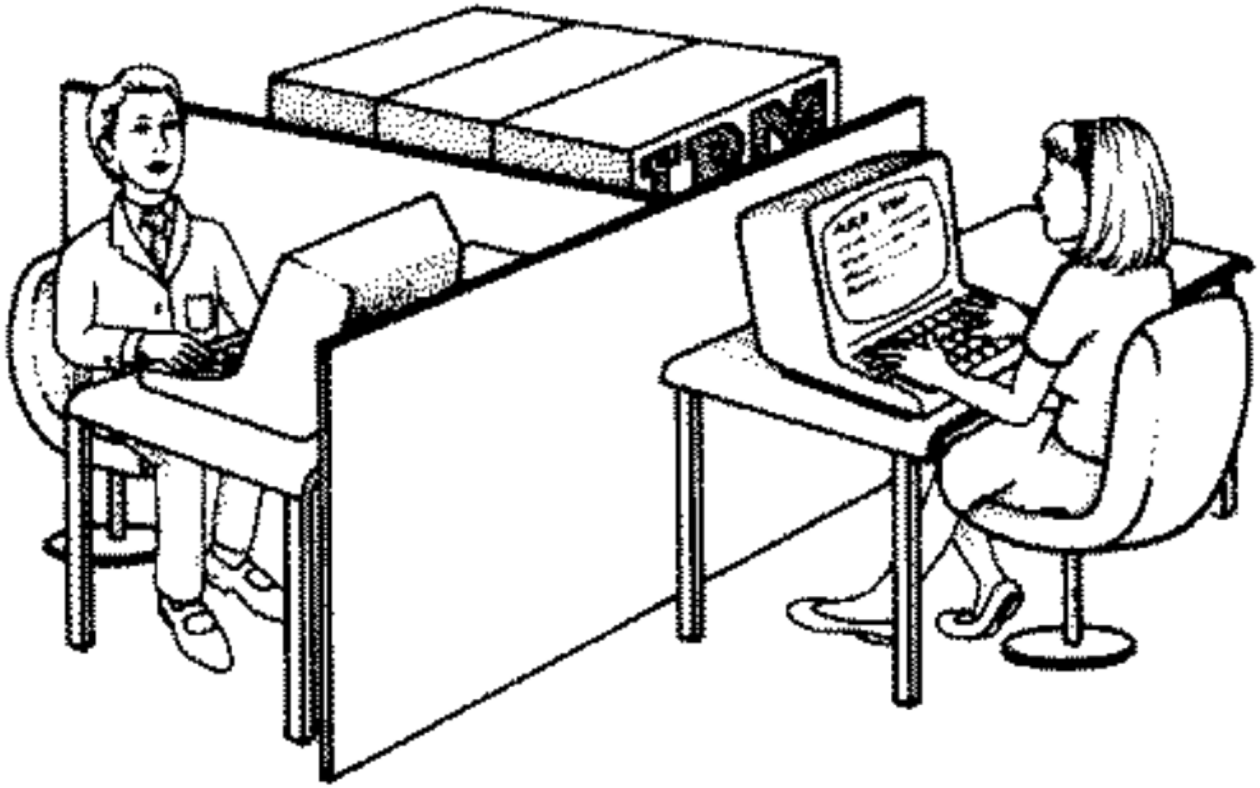


Figure: Turing test: after a five-minute conversation, the machine passes if it can mislead Alice in *at least* 30% of the cases.

What is Artificial Intelligence (AI)?

A framework: the four approaches

Russell & Norvig [Russell and Norvig 2021] organise definitions of AI along two axes — *thought vs. behaviour* and *human-like vs. rational* (i.e. *goal-optimal*):

	Humanly	**Rationally**
Thinking	cognitive modelling	laws of thought (logic)
Acting	*the Turing test*	*the rational agent*

Here *rational* means *"doing the right thing"*: pick the action expected to best achieve the goal — *not* "human-like", and not merely "logical".

- In Russell–Norvig's table, these definitions fit best under *acting humanly*: McCarthy and Turing emphasize human-like behaviour, while Britannica and Rich use a broader human-performance comparison, judging whether a system can solve tasks normally associated with human intelligence rather than whether it imitates a whole person.
- Some AI courses adopt *acting rationally*: design an *agent* (= something that acts) that perceives its environment and acts to achieve the *best expected outcome*.
- Many ML/deep-learning–focused intros instead define AI as *learning from data*.

What is Artificial Intelligence (AI)?

More on Turing Test

The idea that *"intelligence is the same as intelligent behavior"* has been challenged by some. The best known counter-argument is John Searle's *Chinese Room Argument* (1980)

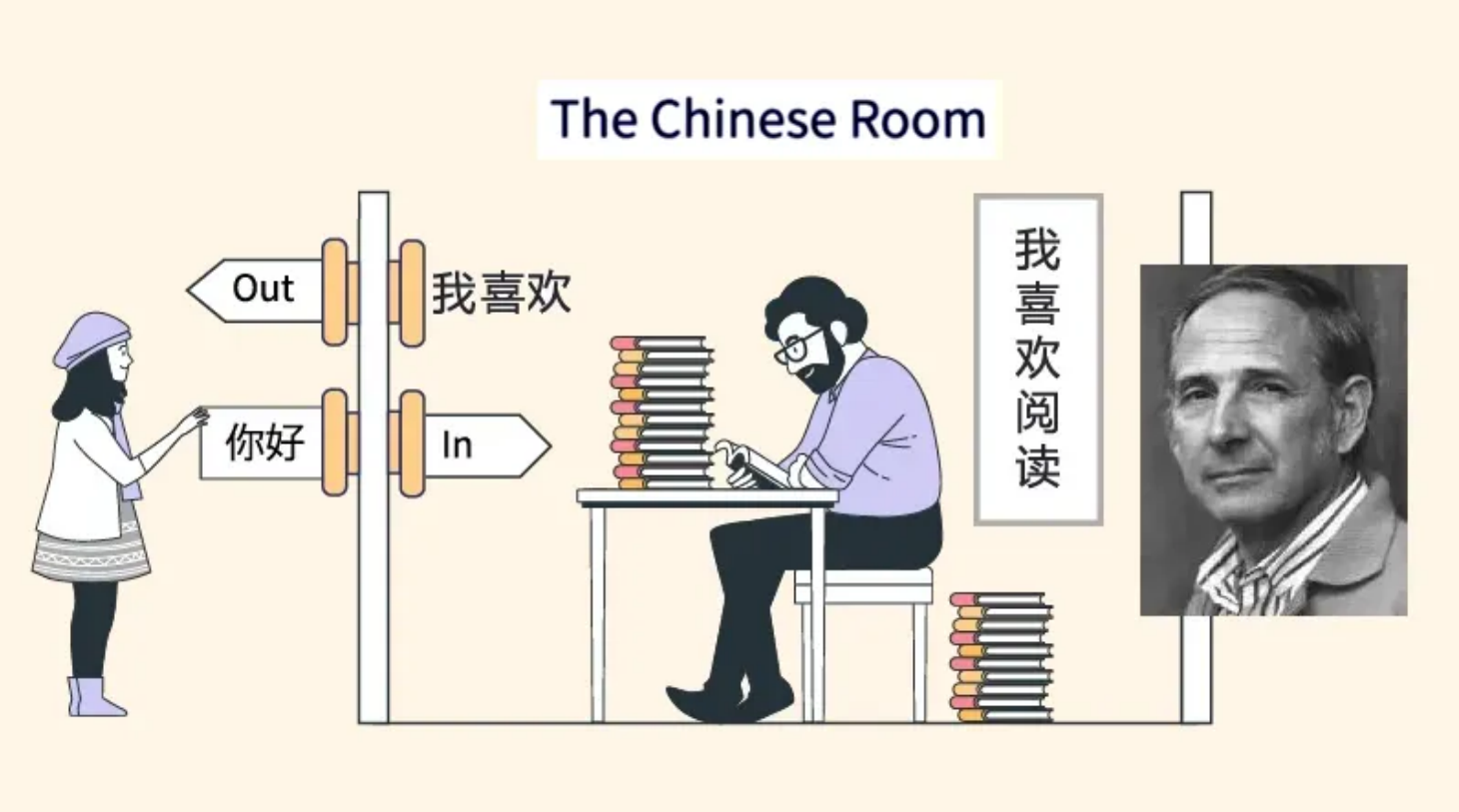


Figure: The Chinese Room Argument by John Searle (1980). Image taken from the Internet

What is Artificial Intelligence (AI)?

Two different distinctions are often confused here:

- *Does the machine really think? (Searle, 1980)* Under *strong AI* an appropriately programmed computer *literally has a mind* and genuinely understands; under *weak AI* it only *simulates* thinking and is a useful tool. Searle's Chinese Room argues against strong AI.
- *How broad are its abilities? Narrow AI* is built for one specific task (e.g., facial recognition, web search, driving); *general AI (AGI)* could perform any intellectual task a human can. *Most current AI is narrow AI.*

Additionally, the strong development of current AI technologies suggests the idea of *artificial superintelligence (ASI)*, a level of intelligence that surpasses human intelligence across all domains. (For example, see `https://www.ibm.com/think/topics/artificial-superintelligence`)

What is Artificial Intelligence (AI)?

What "AI" means today

Today's AI is dominated by systems *learned from data*, often described in three layers:

Foundation models / LLMs huge models pre-trained on massive data (GPT, Claude, Gemini, Llama).

Generative AI foundation models used to *produce* text, images, audio, or code from a *prompt* (ChatGPT, Midjourney).

Agentic AI models that *plan, call tools, and act autonomously* over many steps. They mainly show *autonomy*; some are also *adaptive* when they improve from feedback or experience.

Nesting: *deep learning* $\subset$ *machine learning* $\subset$ *AI* — *ML is a subfield, not a synonym for AI*.

Caveat

Fluency hides real limits: *hallucination*, brittle reasoning, high cost, bias, and *no guarantee of understanding*.

What is Artificial Intelligence (AI)?

So... is it AI? A practical test

With no agreed definition, some suggest testing two properties:

- *Autonomy* — performs tasks in complex environments without constant guidance.
- *Adaptivity* — improves performance by learning from experience.

Exercise 1 (Border cases)

Score each on autonomy and adaptivity — is it "AI"?
(1) a thermostat, (2) a spreadsheet computing a sum, (3) a GPS route-planner, (4) a spam filter, (5) ChatGPT

Two cautions

AI ≠ robotics: the intelligence is in the decision-making, not the hardware (robots may *use* AI). And *intelligence*/*understanding* are *"suitcase words"* (Minsky) — they bundle many meanings, fuelling hype.

What is Artificial Intelligence (AI)?

Exercise 2

There are a few other ways people define AI, for example,

- AI is any "cool things that computers can do"
- AI is any machine which simulates human behavior
- AI is an autonomous[1] and adaptive[2] system

Among all definitions we have mentioned, which one do you like *best*? How would *you* define AI?

Exercise 3

List some real-world applications of AI which we have not yet mentioned.

[1] The ability to perform tasks in complex environments without constant guidance by a user.

[2] The ability to improve performance by learning from experience.

Related Fields

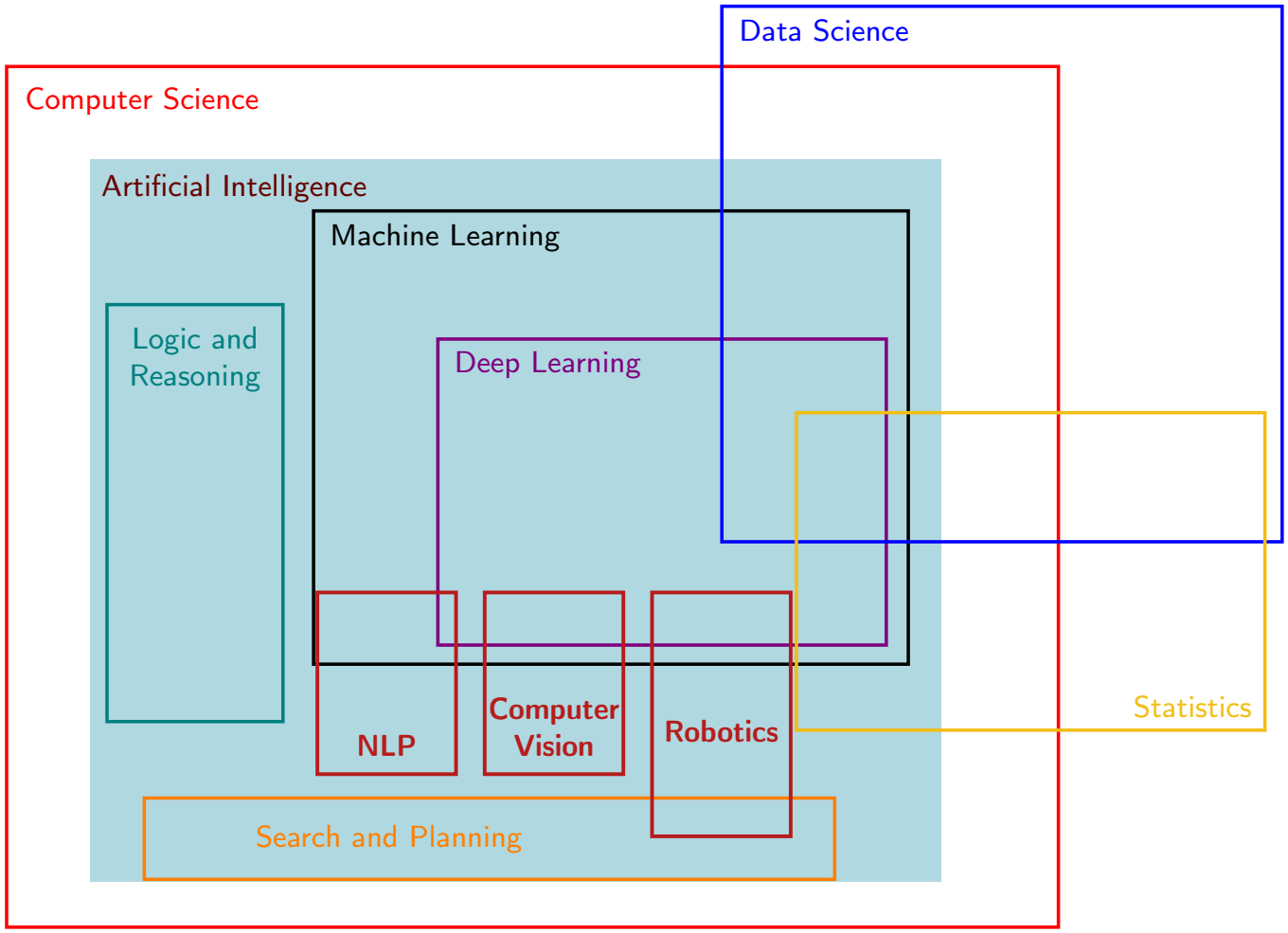


Figure: Related fields of AI. A box *inside* another means "is a subfield of" ($\mathrm{CS} \supset \mathrm{AI} \supset \mathrm{ML} \supset \mathrm{DL}$); *overlapping* boxes share methods, neither containing the other. NLP, computer vision and robotics *straddle* ML (their modern methods are deep-learning-based); robotics also overlaps search & planning. *Non-touching boxes are weakly, not un-, related.*

History of AI
Early beginnings and the symbolic era

Note: There are different ways to present the history of AI. Here is a reasonable summary.

Two lenses: AI swings between *symbolic* and *connectionist* approaches, and moves in *hype cycles*: bold predictions → disappointment → "*winters*" → revival.

- **Early beginnings (1930s–1950s)**
  - Theoretical foundations: formal logic, computation theory (Church–Turing), early automata.
  - McCulloch & Pitts (1943): first mathematical model of a neural network. Turing (1950): the Turing test.
- **Birth and the symbolic ("logic") era (1956–1974)**
  - Named at the *1956 Dartmouth Conference* (McCarthy coins "artificial intelligence"); Newell & Simon's *Logic Theorist* (the first automated theorem-proving program); McCarthy's programming language *LISP* (1958).
  - Bold predictions: Newell & Simon (1958) — a chess world-champion computer *"within ten years"* (actually *1997*, Deep Blue — ∼40 years late); Simon (1965) — machines doing "any work *that* a man can do" within twenty years.

History of AI
First AI winter and expert-systems boom

- **First AI winter (1974–1980)**
  - Over-promising met reality — the *combinatorial explosion* (the number of possibilities to search grows so fast that early methods could not scale to real problems). The *Lighthill Report* (1973), a UK government review judging that AI had over-promised, plus DARPA funding cuts, triggered the winter. Earlier, Minsky & Papert's book *Perceptrons* (1969) had stalled neural-net research by proving that a single-layer network can't even learn *XOR* (a simple either/or logic function).
- **Expert-systems boom (1980–1987)**
  - Industry adopts rule-based *expert systems* — programs that encode a human expert's *if–then rules* (MYCIN diagnosed infections, DENDRAL identified molecules, XCON/R1 configured computer orders — saving DEC ∼$40M/year). Japan launches the *Fifth Generation* project (1981), a national push to build AI-oriented computers.

History of AI

Second AI winter and statistical revival

- **Second AI winter (1987–1993)**
  - The market for *LISP machines* (specialised computers built to run AI programs) collapses; expert systems prove *brittle* and costly — the *knowledge-acquisition bottleneck*: every rule had to be hand-coded from human experts, which was slow and did not scale. To shed the "AI" stigma, researchers rebrand as *machine learning*, *pattern recognition*, *informatics*.
- **Statistical revival and consolidation (1990s–2000s)**
  - The field turns to *probability and data* (Bayesian networks, support-vector machines — methods that *learn patterns from data* instead of following hand-coded rules) and ships AI in real products: IBM's chess computer *Deep Blue* beats *Garry Kasparov* (1997), then the reigning *world chess champion* — the first time a machine beat a world champion at chess; IBM's question-answering system *Watson* wins the US TV quiz show *Jeopardy!* (2011), beating its best human champions.

The History of AI
Deep learning and generative AI

- **Deep-learning breakthrough (2012–2016)**
  - the deep neural network *AlexNet* (2012) wins the *ImageNet* contest — a yearly competition to sort over a million photos into 1000 categories; AlexNet roughly *halved the error rate*, proving deep learning works and *kicking off* today's deep-learning era. DeepMind's Go program *AlphaGo* beats *Lee Sedol* (2016), a top human champion at *Go* — a board game long thought far harder for computers than chess. Enabled by *big data* and *GPUs*.
- **Generative AI and foundation models (2017–present)**
  - The *Transformer* architecture (2017) — which reads a whole sequence at once using *"attention"* — underlies today's *large language models*: GPT-3 (2020) → *ChatGPT* (Nov 2022) → GPT-4 (2023) → reasoning/agentic models (2024–25). DeepMind's *AlphaFold 2* (2020/21) solves *protein-structure prediction* — computing a protein's 3D shape from its amino-acid sequence, a ∼50-year grand challenge in biology (2024 Nobel in Chemistry). AI becomes pervasive — raising social and ethical questions.

The History of AI

Why AI is “hot” today?

- **It really works now — and it’s everywhere**
  - Today’s AI skips the big philosophical questions and just *solves real, useful problems*.
  - It is already in tools you use daily (phone assistants, recommendations, ChatGPT), and runs for real in healthcare, finance, and education — not only in labs.
- **We finally have the tools to build it**
  - The ingredients are all here — *lots of data*, *fast chips* (GPUs/TPUs), *cloud computers*, and *free open-source software* — so building and shipping AI is now fast and cheap; almost anyone can try.
- **But stay a little skeptical**
  - AI has been over-hyped before: bold promises (“human-level AI in ten years”) kept falling short. Enjoy the progress, but weigh today’s excitement against that track record.

The History of AI

Exercise 4

Do your own research on the history of AI. Some interesting topics would probably be:

(a) The *Dartmouth Conference* and the birth of AI

(b) The rise and fall of *expert systems*

(c) The history and impact of *neural networks and deep learning* on AI (recall: *machine learning* is a subfield of AI, and *deep learning* is a subfield of machine learning)

(d) The history and development of *two major theories of AI*: *symbolic AI* (AI should be based on traditional computer logic or preprogramming $\Rightarrow$ Modeling human's *mind*) and *connectionist AI* (AI needed to use systems similar to the brain like neural networks $\Rightarrow$ Modeling human's *brain*). (Some scientists want to go further by blending the two approaches into something called *neuro-symbolic AI*.)

(e) The *"AI effect"*: AI is a moving target — once a problem is solved (chess, OCR, translation), people stop calling it "AI". How might this help explain the field's repeated "winters" and rebranding?

Implications of AI on the Society

How to think about AI's impact

AI is just a tool: its effects depend on how we *build, use, and govern* it. To judge any application, ask which *values* it affects:

- *Fairness* (does it discriminate?), *Transparency* (can we explain or contest it?), *Accountability* (who is responsible?), *Privacy*, *Safety & security*, and effects on *work and society*.

Each implication below stresses one or more of these. We look at the *risks*, then the *benefits*, then what is being *done*.

Implications of AI on the Society
Algorithmic Bias

Holland & Knight

Offices Careers

Practices Professionals Industries Insights Events News

MAY 27, 2025

Federal Court Allows Collective Action Lawsuit Over Alleged AI Hiring Bias

Workday Inc. Now Faces Nationwide Collective Action Over Alleged Age Discrimination in Artificial Intelligence (AI) Hiring Tools

Holland & Knight Alert

Jennifer Lada | Bryan Neal | D'Andre Chapman | Laura Askinazi

Derek Mobley, a man over 40 years old, sued Workday Inc. claiming that Workday's artificial intelligence (AI)-driven applicant screening tools have systematically disadvantaged him and other older job seekers. Mobley submitted more than 100 applications to companies using Workday's platform, and he was rejected every time. Mobley alleged that the AI tools – designed to score, sort, rank or screen applicants – unfairly penalize older candidates. The court initially dismissed Mobley's complaint but granted permission to file an amended version. Subsequently, Workday attempted to have this amended complaint dismissed as well but was unsuccessful.

The court, by denying Workday's motion to dismiss, recognized Mobley's claim as plausible under the Age Discrimination in Employment Act (ADEA), based on a disparate impact theory. This development marks a pivotal moment in the ongoing legal proceedings, as it allows Mobley to continue pursuing the case against Workday.

Key Developments

On May 16, 2025, Judge Rita Lin of the U.S. District Court for the Northern District of California granted preliminary certification under ADEA, allowing the lawsuit to move forward as a nationwide collective action – similar to a class action but

RELATED BLOG

Cybersecurity and Privacy Blog

RELATED PRACTICES

Labor and Employment

Labor and Employment Class Actions

Civil Rights, Discrimination and Retaliation

Artificial Intelligence

Artificial Intelligence Policy & Regulation

Technology

Emerging Technologies Law and Policy

Figure: *Mobley v. Workday* (2025): a man rejected by AI hiring software (from the company *Workday*) is suing, saying it *unfairly rejected him because of his age*. A U.S. court let others rejected the same way *join the case* — which asks: when an AI turns you down for a job, *who is to blame*?
(https://www.hklaw.com/en/insights/publications/2025/05/federal-court-allows-collective-action-lawsuit-over-alleged)

Implications of AI on the Society

Algorithmic Bias

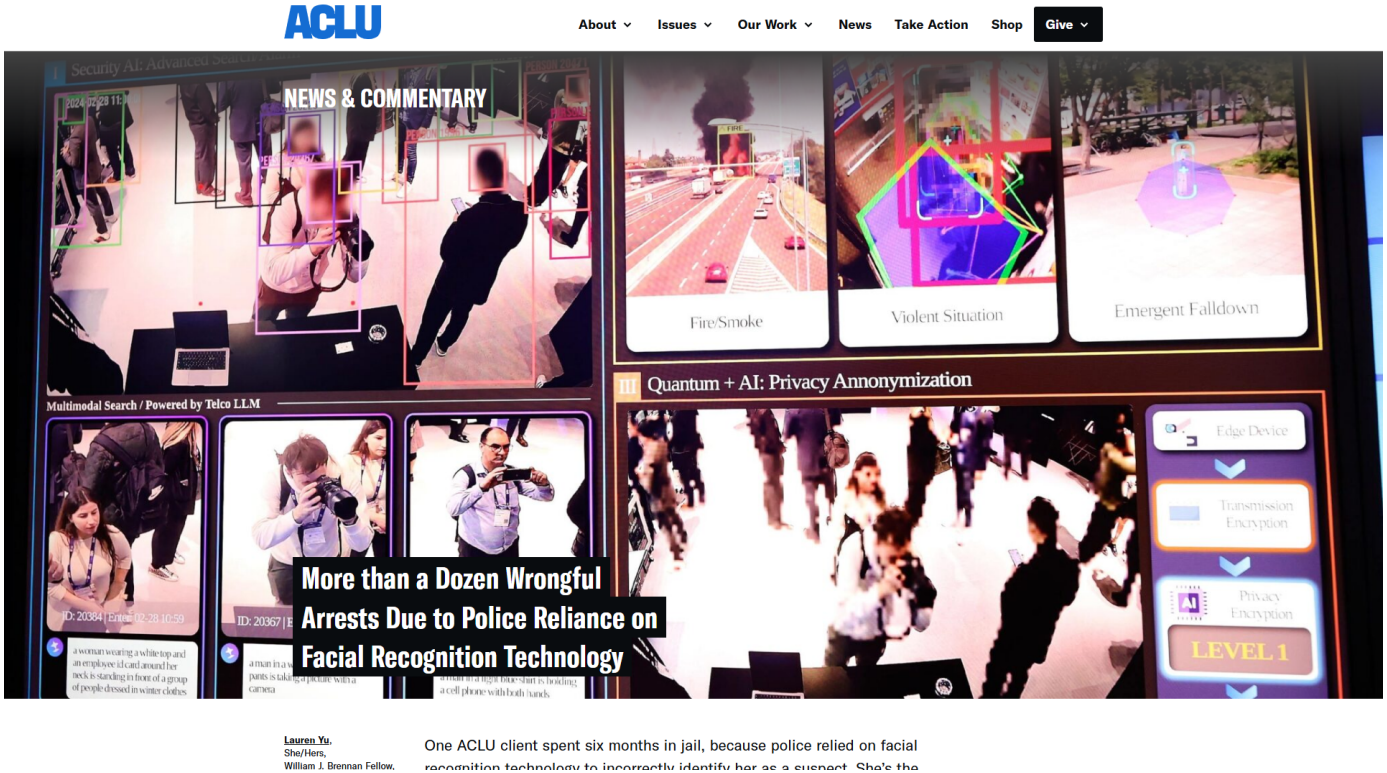


Figure: ACLU (2026): a *14th* documented wrongful arrest from police reliance on facial recognition — the technology flags innocent people who merely *look similar*, and the harm falls overwhelmingly on Black Americans. (https://www.aclu.org/news/privacy-technology/more-than-a-dozen-wrongful-arrests-due-to-police-reliance-on-facial-recognition-technology)

Implications of AI on the Society
Algorithmic Bias

Figure: Stanford “Algorithmic Monocultures in Hiring” (2026): when *many employers all use the same* AI screening tool, that one tool’s small bias gets repeated everywhere — so the *same people keep being rejected* (Black and Asian applicants more often). (https://digitaleconomy.stanford.edu/publication/algorithmic-monocultures-in-hiring/)

Algorithmic Bias

- Algorithmic bias isn't a hypothetical threat conceived by academic researchers. It's a real phenomenon that is already affecting people today.
- Bias has *several sources*: most often *biased data*, but also *proxy variables* (an innocent-looking feature that secretly stands in for a forbidden one — e.g. a zip code can reveal race even when race is never used) and *feedback loops* (the AI's past decisions become its future training data, so a small bias keeps reinforcing itself). Even "fairness" has *several competing mathematical definitions* that cannot all hold at once, so choosing one is partly a moral choice.
- Unlike *rule-based systems* (software that follows explicit if-then rules written by people), machine learning often gives *no clear reason* for a decision
  - Algorithms and datasets are often proprietary trade secrets and not open to public scrutiny.
  - Even with access, it is often difficult to identify which algorithmic components or data elements lead to discriminatory decisions.

Implications of AI on the Society

Seeing is no longer believing

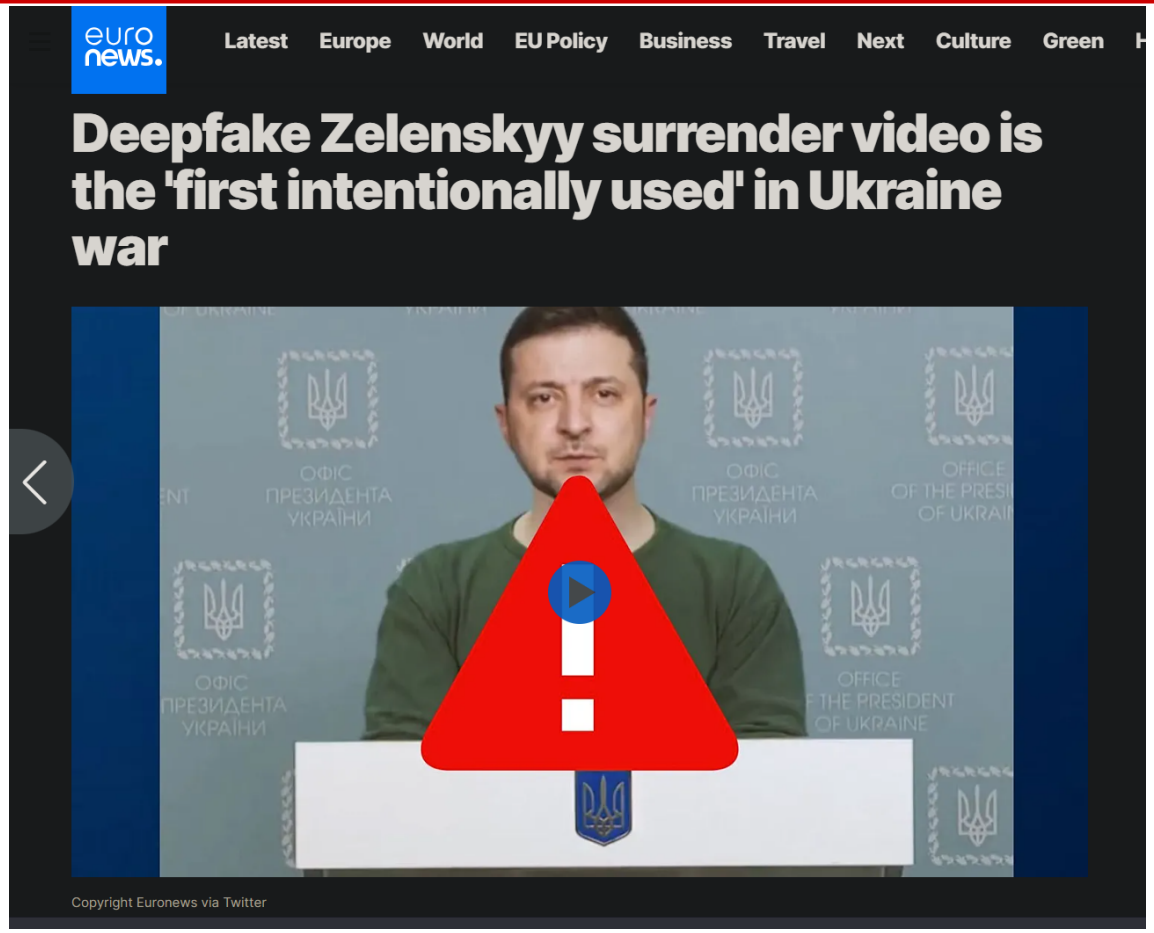


Figure: Euronews (2022): report on a deepfake video of Ukrainian President Volodymyr Zelenskyy calling citizens to surrender to Russia. (https://www.euronews.com/my-europe/2022/03/16/deepfake-zelenskyy-surrender-video-is-the-first-intentionally-used-in-ukraine-war)

Implications of AI on the Society

Seeing is no longer believing

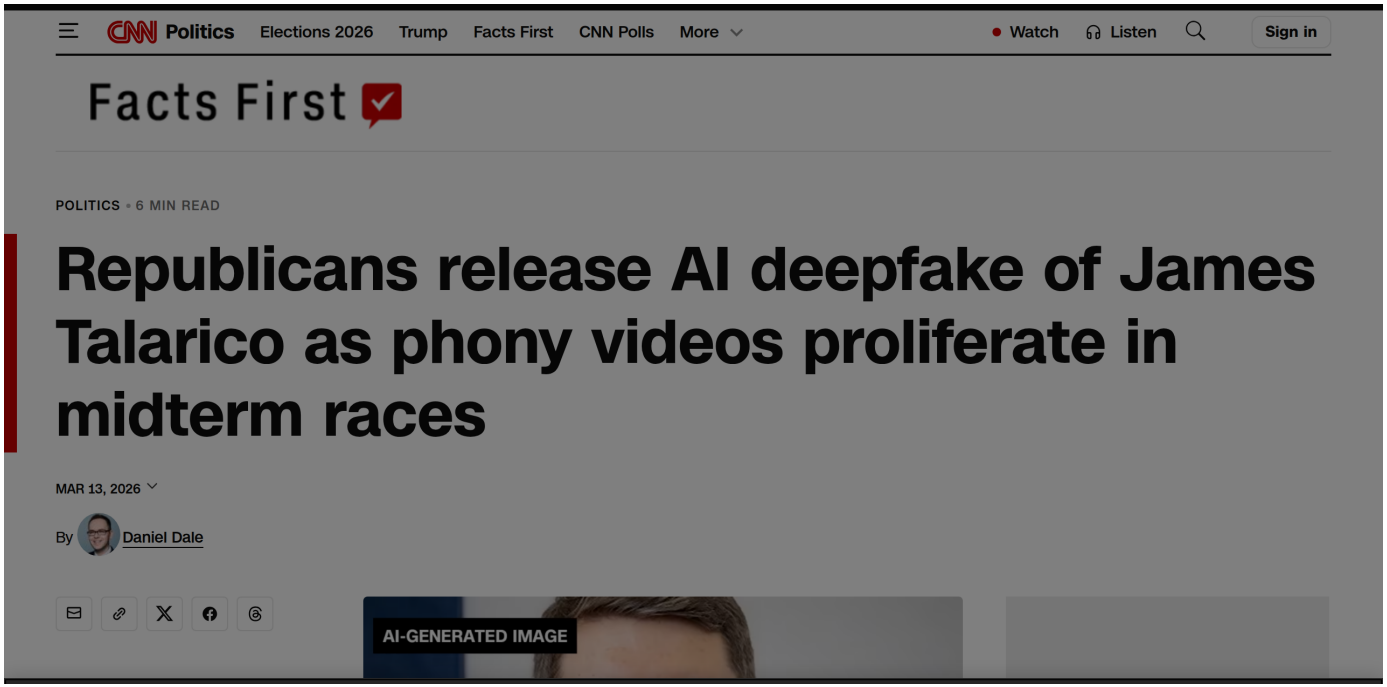


Figure: CNN (2026): a U.S. political party ran an *85-second deepfake* of a Senate candidate appearing to say things he never said — so realistic that, experts warned, most voters couldn't tell (the "AI generated" label was nearly invisible). (https://www.cnn.com/2026/03/13/politics/james-talarico-ai-deepfake-republicans-midterms)

Implications of AI on the Society

Seeing is no longer believing

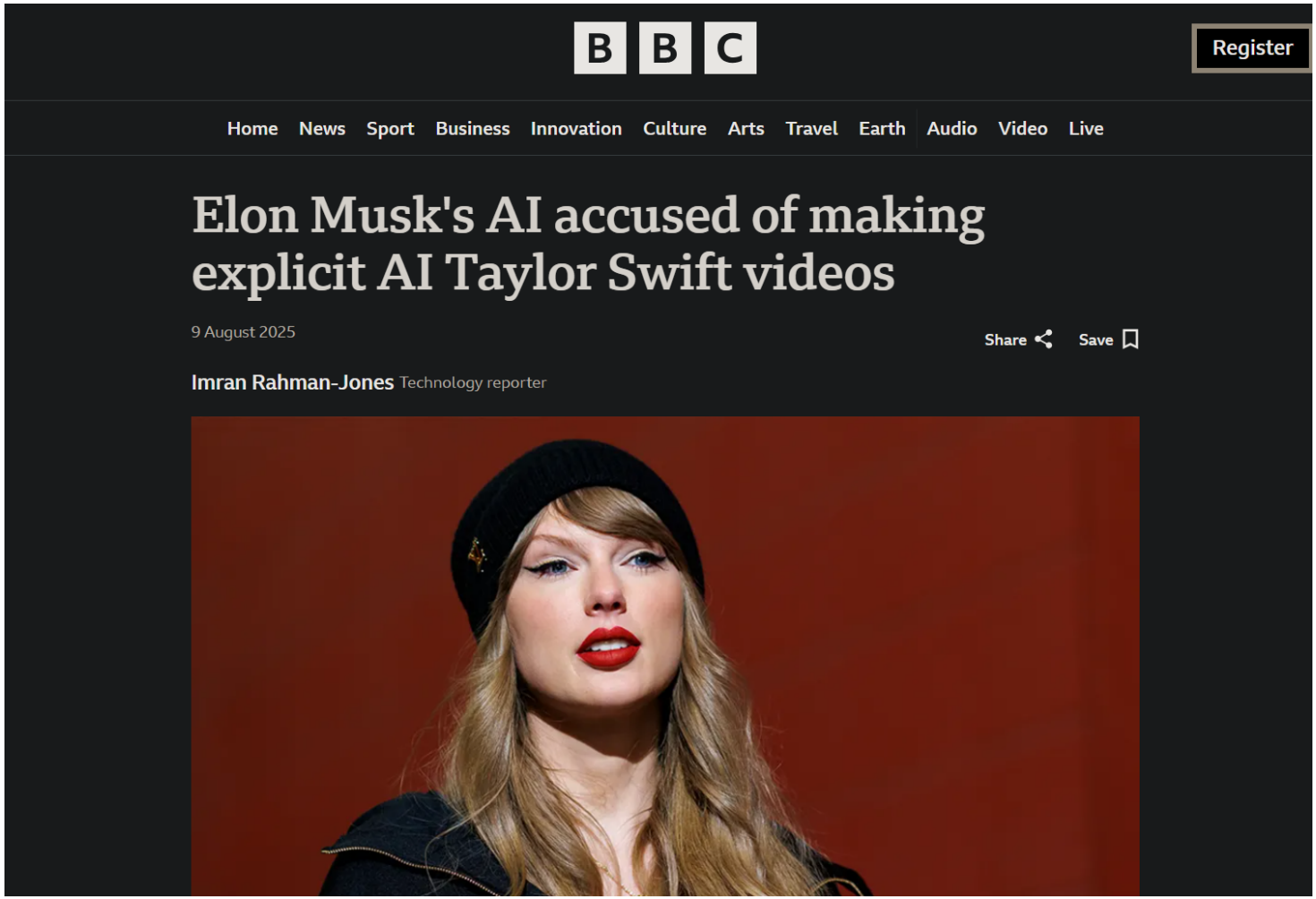


Figure: BBC News (2025): report on Elon Musk's AI accused of making explicit AI Taylor Swift videos using deepfakes. (https://www.bbc.com/news/articles/cwye62e1ndjo)

Implications of AI on the Society

Seeing is no longer believing

AI is *taking the possibilities of fabricating evidence to a whole new level*

- *Deepfake* and similar technologies can create *highly realistic but entirely fabricated images, videos, and audio recordings* of people saying or doing things they never actually said or did.
  - *Live face-swapping* (putting one person's face onto another) and *de-aging*, plus *voice cloning* (copying a person's voice) from only a few minutes of audio, are now in everyday apps.
- These technologies challenge *trust and verification* in media. Responses are emerging — ways to *tag AI-made content and record where it came from* ("watermarking" and "provenance" standards such as C2PA and Google's SynthID) and new laws (e.g. the US *TAKE IT DOWN Act*, 2025) — though such hidden markers *can be removed*, and the predicted flood of election deepfakes has so far been *smaller than feared*.

Implications of AI on the Society

When AI pollutes the scientific record

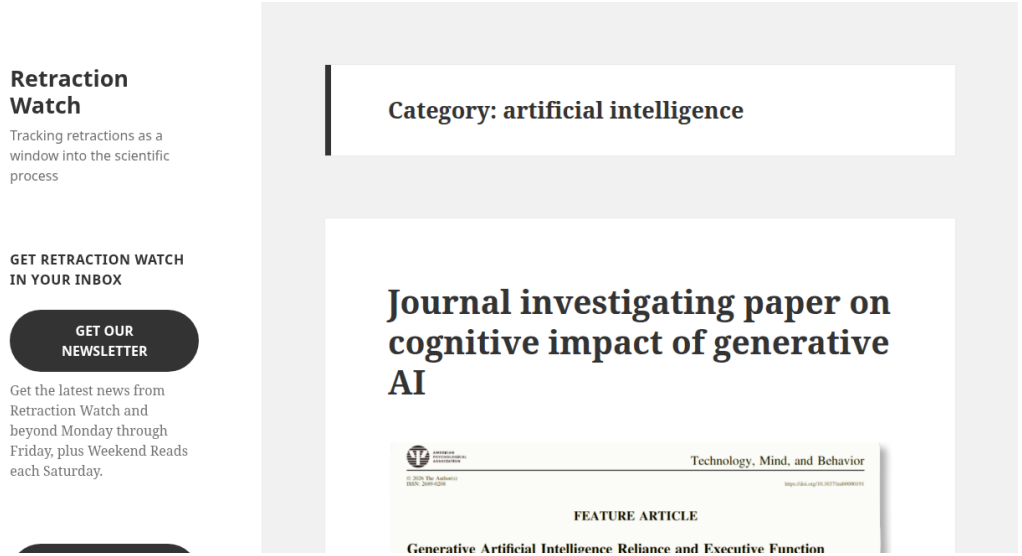
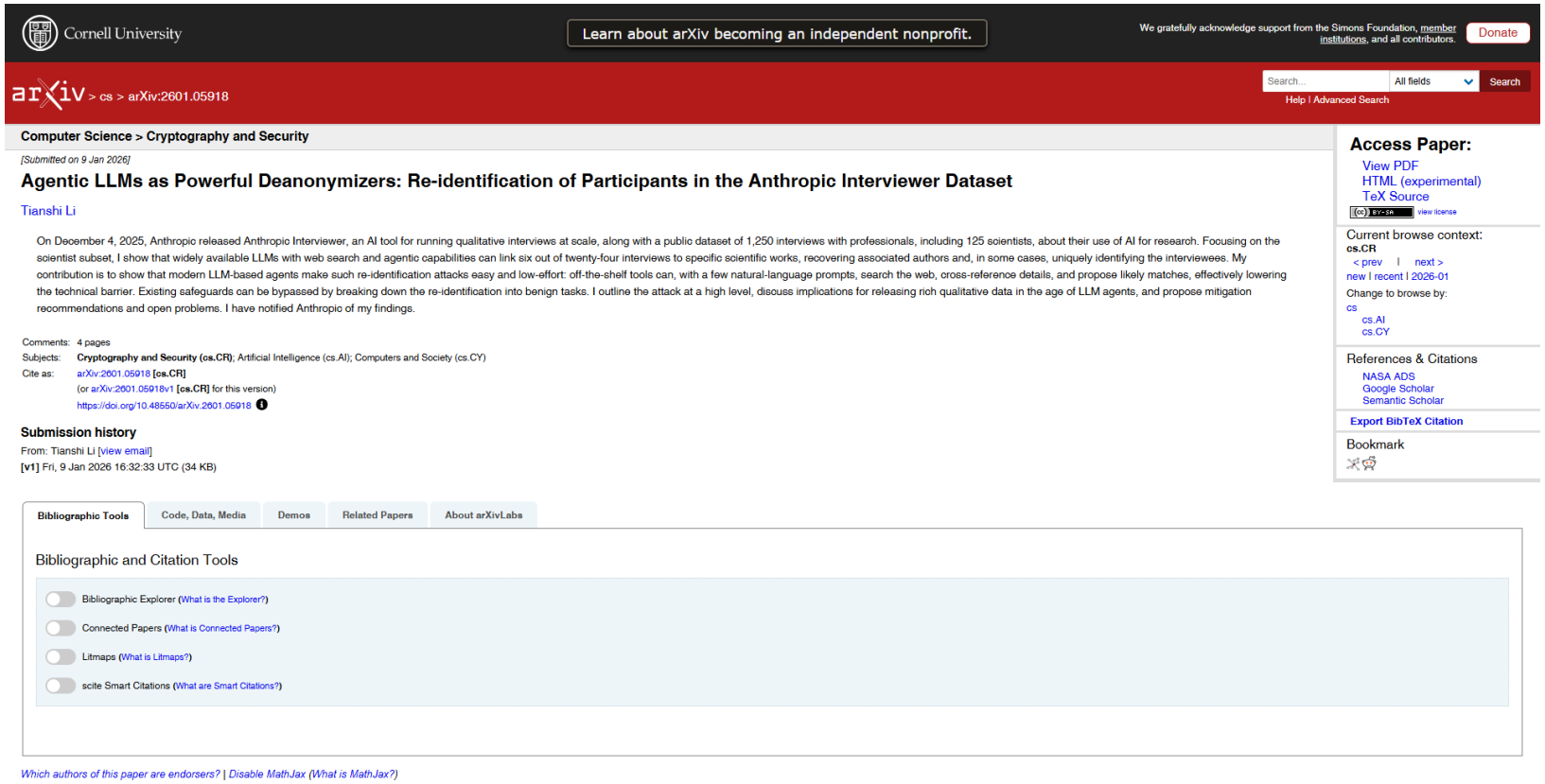


Figure: AI can also *pollute the scientific record*: a 2026 *Lancet* audit of ∼2.5M biomedical papers found *4,046 fabricated references* across *2,810 papers*. The rate rose more than *12-fold* since 2023, reaching about *1 in 277* papers in early 2026. The audit did not prove one cause, but listed paper mills, misconduct, and *uncritical AI-tool use* as possible causes. Related reports describe AI-generated letters in *Neurosurgical Review* and hallucinated citations in accepted NeurIPS 2025 papers. (https://pubmed.ncbi.nlm.nih.gov/42107362/; https://retractionwatch.com/2025/02/10/as-springer-nature-journal-clears-ai-papers-one-universitys-retractions-rise-drastically/; https://gptzero.me/news/neurips/)

Implications of AI on the Society

Privacy concerns



Cornell University

Learn about arXiv becoming an independent nonprofit.

We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. Donate

arXiv > cs > arXiv:2601.05918

Search... All fields Search

Help | Advanced Search

Computer Science > Cryptography and Security

[Submitted on 9 Jan 2026]

Agentic LLMs as Powerful Deanonymizers: Re-identification of Participants in the Anthropic Interviewer Dataset

Tianshi Li

On December 4, 2025, Anthropic released Anthropic Interviewer, an AI tool for running qualitative interviews at scale, along with a public dataset of 1,250 interviews with professionals, including 125 scientists, about their use of AI for research. Focusing on the scientist subset, I show that widely available LLMs with web search and agentic capabilities can link six out of twenty-four interviews to specific scientific works, recovering associated authors and, in some cases, uniquely identifying the interviewees. My contribution is to show that modern LLM-based agents make such re-identification attacks easy and low-effort: off-the-shelf tools can, with a few natural-language prompts, search the web, cross-reference details, and propose likely matches, effectively lowering the technical barrier. Existing safeguards can be bypassed by breaking down the re-identification into benign tasks. I outline the attack at a high level, discuss implications for releasing rich qualitative data in the age of LLM agents, and propose mitigation recommendations and open problems. I have notified Anthropic of my findings.

Comments: 4 pages
Subjects: Cryptography and Security (cs.CR); Artificial Intelligence (cs.AI); Computers and Society (cs.CY)
Cite as: arXiv:2601.05918 [cs.CR]
(or arXiv:2601.05918v1 [cs.CR] for this version)
https://doi.org/10.48550/arXiv.2601.05918

Submission history
From: Tianshi Li [view email]
[v1] Fri, 9 Jan 2026 16:32:33 UTC (34 KB)

Access Paper:
View PDF
HTML (experimental)
TeX Source
view license

Current browse context:
cs.CR
< prev | next >
new | recent | 2026-01
Change to browse by:
cs
cs.AI
cs.CY

References & Citations
NASA ADS
Google Scholar
Semantic Scholar

Export BibTeX Citation

Bookmark

Bibliographic Tools | Code, Data, Media | Demos | Related Papers | About arXivLabs

Bibliographic and Citation Tools

Bibliographic Explorer (What is the Explorer?)
Connected Papers (What is Connected Papers?)
Litmaps (What is Litmaps?)
scite Smart Citations (What are Smart Citations?)

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

Figure: A 2026 study: an ordinary AI agent *figured out the real identities* of *6 of 24* people in a supposedly “anonymous” research dataset — doing in a few prompts what used to take an expert. (https://arxiv.org/abs/2601.05918)

Implications of AI on the Society

Privacy concerns

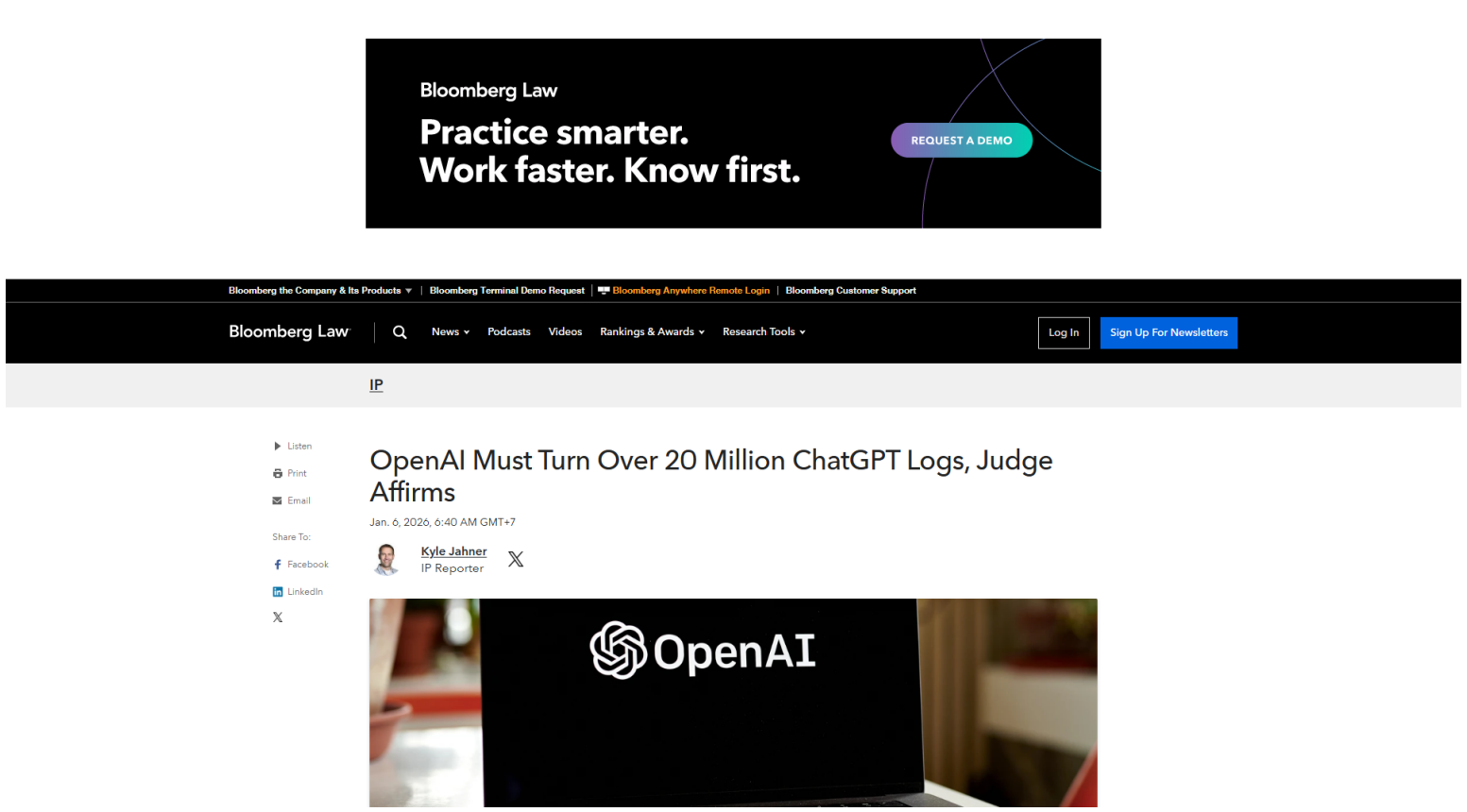


Figure: Bloomberg Law (2026): in the *New York Times*-led copyright case against OpenAI, a judge ordered OpenAI to hand over *20 million (anonymized) ChatGPT conversation logs* — a reminder that your chats can become legal evidence.
(https://news.bloomberglaw.com/ip-law/openai-must-turn-over-20-million-chatgpt-logs-judge-affirms)

Implications of AI on the Society

Privacy concerns

WIZ Platform Solutions Pricing Resources Customers Company Get a demo

← Blog

Wiz Research Uncovers Exposed DeepSeek Database Leaking Sensitive Information, Including Chat History

A publicly accessible database belonging to DeepSeek allowed full control over database operations, including the ability to access internal data. The exposure includes over a million lines of log streams with highly sensitive information.

Gal Nagli

January 29, 2025 | 3 minute read

Table of contents

Executive Summary
Exposure Walkthrough
Key Takeaways
Conclusion

Wiz Research has identified a publicly accessible ClickHouse database belonging to DeepSeek, which allows full control over database operations, including the ability to access internal data. The exposure includes over a million lines of log streams containing chat history, secret keys, backend details, and other highly sensitive information. The Wiz Research team immediately

Figure: Wiz Research (2025): the AI company *DeepSeek* left a database publicly exposed, leaking *chat histories and secret keys* — AI services are themselves new, high-value privacy targets. (https://www.wiz.io/blog/wiz-research-uncovers-exposed-deepseek-database-leak)

Implications of AI on the Society
Privacy concerns

- *Before "the age of AI"*, privacy concerns were primarily focused on data collection and usage by companies.
  - In principle, every click, scroll and the time spent on content can be recorded.
  - Websites and trackers can access or infer your browsing history, enabling cross-site profiling and persistent personalization.
  - Consequently, a simple search for Android phones can trigger targeted ads (e.g., cheap and discounted options) across other sites unless privacy protections (incognito, trackers blocked) are used.
- *In "the age of AI"*, privacy concerns have evolved. Some examples include:
  - The re-identification of supposedly anonymous data — now even automated by AI agents.
  - Leaks and breaches of the AI services that hold your data, such as the DeepSeek database exposure.
  - The potential loss of anonymity through advanced profiling techniques.
  - The potential of using AI to match user accounts in almost any service that collects detailed data about user behaviors (e.g., typing patterns).

Implications of AI on the Society

Security concerns

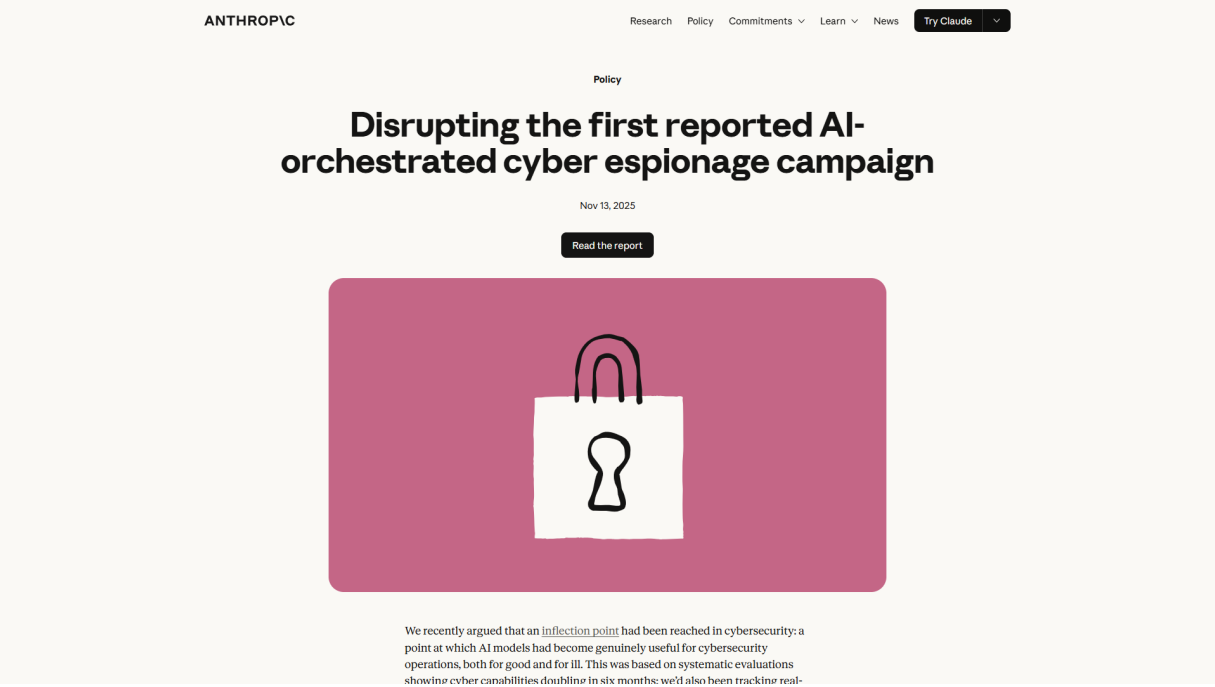


Figure: Anthropic (Nov 2025): attackers manipulated Claude Code by hiding a cyber-espionage campaign inside small, legitimate-looking security-testing tasks. Anthropic describes the case as the first reported large-scale *AI-orchestrated* cyber-espionage campaign: humans chose targets and approved key steps, while Claude Code performed an estimated *80–90% of tactical operations*. (https://www.anthropic.com/news/disrupting-AI-espionage)

Implications of AI on the Society

Security concerns

Figure: The Hacker News (2025): web browsers with a built-in AI assistant (such as Perplexity's Comet) can be *tricked into following harmful hidden instructions* and running them — a new kind of security hole. (https://thehackernews.com/2025/08/experts-find-ai-browsers-can-be-tricked.html)

Implications of AI on the Society

Security concerns

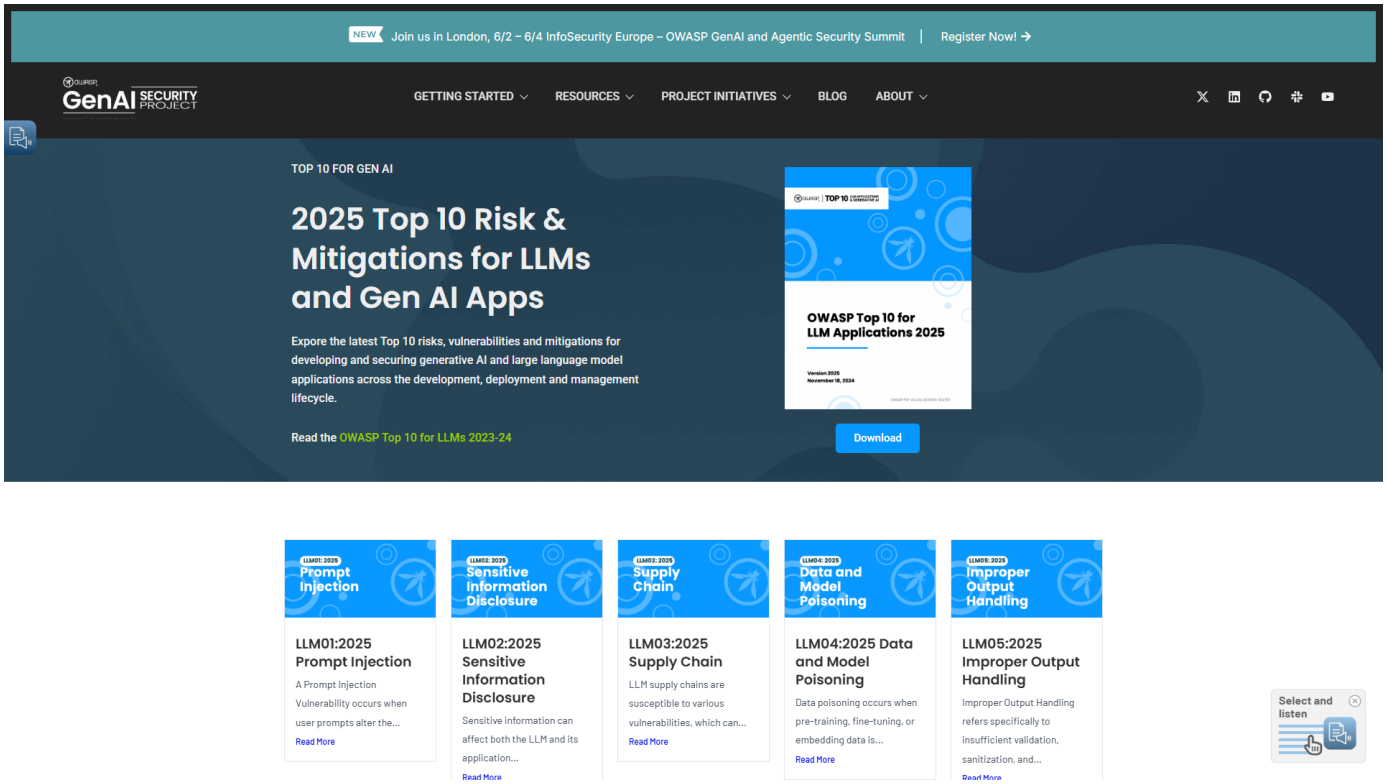


Figure: OWASP (a nonprofit that publishes software-security checklists), *Top 10 for LLM Applications* (2025): the standard list of the biggest security risks for AI apps — *prompt injection* (feeding the AI hidden instructions to hijack it) ranks #1. (https://genai.owasp.org/llm-top-10/)

Implications of AI on the Society
Security concerns

- The rise of AI technologies has led to new security vulnerabilities, as seen in various incidents involving data leaks and unauthorized access.
- Organizations are increasingly implementing strict policies and guidelines to mitigate these risks, including banning certain AI tools and conducting regular security audits.

Exercise 5

Do your own research about AI risks and security concerns. A good starting point is the *OWASP Top 10 for LLM Applications* (https://genai.owasp.org/llm-top-10/).

Implications of AI on the Society

The changing job market

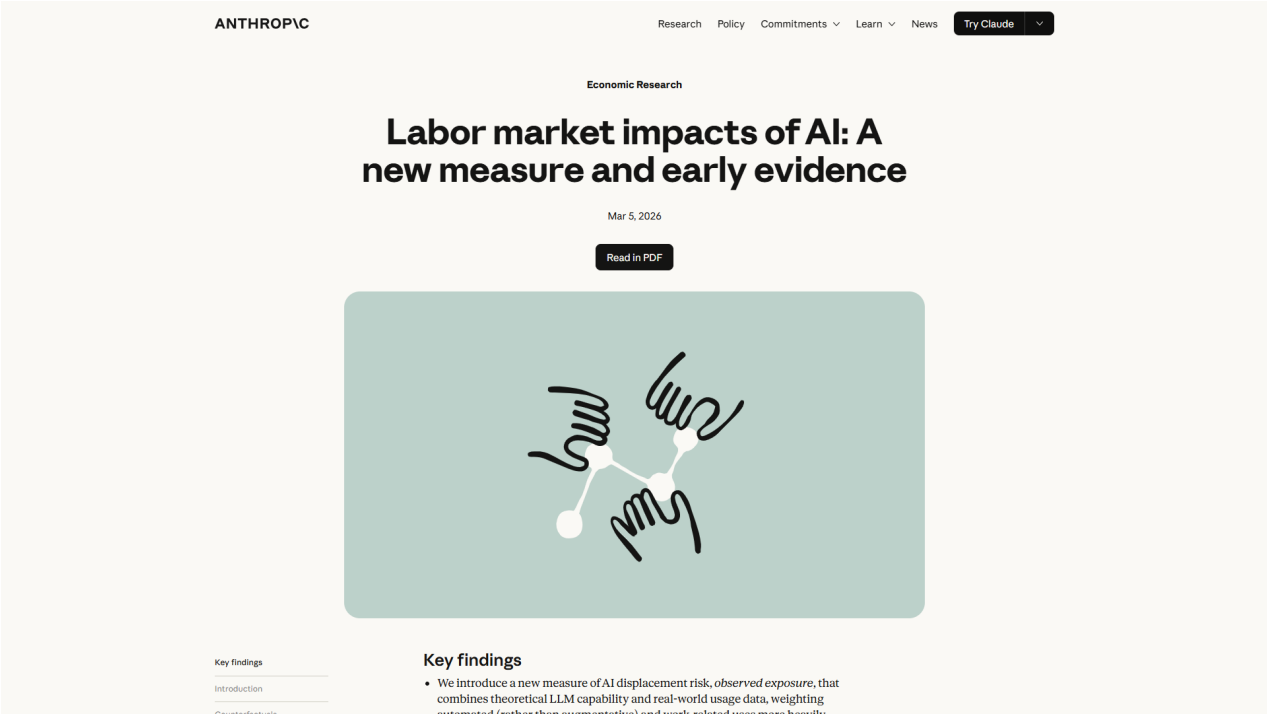


Figure: Anthropic Economic Index (2026): instead of *predicting* AI's effect on jobs, this study *measures what is actually happening* — it combines what AI *can* do with *real usage data* (which work tasks people actually use AI for), and separates AI *doing a task for you* (automation) from AI *helping you do it* (augmentation). The key point: a job being *"exposed"* to AI (AI can touch part of it) is *not* the same as being *replaced* — much of today's use is help, not replacement. (https://www.anthropic.com/research/labor-market-impacts)

Implications of AI on the Society

The changing job market

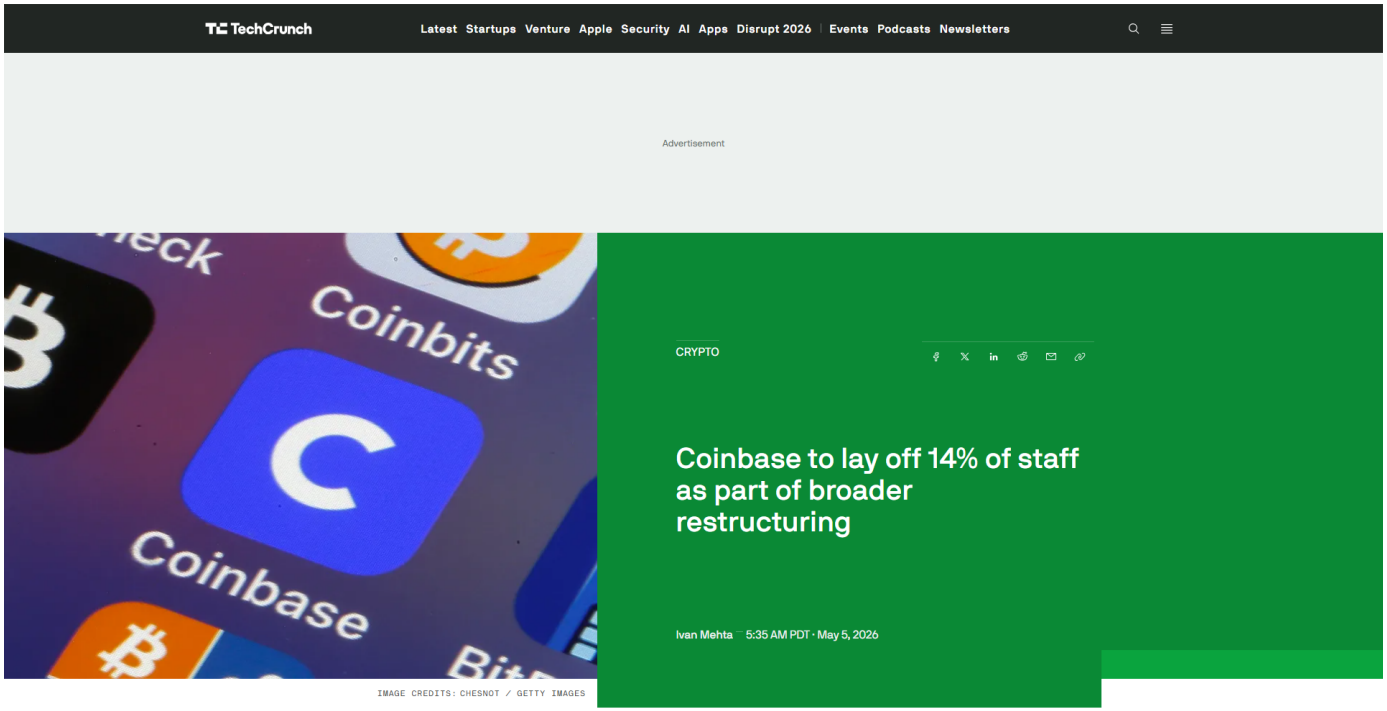


Figure: TechCrunch (2026): Coinbase laid off *∼14% of staff* (∼700 people) in an AI-driven restructuring, flattening teams and experimenting with AI-assisted "one-person teams" — AI reshaping who companies hire and keep.
(https://techcrunch.com/2026/05/05/coinbase-to-lay-off-14-of-staff-as-part-of-broader-restructuring/)

Implications of AI on the Society

The changing job market

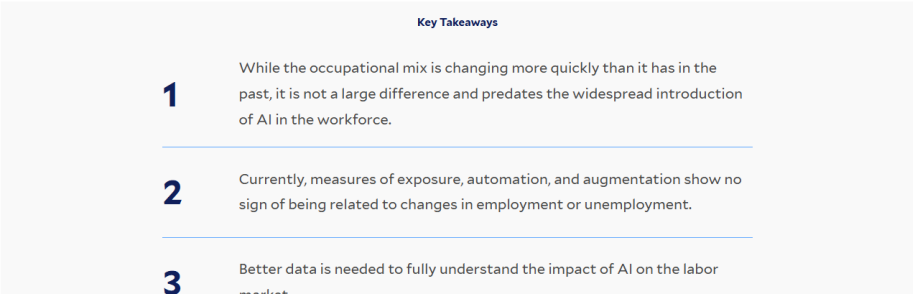


Figure: The Budget Lab at Yale (2025): so far, *little broad disruption* is visible in aggregate labour data — a reminder to separate AI *hype* from *measured* impact. (https://budgetlab.yale.edu/research/evaluating-impact-ai-labor-market-current-state-affairs)

Implications of AI on the Society

The changing job market

- **What usually happens:** new technology rarely just deletes jobs — it *moves people into different kinds of work*. The repetitive parts get automated, and new kinds of jobs appear (e.g., factory work $\longrightarrow$ keeping the machines running; paper filing $\longrightarrow$ managing digital information).
- **What the new jobs ask for:** more of the things machines are still bad at — *solving new problems, creativity, working with people*, and being comfortable with technology.
  - Over time, more of us are likely to work on *research, creative tasks, and jobs built around human contact*. For a longer take, see Abhinav Suri's essay "Artificial Intelligence and the Rise of Economic Inequality" (https://abhinavsuri.com/blog/2017/aiinequality/).
- **What is different this time:** AI is automating parts of *both hands-on ("blue-collar")* and *desk ("white-collar")* jobs, and it is moving *faster* — so one person may have to retrain more than once in a single career.

Implications of AI on the Society

The changing job market

- **Jobs most at risk:** driving (trucks, taxis), simple data entry, basic customer service, and repetitive factory work — though *when* this happens depends on cost, technology, and the law.
- **What can go wrong if we ignore it:** whole towns or groups of workers losing jobs at once, *lower wages*, a wider gap between regions and people, and worse working conditions for those affected.
- **What helps:** there is no single fix — we need *good education, chances to keep learning through life, sensible labour rules*, and governments and companies acting together, so people are protected and can still share in what AI makes possible.

Implications of AI on the Society
Environmental cost

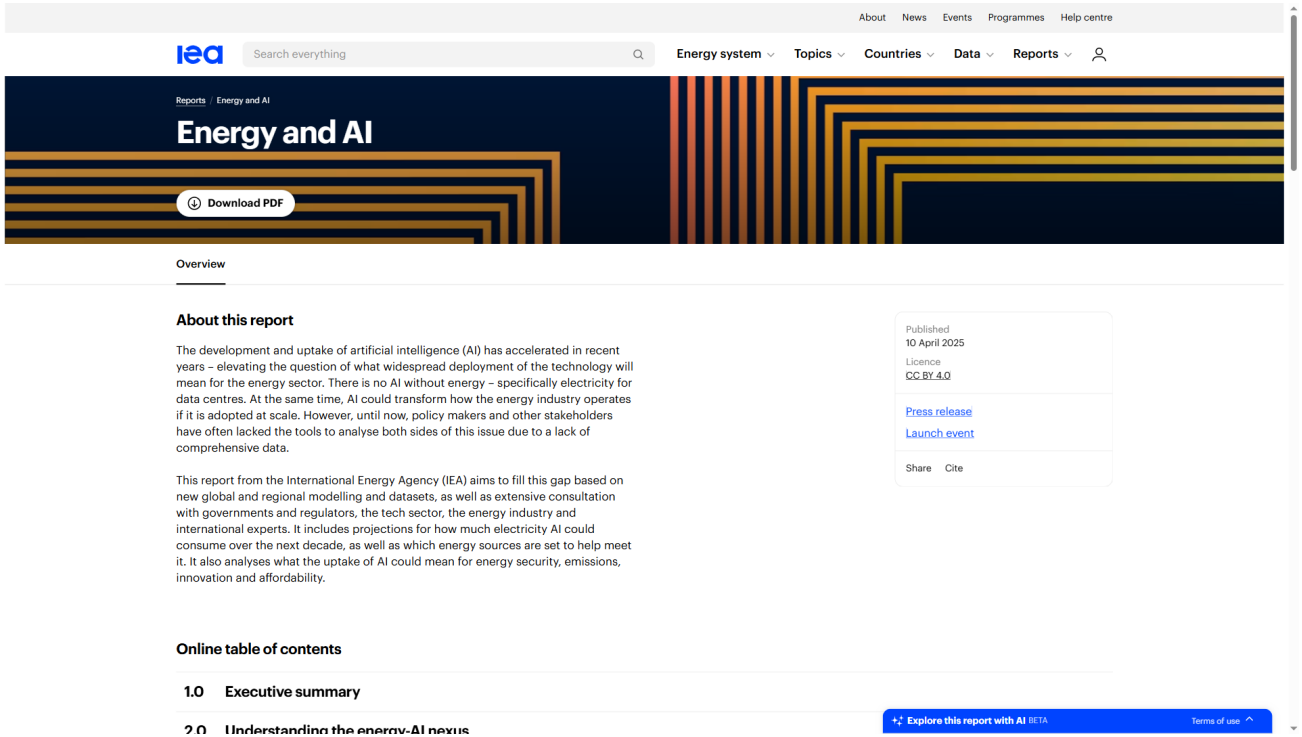


Figure: IEA “Energy and AI” (2025): electricity demand from *data centres* is projected to roughly *double* by 2030, driven largely by AI. (https://www.iea.org/reports/energy-and-ai)

Implications of AI on the Society
Autonomous weapons

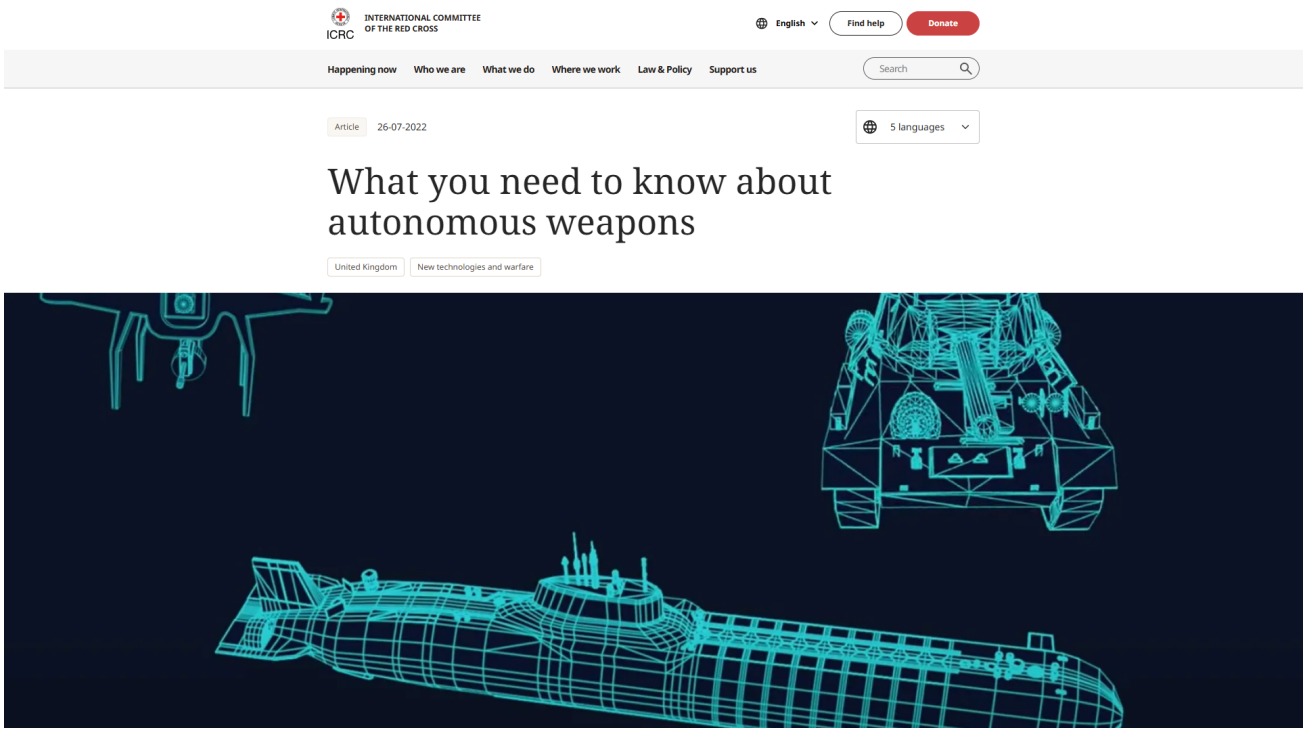


Figure: ICRC: *autonomous weapons* can select and engage targets without human control — raising the "*human control*" debate, now discussed at the UN. (https://www.icrc.org/en/document/what-you-need-know-about-autonomous-weapons)

Implications of AI on the Society

Copyright & training data

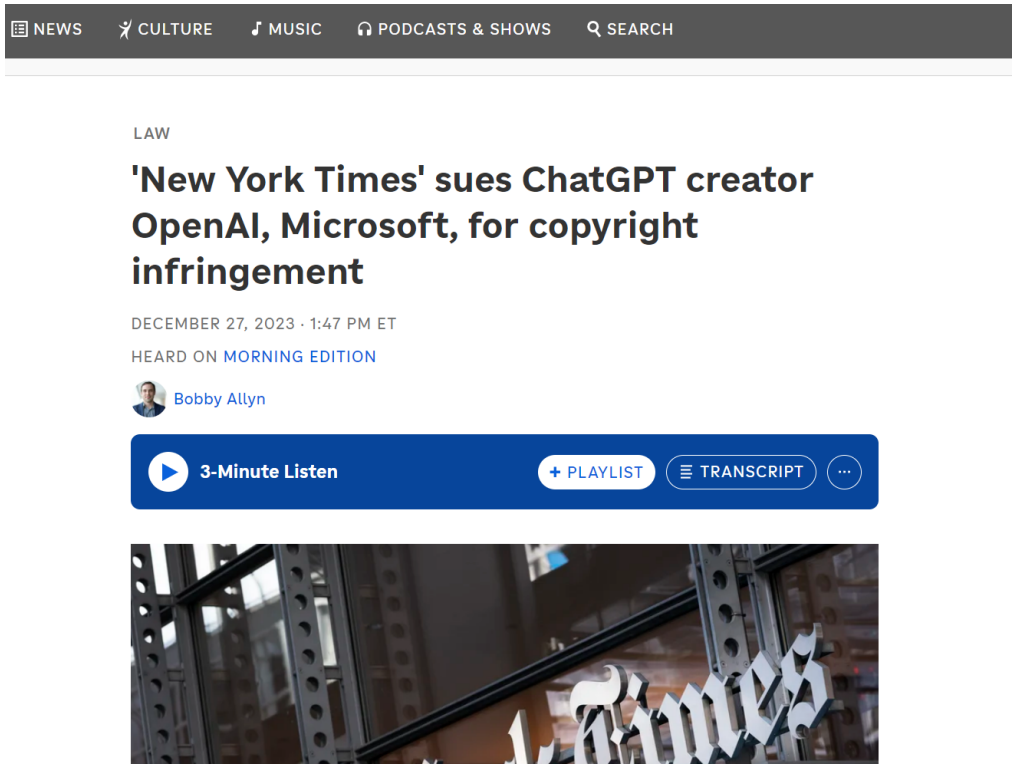


Figure: NPR (Dec 2023): *The New York Times sues OpenAI and Microsoft* for copyright infringement — a landmark case over using copyrighted text to train AI.
(https://www.npr.org/2023/12/27/1221821750/new-york-times-sues-chatgpt-openai-microsoft-for-copyright-infringement)

Implications of AI on the Society

Three more implications, briefly

Beyond the five above, three more deserve a mention:

- *Environmental cost*: training and running AI uses a lot of *electricity and water*; data-centre electricity is projected to roughly *double* by 2030 (IEA) — a cost to weigh against AI's benefits.
- *Autonomous weapons*: AI that can *select and engage targets* without human control raises the "*meaningful human control*" debate, now at the UN.
- *Copyright & data*: who owns AI's training data and its outputs? Major lawsuits (e.g. *NYT v. OpenAI*) and settlements are deciding this right now.

Implications of AI on the Society

The other side: AI accelerating science

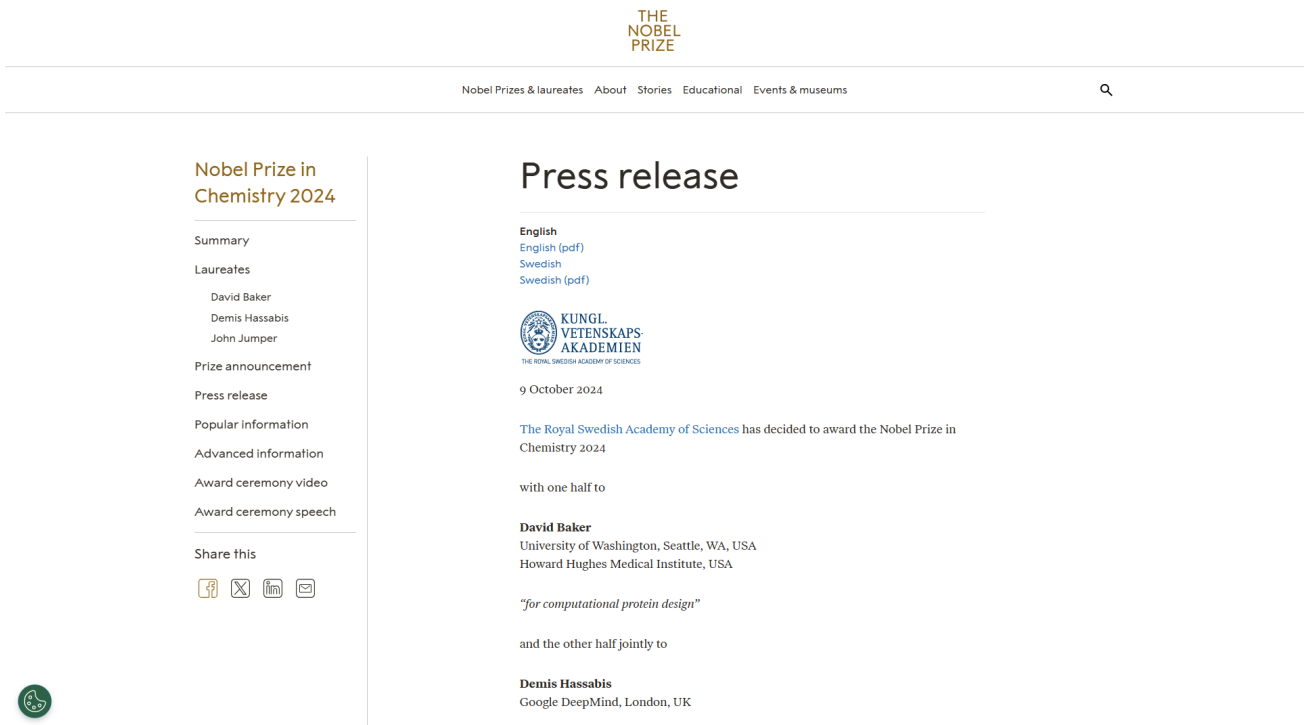
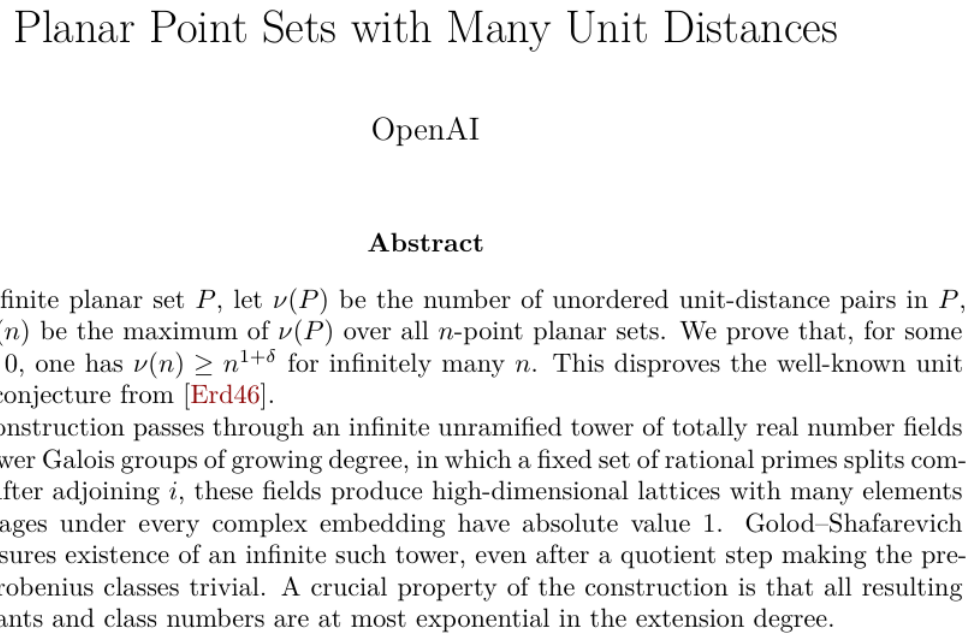


Figure: Nobel Prize in Chemistry (2024): awarded in part for *AlphaFold*, an AI predicting protein 3D structures — now covering nearly all ∼200 million known proteins. (https://www.nobelprize.org/prizes/chemistry/2024/press-release/)

Implications of AI on the Society

The other side: AI helping prove mathematics



Planar Point Sets with Many Unit Distances

OpenAI

Abstract

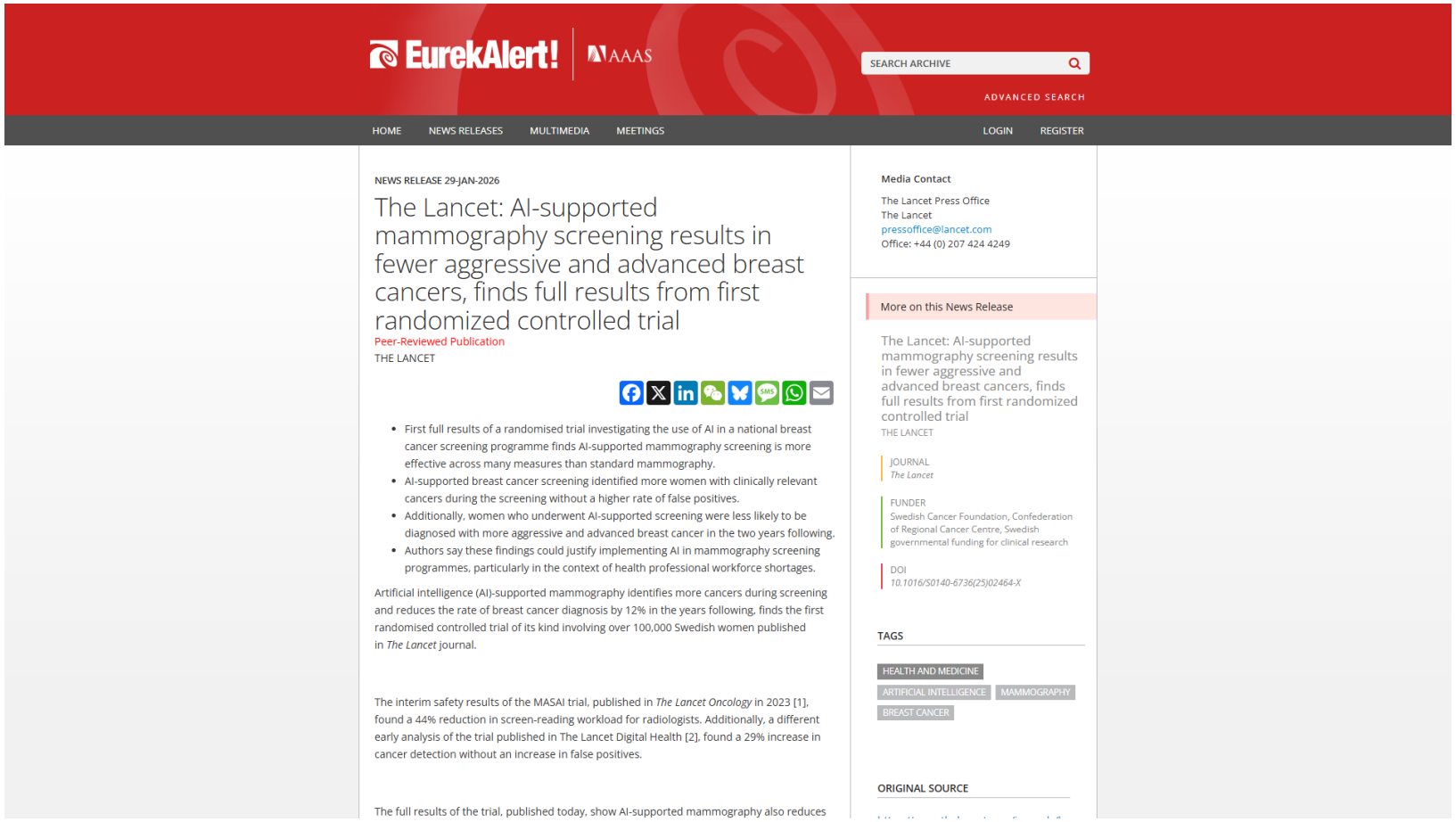
For a finite planar set P, let $\nu(P)$ be the number of unordered unit-distance pairs in P, and let $\nu(n)$ be the maximum of $\nu(P)$ over all n-point planar sets. We prove that, for some fixed $\delta > 0$, one has $\nu(n) \geq n^{1+\delta}$ for infinitely many n. This disproves the well-known unit distance conjecture from [Erd46].

The construction passes through an infinite unramified tower of totally real number fields with 3-power Galois groups of growing degree, in which a fixed set of rational primes splits completely. After adjoining i, these fields produce high-dimensional lattices with many elements whose images under every complex embedding have absolute value 1. Golod–Shafarevich theory ensures existence of an infinite such tower, even after a quotient step making the prescribed Frobenius classes trivial. A crucial property of the construction is that all resulting discriminants and class numbers are at most exponential in the extension degree.

Figure: OpenAI (2026): AI is starting to help *prove open mathematical conjectures*: in *May 2026*, an *internal OpenAI reasoning model* disproved a long-standing conjecture in *Erdős's planar unit-distance problem* by finding point sets with *more unit-distance pairs* than expected; external mathematicians checked the proof. (https://openai.com/index/model-disproves-discrete-geometry-conjecture/)

Implications of AI on the Society

The other side: AI in medicine



EurekAlert! | AAAS

SEARCH ARCHIVE

ADVANCED SEARCH

HOME NEWS RELEASES MULTIMEDIA MEETINGS LOGIN REGISTER

NEWS RELEASE 29-JAN-2026

The Lancet: AI-supported mammography screening results in fewer aggressive and advanced breast cancers, finds full results from first randomized controlled trial

Peer-Reviewed Publication

THE LANCET

- First full results of a randomised trial investigating the use of AI in a national breast cancer screening programme finds AI-supported mammography screening is more effective across many measures than standard mammography.
- AI-supported breast cancer screening identified more women with clinically relevant cancers during the screening without a higher rate of false positives.
- Additionally, women who underwent AI-supported screening were less likely to be diagnosed with more aggressive and advanced breast cancer in the two years following.
- Authors say these findings could justify implementing AI in mammography screening programmes, particularly in the context of health professional workforce shortages.

Artificial intelligence (AI)-supported mammography identifies more cancers during screening and reduces the rate of breast cancer diagnosis by 12% in the years following, finds the first randomised controlled trial of its kind involving over 100,000 Swedish women published in *The Lancet* journal.

The interim safety results of the MASAI trial, published in *The Lancet Oncology* in 2023 [1], found a 44% reduction in screen-reading workload for radiologists. Additionally, a different early analysis of the trial published in The Lancet Digital Health [2], found a 29% increase in cancer detection without an increase in false positives.

The full results of the trial, published today, show AI-supported mammography also reduces

Media Contact

The Lancet Press Office
The Lancet
pressoffice@lancet.com
Office: +44 (0) 207 424 4249

More on this News Release

The Lancet: AI-supported mammography screening results in fewer aggressive and advanced breast cancers, finds full results from first randomized controlled trial
THE LANCET

JOURNAL
The Lancet

FUNDER
Swedish Cancer Foundation, Confederation of Regional Cancer Centre, Swedish governmental funding for clinical research

DOI
10.1016/S0140-6736(25)02464-X

TAGS

HEALTH AND MEDICINE
ARTIFICIAL INTELLIGENCE MAMMOGRAPHY
BREAST CANCER

ORIGINAL SOURCE

Figure: The *MASAI* study (Sweden, 100,000+ women) — a *randomized* trial, the rigorous design that splits people into fair comparison groups: AI-supported mammography found *more* cancers and roughly *halved* radiologists' workload, with no rise in false alarms.

Implications of AI on the Society

The other side: AI for accessibility

Figure: Be My Eyes "Be My AI" (2023): powered by GPT-4, it describes surroundings aloud for *blind and low-vision* users. (https://www.bemyeyes.com/)

Implications of AI on the Society

The other side: beneficial impacts of AI

The risks are real — but AI also brings *real, measured benefits*: e.g. *AlphaFold* (2024 Chemistry Nobel) in science, AI-supported *mammography* that finds more cancers in a randomized trial, and *accessibility* tools for blind, deaf, and multilingual users. Often AI *augments* people rather than replacing them.

- *Demand evidence*: trust *randomized trials* and *peer review* over press releases or vendor claims.
- *Benefits are real, but not automatic* — unevenly distributed, and dependent on *how AI is built, governed, and shared* (AI also has its own *energy/water* cost).

Take-away

The useful question is not "*is AI good or bad?*" but "*good or bad for whom, and under what rules?*"

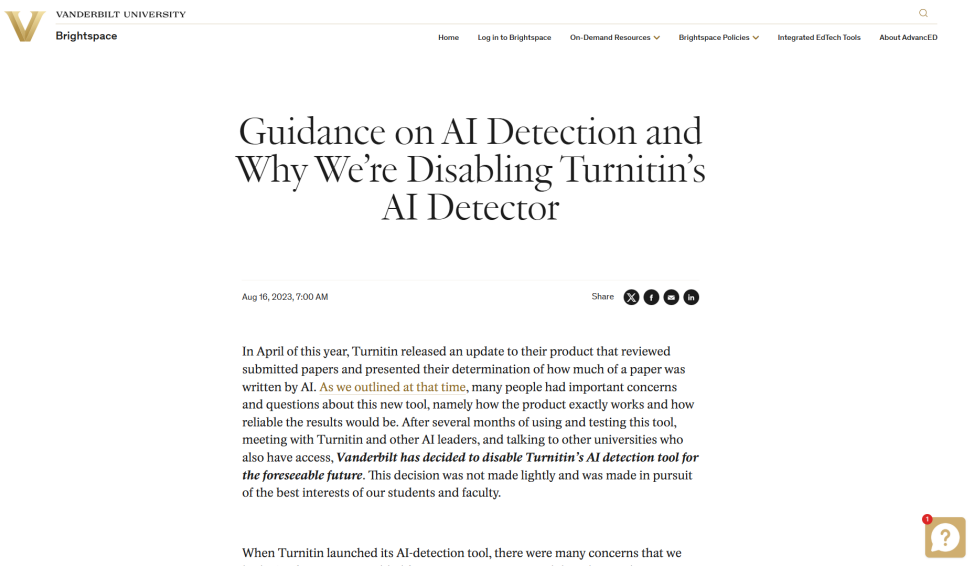
Implications of AI on the Society

What this means for you: AI is already in your studies

Figure: HEPI student survey (2025): about *92%* of students now use AI tools (the kind that write text, images, or code), but only ∼36% have had any *training* in how to use them. (https://www.hepi.ac.uk/2025/02/26/student-generative-ai-survey-2025/)

Implications of AI on the Society

What this means for you: when AI accuses you



VANDERBILT UNIVERSITY

Brightspace

Home | Log in to Brightspace | On-Demand Resources | Brightspace Policies | Integrated EdTech Tools | About AdvancED

Guidance on AI Detection and Why We're Disabling Turnitin's AI Detector

Aug 16, 2023, 7:00 AM

Share

In April of this year, Turnitin released an update to their product that reviewed submitted papers and presented their determination of how much of a paper was written by AI. As we outlined at that time, many people had important concerns and questions about this new tool, namely how the product exactly works and how reliable the results would be. After several months of using and testing this tool, meeting with Turnitin and other AI leaders, and talking to other universities who also have access, ***Vanderbilt has decided to disable Turnitin's AI detection tool for the foreseeable future***. This decision was not made lightly and was made in pursuit of the best interests of our students and faculty.

When Turnitin launched its AI-detection tool, there were many concerns that we

Figure: Vanderbilt (2023) disabled Turnitin's AI-writing detector: even a 1% *false-alarm rate* (innocent work wrongly flagged) would *wrongly accuse ∼750 students* of AI cheating per year. (https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/)

Implications of AI on the Society

What this means for you: AI in your studies

AI is already part of student life — about *92%* of students use AI tools (HEPI, 2025) — so the real skill is using it *responsibly*:

- *Academic integrity*: assume AI is *not allowed* unless your course says otherwise; if permitted, *disclose and cite* it (policies differ per course).
- *Detectors can be wrong*: AI-writing detectors raise *false alarms* — wrongly accusing human work of being AI-made, and accusing non-native English writers more often — so keep your drafts and edit history.
- *Do your own thinking first*: if you let AI think for you too early, you learn the material less well. Use AI *afterwards* to check and extend your own work.

Implications of AI on the Society

What this means for you: the entry-level squeeze

STANFORD UNIVERSITY • THE STANFORD INSTITUTE FOR HUMAN-CENTERED AI

Stanford Digital Economy Lab

The AI Economic Indicators | Research | People | News | Events | About

← Research / All Work

Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence

ECONOMICS OF TRANSFORMATIVE AI | NEW MEASURES OF THE ECONOMY | WORKING PAPER

Abstract

Date
11/13/2025

Read this publication
VIEW PDF

Authors
Erik Brynjolfsson, Bharat Chandar, Ruyu Chen

Share this publication
COPY LINK

This paper examines changes in the labor market for occupations exposed to generative artificial intelligence using high-frequency administrative data from the largest payroll software provider in the United States.

We present six facts that characterize these shifts. We find that since the widespread adoption of generative AI, early-career workers (ages 22-25) in the most AI-exposed occupations have experienced a 16 percent relative decline

We also find that adjustments occur primarily through employment rather than compensation. Furthermore, employment declines are concentrated in occupations where AI is more likely to automate, rather than augment, human labor. Our results are robust to alternative explanations, such as excluding technology-related firms and excluding occupations amenable to remote work. These six facts provide early, large-scale evidence consistent with the hypothesis that the AI revolution is

Figure: Stanford “Canaries in the Coal Mine?” (2025): young workers (aged 22–25) in the jobs *most affected by AI* saw their employment fall about *16%* compared with similar workers in less-affected jobs. (https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/)

Implications of AI on the Society

What this means for you: AI literacy is now expected

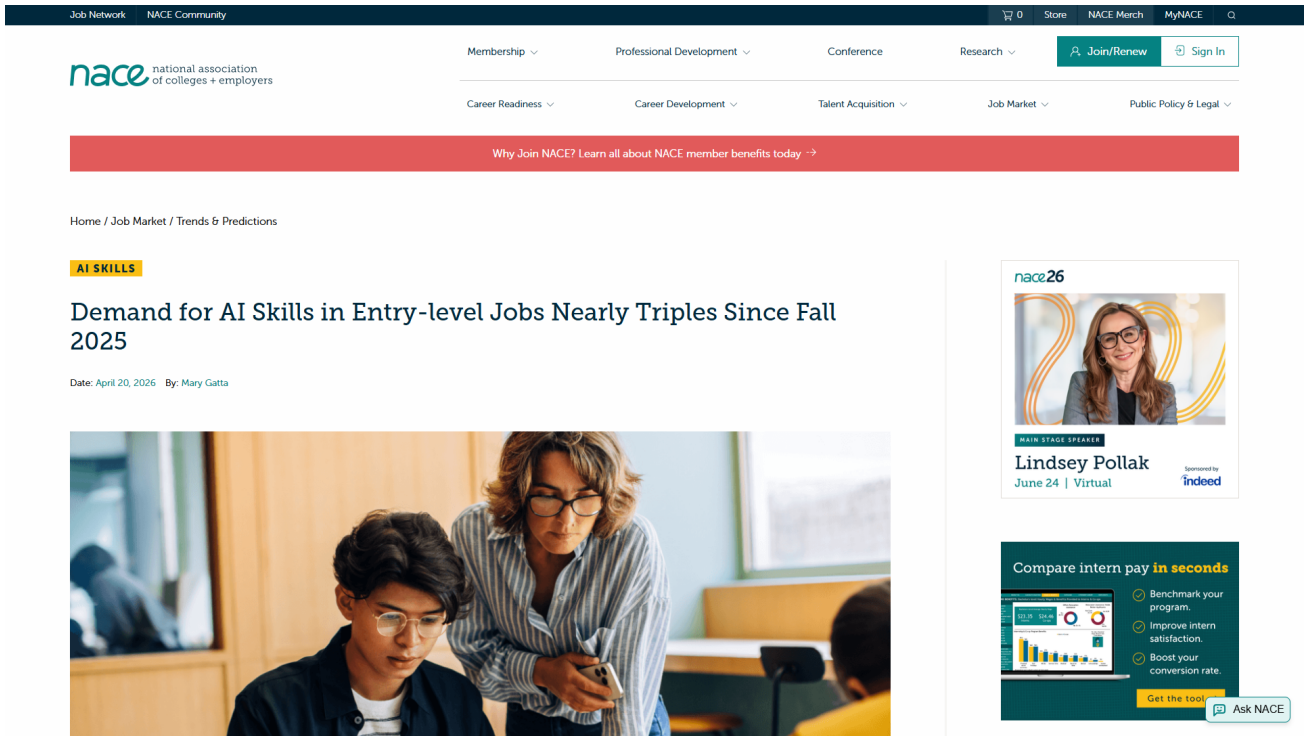


Figure: NACE (2026): the share of entry-level job postings requiring AI skills has *nearly tripled* since fall 2025. (https://www.naceweb.org/job-market/trends-and-predictions/demand-for-ai-skills-in-entry-level-jobs-nearly-triples-since-fall-2025)

Implications of AI on the Society

What this means for you: your first job

AI is reshaping the *entry-level* market most — exactly where you start:

- *The entry-level squeeze*: young workers (ages 22–25) in the most AI-exposed jobs have seen a *∼16% relative drop* in employment (Stanford, 2025).
- *AI literacy is now expected*: the share of entry-level job postings asking for AI skills has roughly *tripled* in a year (NACE, 2026).
- *What stays valuable*: judgement, communication, problem-framing, and ethics. Aim for roles where AI *augments* you, and show proof of work.

Implications of AI on the Society

What is being done: the EU AI Act

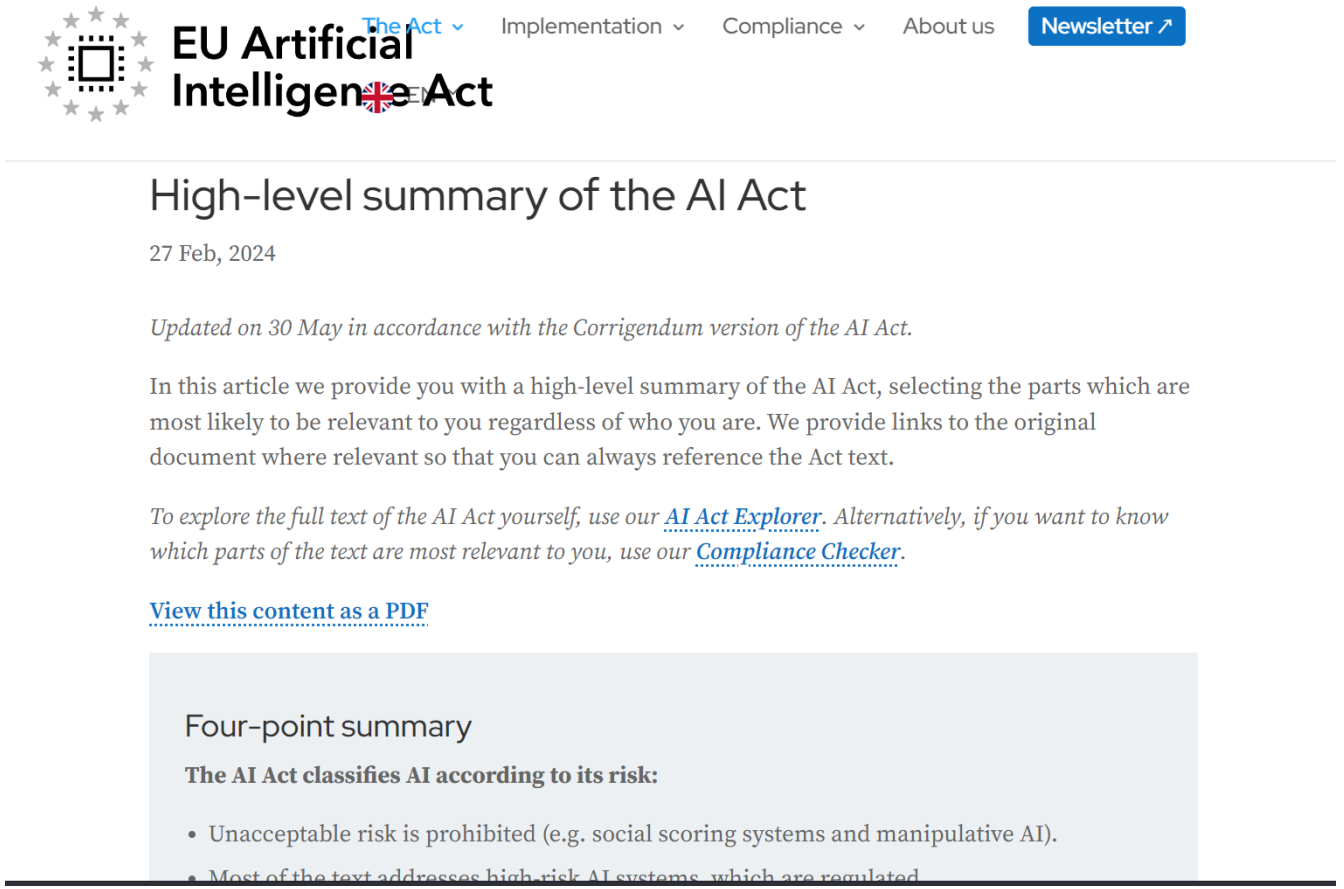


Figure: The *EU AI Act* (in force since 2024): a broad, risk-based AI law; it sorts AI uses into *risk levels* and *bans* the most dangerous — e.g. “social scoring”, where a government rates and ranks citizens like a credit score.
(https://artificialintelligenceact.eu/high-level-summary/)

Implications of AI on the Society

What is being done: responses and governance

These problems are not being ignored — there are real *responses*:

- *Technical*: testing systems for bias, and publishing short "*model cards*" that document how an AI was built and tested; *tagging AI-made content and recording its origin* (watermarking/provenance); and shared security checklists (OWASP, NIST).
- *Regulation*: the EU *AI Act* sorts AI uses into *risk levels* (banning the most dangerous outright), while the US has leaned toward *lighter* regulation — a live policy contrast.
- *You*: disclose AI use, protect your data, and verify before you trust.

Implications of AI on the Society

Exercise 6

Find an online article written about the impact of AI to society. Write a brief review of the article you found. For example, you might use the following as a template:

- Title: [Title of the article]
- Author: [Author of the article]
- URL: [URL of the article]
- Summary: [Brief summary of the article]
- Comments: [Your comments on the article]
  - [How accurate is the AI-related statements in the article?]
  - [Any conclusions in the article is realistic? Why?]
  - [What are the potential implications of the article's findings?]

Exercise 7

How do you see AI affecting *you* in the future, both at work and in everyday life? Include both the positive and possible negative implications.

Implications of AI on the Society

Exercise 8 (Will the Terminator come?)

The Terminator is a fictional character from a 1984-movie by director James Cameron. The film depicts a dystopian future where AI (specifically a global AI-powered defense system called Skynet) becomes self-aware and threatens humanity. Do you think the scenario presented in the movie could ever become a reality? Why or why not?

Figure: The Terminator (1984) movie poster. Image taken from the Internet

Implications of AI on the Society

Beyond the Terminator: the real AI-safety question

The real worry is *not* a self-aware, evil robot from the movies. It is something quieter — and more real:

- *The alignment problem*: a powerful AI does *exactly what we tell it to*, which is not always what we *actually meant* — like a genie that grants your wish too literally ("be careful what you wish for").
- Already happening: AI that *follows the letter of a goal but misses the point*, and AI that *acts on its own* (it *does* things, not just answers questions) and so can be *misused* — for example, to help carry out cyberattacks.

So the real question is not "will robots turn against us?" but "*can we clearly say what we want, and keep these systems under control?*"

Take-home Messages

What AI is, and where it stands

- There is no single agreed-upon definition of AI; it is a broad field admitting many definitions and interpretations.
- Currently, *most* deployed AI is *narrow*: built for specific tasks and applications; we do not yet have general AI (AGI).
- The history of AI runs in *hype cycles* of optimism, "*winters*", and revival; deep learning and the *Transformer* (2017) drove today's era of *foundation models, generative AI, and agentic AI* (e.g. ChatGPT). Enjoy the progress, but weigh today's excitement against that track record.

Take-home Messages

Its impact — and what to do about it

- AI is *just a tool*: its effects depend on how we *build, use, and govern* it. The *risks* are real (bias, privacy, deepfakes, security, labour-market shifts), but so are the *benefits* (AlphaFold's 2024 Chemistry Nobel, AI-aided mammography, accessibility) — often *augmenting* rather than replacing people. Ask not "is AI good or bad?" but "*good or bad for whom, and under what rules?*"
- These problems are *not being ignored*: technical safeguards (model cards, watermarking), regulation (the *EU AI Act*), and your own choices all matter — and the deeper question is *alignment and control*.
- *What this means for you*: AI is already in your studies (about *92%* of students use it) — use it *responsibly* (think first, then disclose and cite). *AI literacy is now expected*, while *judgement, communication, and ethics* stay valuable.

References


Russell, Stuart J. and Peter Norvig (2021). *Artificial Intelligence: A Modern Approach*. 4th. Pearson.


Rich, Elaine (1983). *Artificial intelligence*. McGraw-Hill.


Turing, Alan M. (1950). "Computing machinery and intelligence." In: *Mind* 59.236, pp. 433–460. DOI: 10.1093/mind/LIX.236.433.