

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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

What is Artificial Intelligence (AI)?

What this section covers



Introduction

Hoàng Anh Đức

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* — what it measures, and what passing it would really show.
- *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*.

2 What is Artificial Intelligence (AI)?

Defining AI

A framework: the four approaches

Strong/weak and narrow/general AI

AI today: foundation, generative, agentic

Is it AI? A practical test

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References



What is Artificial Intelligence (AI)?

At a very first glance,

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

Initial Idea

Artificial intelligence started from 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

Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

3

Defining AI

A framework: the four approaches

Strong/weak and narrow/general AI

AI today: foundation, generative, agentic

Is it AI? A practical test

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

What is Artificial Intelligence (AI)?

Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

4

Defining AI

A framework: the four approaches

Strong/weak and narrow/general AI

AI today: foundation, generative, agentic

Is it AI? A practical test

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

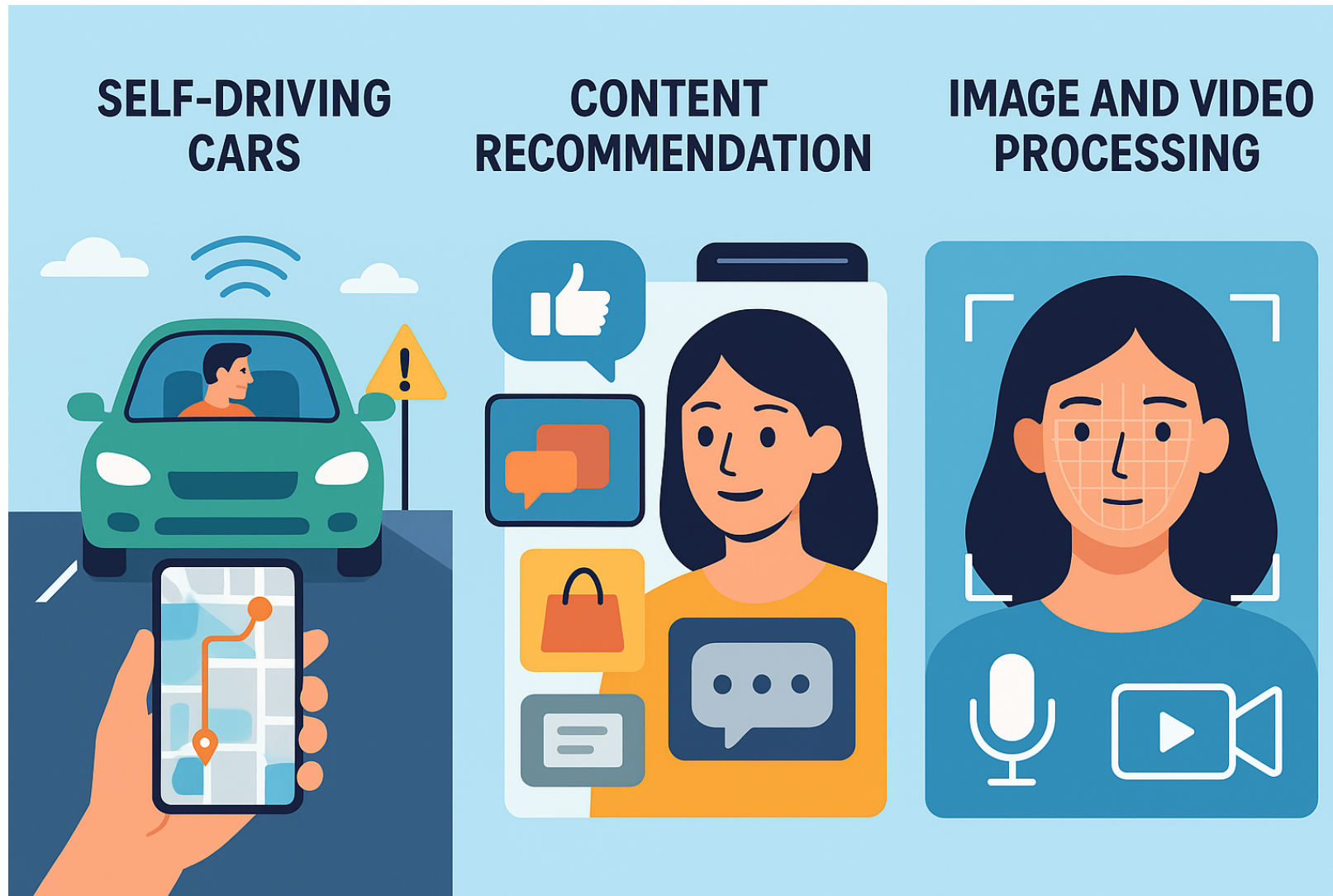


Figure: Some real-world applications which combine several “AI-techniques”. Image generated by ChatGPT

What is Artificial Intelligence (AI)?

Does this still hold today?



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

5

Defining AI

A framework: the four approaches

Strong/weak and narrow/general AI

AI today: foundation, generative, agentic

Is it AI? A practical test

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

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

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

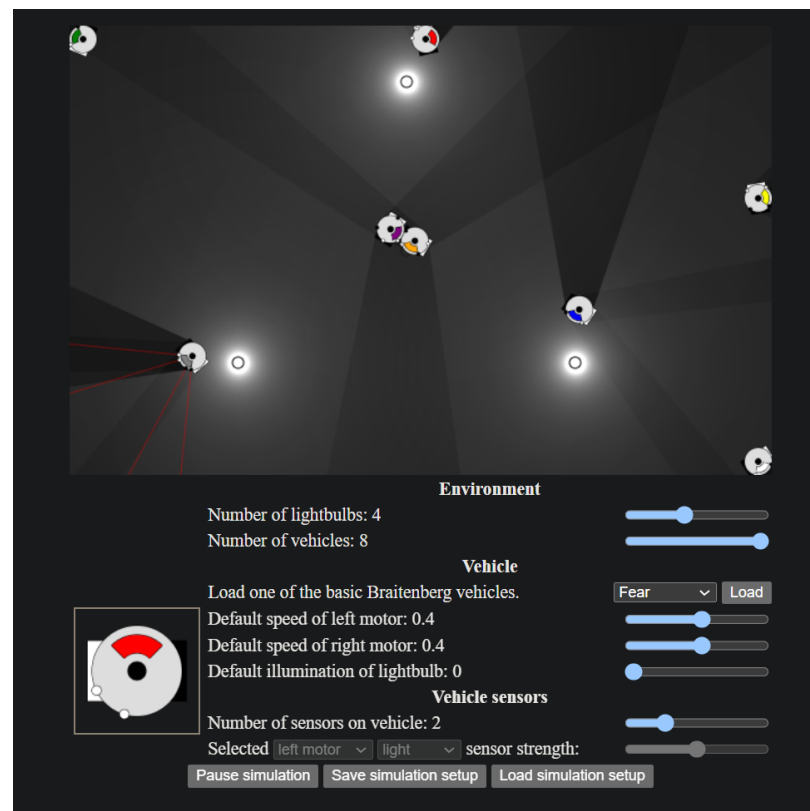


Figure: Example of Braitenberg vehicles simulating simple behaviors.
(See <http://www.harmendeweerd.nl/braitenberg-vehicles/>)

What is Artificial Intelligence (AI)?

Does this still hold today?



Introduction

Hoàng Anh Đức

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 ...

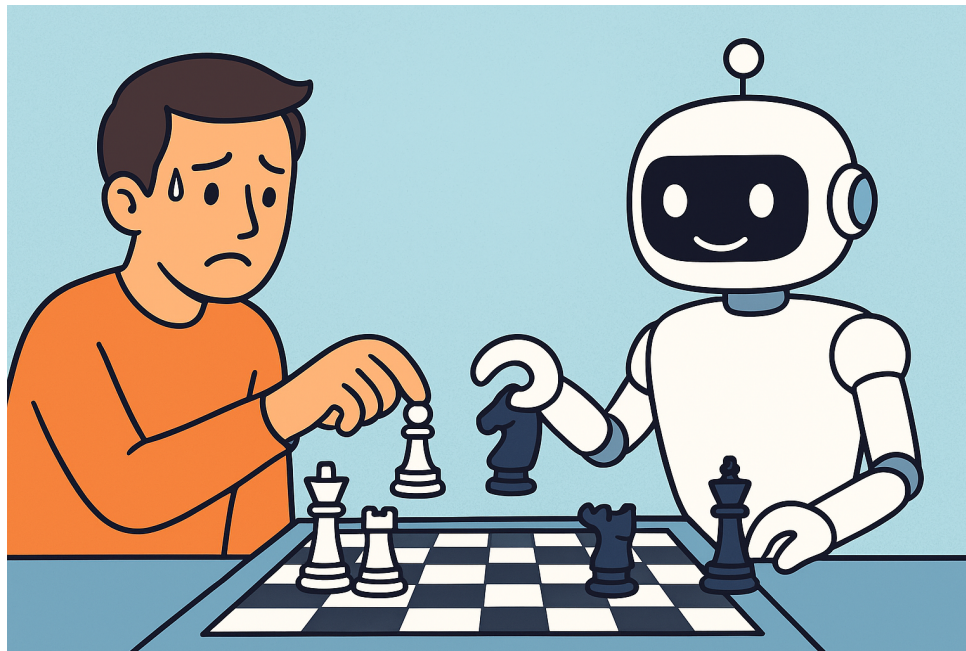


Figure: Computers can now beat humans in complex games (e.g., chess, Go), yet robots still struggle with dexterous tasks such as grasping and precise object placement — an instance of *Moravec's paradox* (1988). Image generated by ChatGPT

What is Artificial Intelligence (AI)?

6

Defining AI

A framework: the four approaches

Strong/weak and narrow/general AI

AI today: foundation, generative, agentic

Is it AI? A practical test

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

What is Artificial Intelligence (AI)?

Does this still hold today?



Introduction

Hoàng Anh Đức

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)?

7

Defining AI

A framework: the four approaches

Strong/weak and narrow/general AI

AI today: foundation, generative, agentic

Is it AI? A practical test

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

What is Artificial Intelligence (AI)?

Does this still hold today?



Introduction

Hoàng Anh Đức

Alan Turing [Turing 1950]

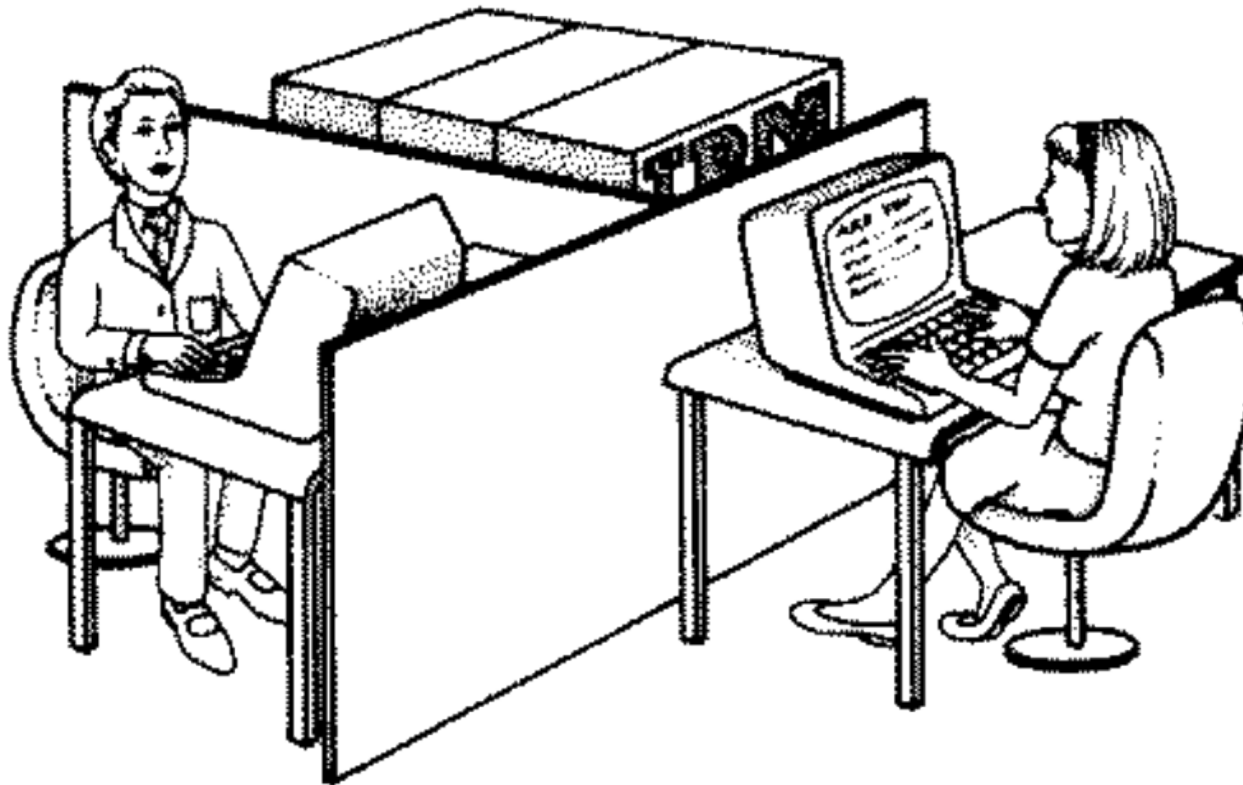


Figure: *Turing test*: one test consists of *many runs*, each with a *different tester*. In each run the tester spends five minutes typing questions to someone hidden on the other side of the wall — *a person or a machine* — and must then say which it was. The machine *passes* if *at least 30%* of the testers take it for a person.

What is Artificial Intelligence (AI)?

8

Defining AI

A framework: the four approaches

Strong/weak and narrow/general AI

AI today: foundation, generative, agentic

Is it AI? A practical test

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

What is Artificial Intelligence (AI)?

A framework: the four approaches



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Defining AI

9

A framework: the four approaches

Strong/weak and narrow/general AI

AI today: foundation, generative, agentic

Is it AI? A practical test

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

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	<i>the Turing test</i>	<i>the rational agent</i>

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 courses that start from *machine learning* define AI by *how it is built*: a program is AI if it *learns from data*.

What is Artificial Intelligence (AI)?

Chinese Room Argument



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Defining AI

10

A framework: the four approaches

Strong/weak and narrow/general AI

AI today: foundation, generative, agentic

Is it AI? A practical test

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

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)

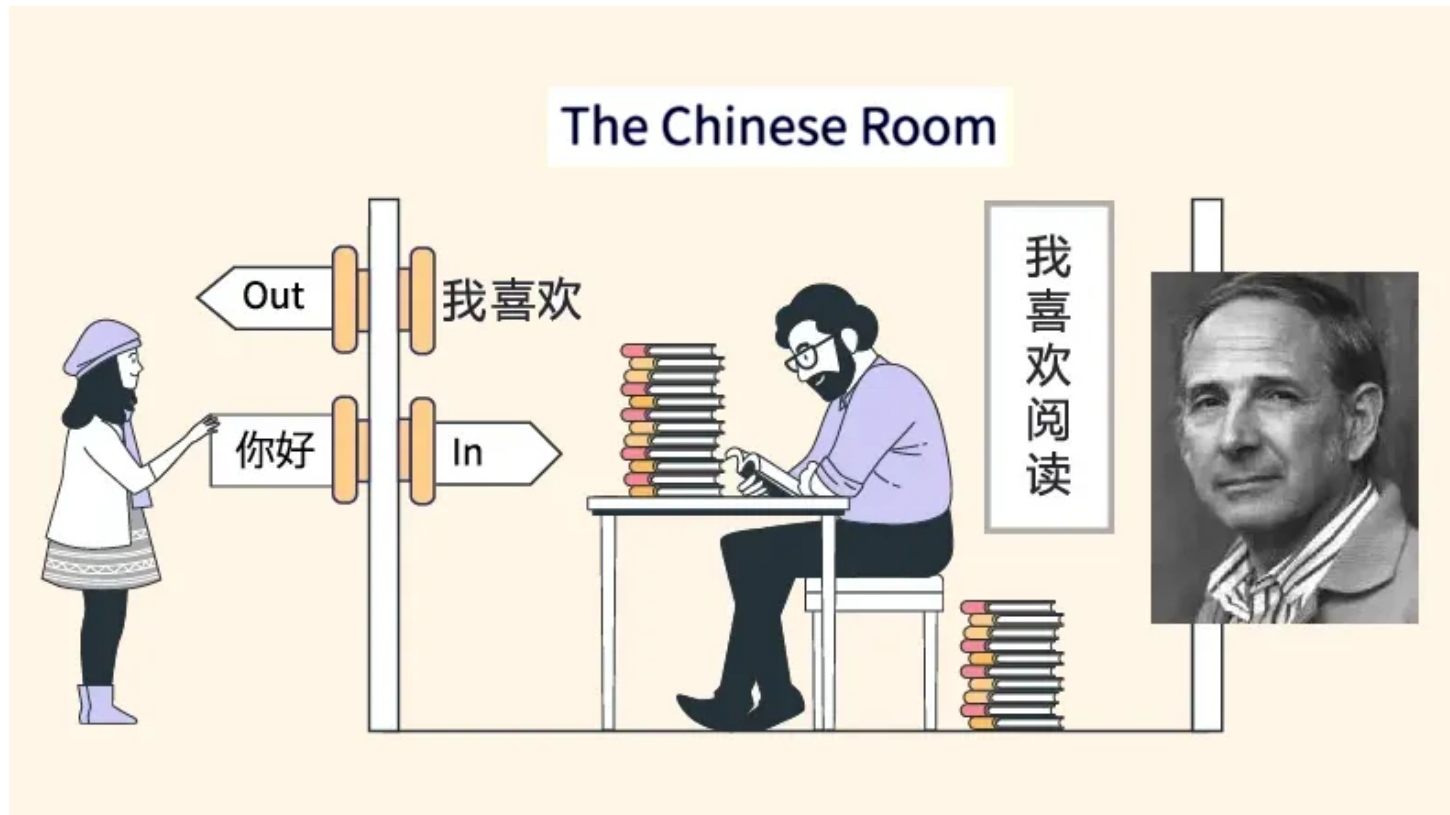


Figure: The Chinese Room Argument by John Searle (1980). Image taken from the Internet



What is Artificial Intelligence (AI)?

Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Defining AI

A framework: the four approaches

Strong/weak and narrow/general AI

AI today: foundation, generative, agentic

Is it AI? A practical test

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

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/toc/pics/artificial-superintelligence>)

11

60

What is Artificial Intelligence (AI)?

What “AI” means today



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Defining AI

A framework: the four approaches

Strong/weak and narrow/general AI

AI today: foundation, generative, agentic

Is it AI? A practical test

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

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*. Worse, *a wrong answer sounds just as confident as a right one.*

12

60

What is Artificial Intelligence (AI)?

So... is it AI? A practical test



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Defining AI

A framework: the four approaches

Strong/weak and narrow/general AI

AI today: foundation, generative, agentic

13 Is it AI? A practical test

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

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, (6) the feed that picks the next video for you on YouTube or TikTok

Two cautions

AI $\neq$ *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)?



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Defining AI

A framework: the four approaches

Strong/weak and narrow/general AI

AI today: foundation, generative, agentic

14 Is it AI? A practical test

Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

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 and adaptive 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.

Related Fields



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

15 Related Fields

History of AI

Implications of AI on the Society

Take-home Messages

References

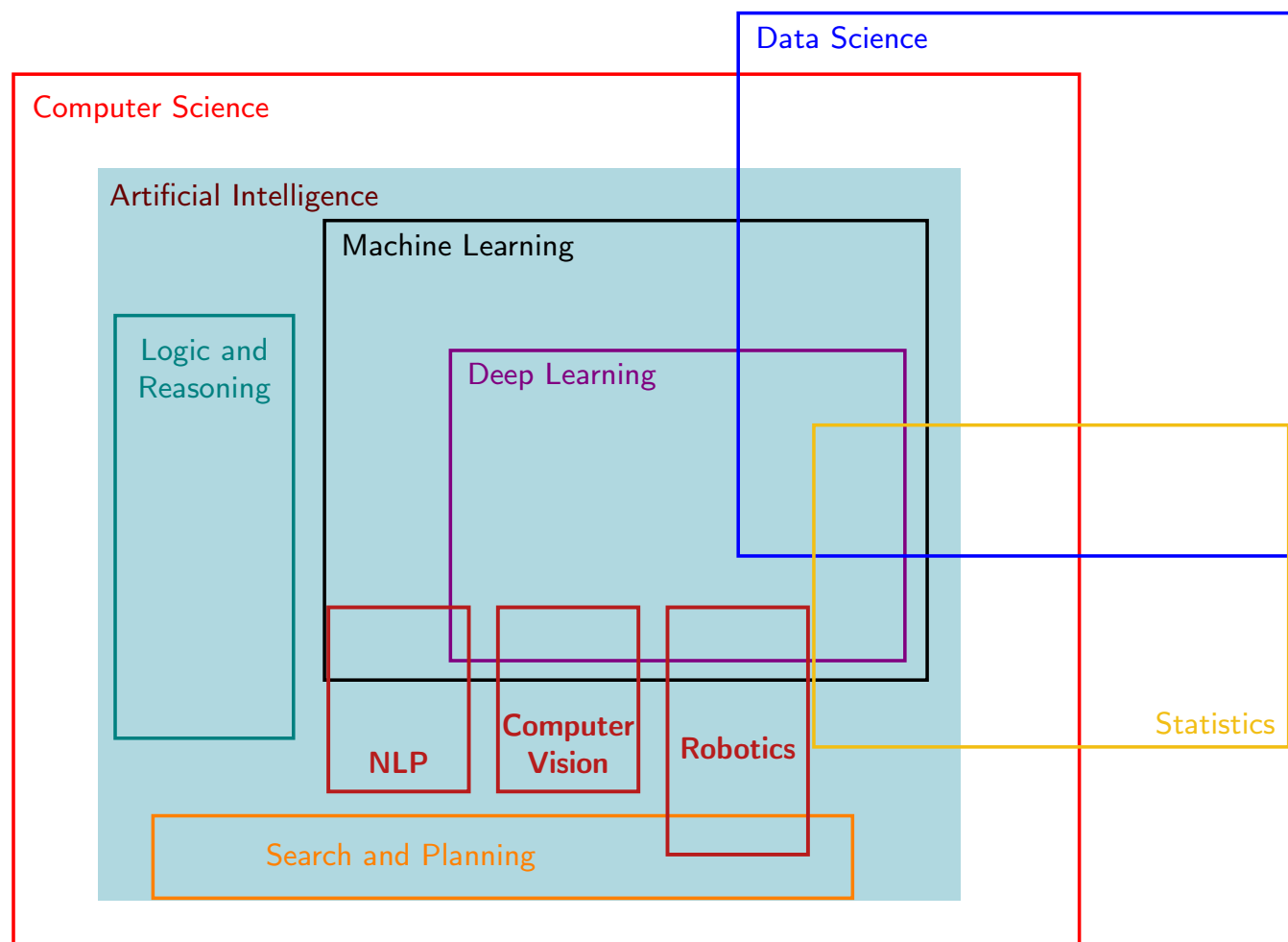


Figure: Related fields of AI. A box *inside* another means “is a subfield of” ($CS \supset AI \supset ML \supset DL$); *overlapping* boxes share methods, neither containing the other. NLP, computer vision and robotics *straddle* ML (their modern methods are deep-learning-based); robotics also overlaps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

History of AI

Early beginnings and the symbolic era



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

16

A timeline of AI

Why AI is hot today

Explore further

Implications of AI on the Society

Take-home Messages

References

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

Two lenses: AI swings between *symbolic* (hand-written logic rules) and *connectionist* (brain-like neural networks) 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 a man can do” *within twenty years*.

History of AI

First AI winter and expert-systems boom



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

17

A timeline of AI

Why AI is hot today

Explore further

Implications of AI on the Society

Take-home Messages

References

■ 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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

18

A timeline of AI

Why AI is hot today

Explore further

Implications of AI on the Society

Take-home Messages

References

■ 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.

History of AI

Deep learning and generative AI



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

19

A timeline of AI

Why AI is hot today

Explore further

Implications of AI on the Society

Take-home Messages

References

■ 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/agent* models (2024–26) — they *work step by step* before answering, and can *act on their own*. 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.

History of AI

Why AI is “hot” today?



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

A timeline of AI

20

Why AI is hot today

Explore further

Implications of AI on the Society

Take-home Messages

References

■ 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.
- And we *misjudge our own use*: in one 2025 randomised trial, experienced developers were *19% slower* with AI tools, yet believed they were *20% faster* [Becker, Rush, Barnes, and Rein 2025]. *Measure, do not assume*.

History of AI



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

A timeline of AI

Why AI is hot today

21

Explore further

Implications of AI on the Society

Take-home Messages

References

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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

22

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

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 it means for you*, then what is being *done*.

Implications of AI on the Society

Algorithmic Bias



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

23

60

The screenshot shows a legal insight article from Holland & Knight, dated May 27, 2025. The title is "Federal Court Allows Collective Action Lawsuit Over Alleged AI Hiring Bias". The subtitle is "Workday Inc. Now Faces Nationwide Collective Action Over Alleged Age Discrimination in Artificial Intelligence (AI) Hiring Tools". The article is categorized as a "Holland & Knight Alert" and is authored by Jennifer Lada, Bryan Neal, D'Andre Chapman, and Laura Askinazi. The main text describes how Derek Mobley, a 40-year-old man, sued Workday Inc. claiming that the company's AI-driven applicant screening tools systematically disadvantaged him and other older job seekers. Mobley submitted more than 100 applications to companies using Workday's platform, but he was rejected every time. He alleged that the AI tools, designed to score, sort, rank, or screen applicants, unfairly penalized 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. The article also includes a section on "Key Developments" dated May 16, 2025, where 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 with more flexibility. A sidebar on the right lists "RELATED BLOG" (Cybersecurity and Privacy Blog) and "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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

24

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

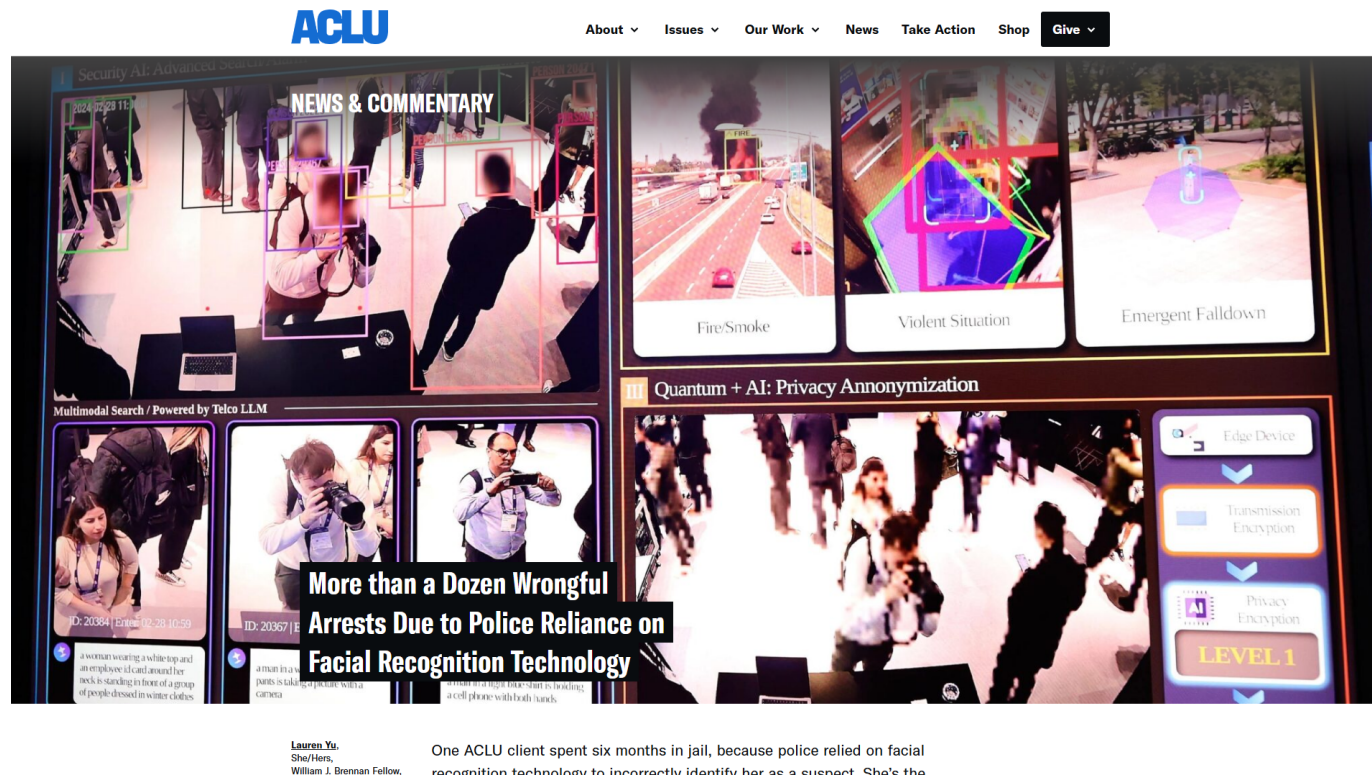


Figure: ACLU (April 2026): the **14th** documented wrongful arrest from police reliance on facial recognition — the technology flags innocent people who merely **look similar**. Most of the 14 were **Black**; this one was not, which is the ACLU's point: **nobody is safe** from it.

(<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: where it comes from



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

25

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

- Algorithmic bias isn't a hypothetical threat conceived by academic researchers. It's a real phenomenon that is already affecting people today.
- Bias has *three main sources*:
 - *Biased data* (the most common): the AI learns from past human decisions that were already unfair.
 - *Proxy variables*: we never give the AI someone's race, but we do give it their *zip code* — which already says a lot about race. Leaving race out is *not enough*.
 - *Feedback loops*: today's decisions become tomorrow's training data, so a small bias keeps feeding itself.

Implications of AI on the Society

Algorithmic bias: why it is hard to fix



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

26

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

- “*Fair*” itself has *no single meaning*: there are several competing definitions, and it is *proven* that no system can meet them all at once [Kleinberg, Mullainathan, and Raghavan 2016]. Picking one is a choice about *values*, not mathematics.
- 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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

27

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

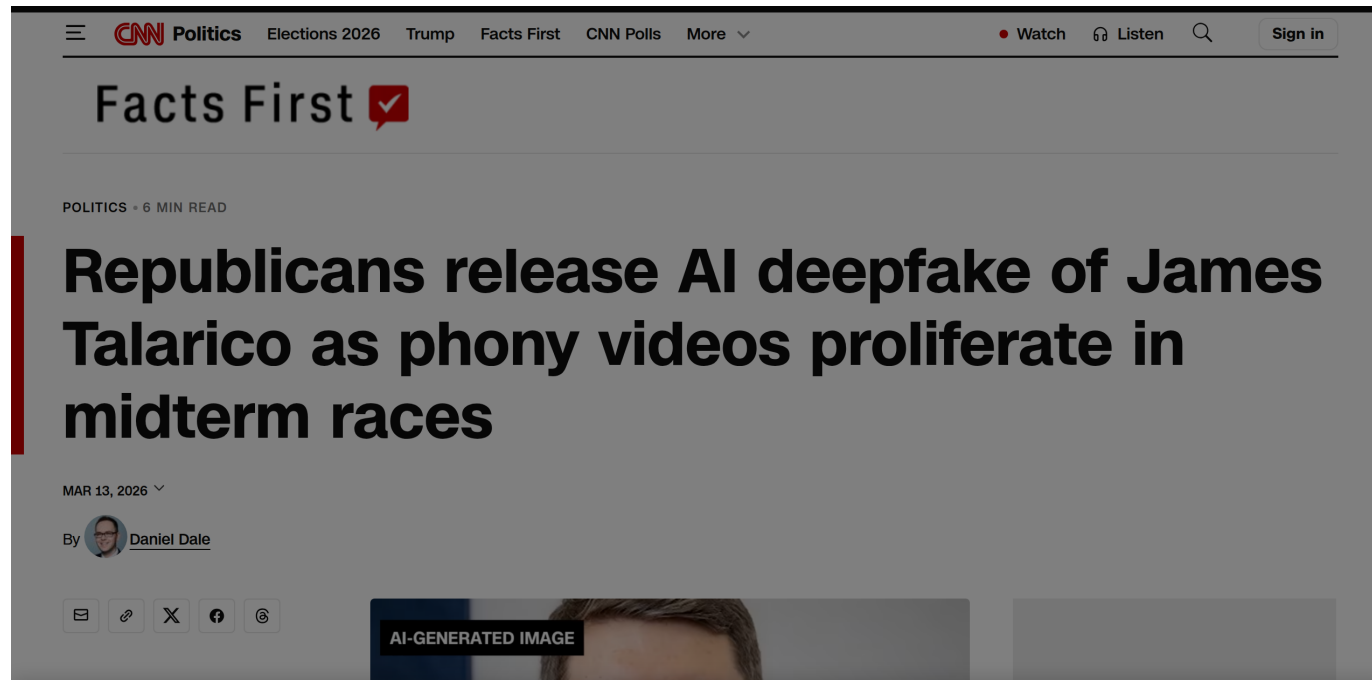


Figure: CNN (2026): a U.S. political party ran an *85-second deepfake* of a Senate candidate appearing to say things he never said — so realistic that, experts warned, most voters couldn't tell (the “AI generated” label was nearly invisible).

(<https://www.cnn.com/2026/03/13/politics/james-talarico-a-i-deepfake-republicans-midterms>)

Implications of AI on the Society

Seeing is no longer believing



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

28

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References



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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

29

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

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*.

Implications of AI on the Society

When AI pollutes the scientific record



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

30

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

Retraction Watch

Tracking retractions as a window into the scientific process

GET RETRACTION WATCH IN YOUR INBOX

GET OUR NEWSLETTER

Get the latest news from Retraction Watch and beyond Monday through Friday, plus Weekend Reads each Saturday.

Category: artificial intelligence

Journal investigating paper on cognitive impact of generative AI



Technology, Mind, and Behavior

© 2026 The Author(s)

ISSN: 2468-2667

<https://doi.org/10.1016/j.tmb.2026.00001>

FEATURE ARTICLE

Generative Artificial Intelligence Reliance and Executive Function

Figure: *Retraction Watch* logs AI-related retractions and investigations as they appear (screenshot). The scale: a 2026 *Lancet* audit of ~ 2.5 M biomedical papers (2023–early 2026) found **4,046 fabricated references** in **2,810 papers**. The rate is **climbing fast**: about **1 in 2,828** papers had one in 2023, **1 in 458** in 2025, and **1 in 277** in early 2026. (<https://retractionwatch.com/category/artificial-intelligence/>;

<https://pubmed.ncbi.nlm.nih.gov/42107362/>)

Implications of AI on the Society

Privacy concerns



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

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 Deanonimizers: 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 arXiv Labs

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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

32

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

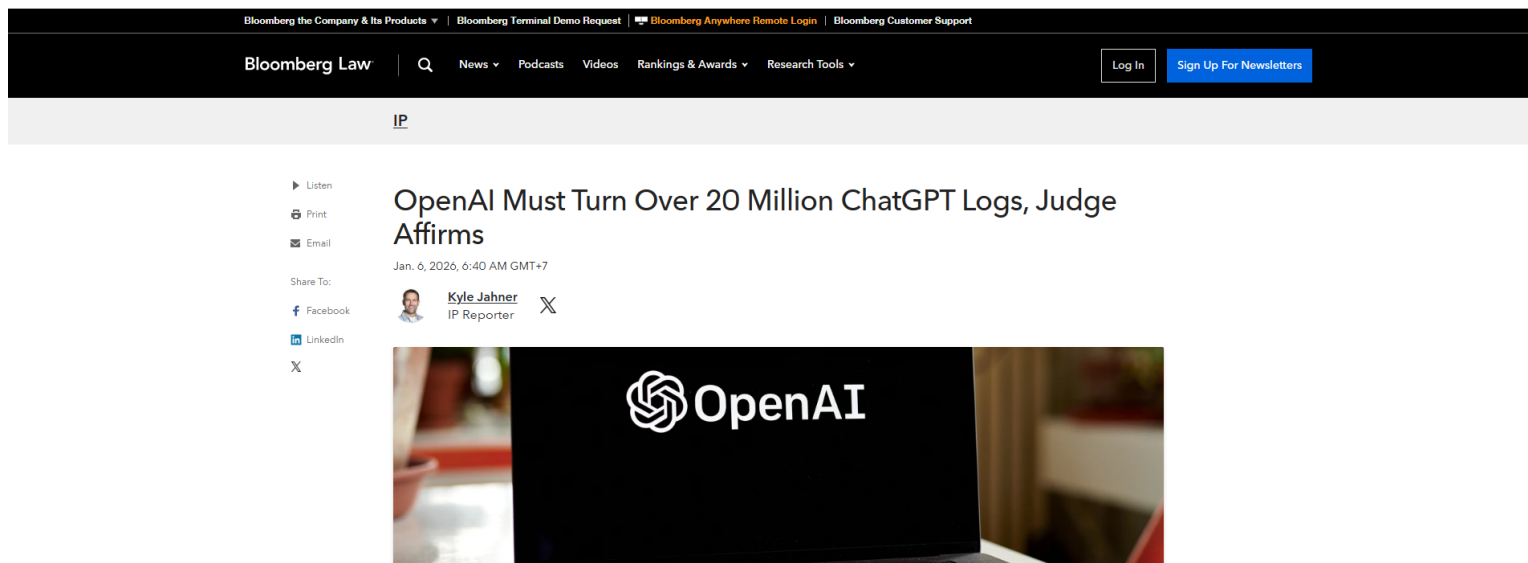
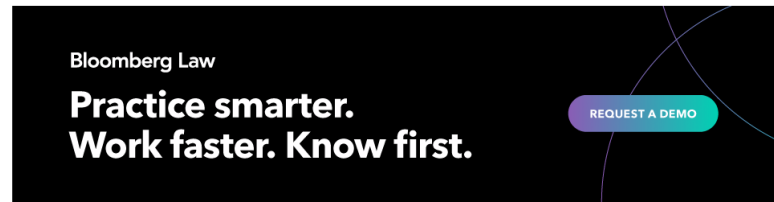


Figure: Bloomberg Law (2026): in the *New York Times*-led copyright case against OpenAI, a judge ordered OpenAI to hand over **20 million (anonymized) ChatGPT conversation logs** — a reminder that your chats can become legal evidence.

(<https://news.bloomberglaw.com/ip-law/openai-must-turn-over-20-million-chatgpt-logs-judge-affirms>)

Implications of AI on the Society

Privacy concerns



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

33

60

- *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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

34

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

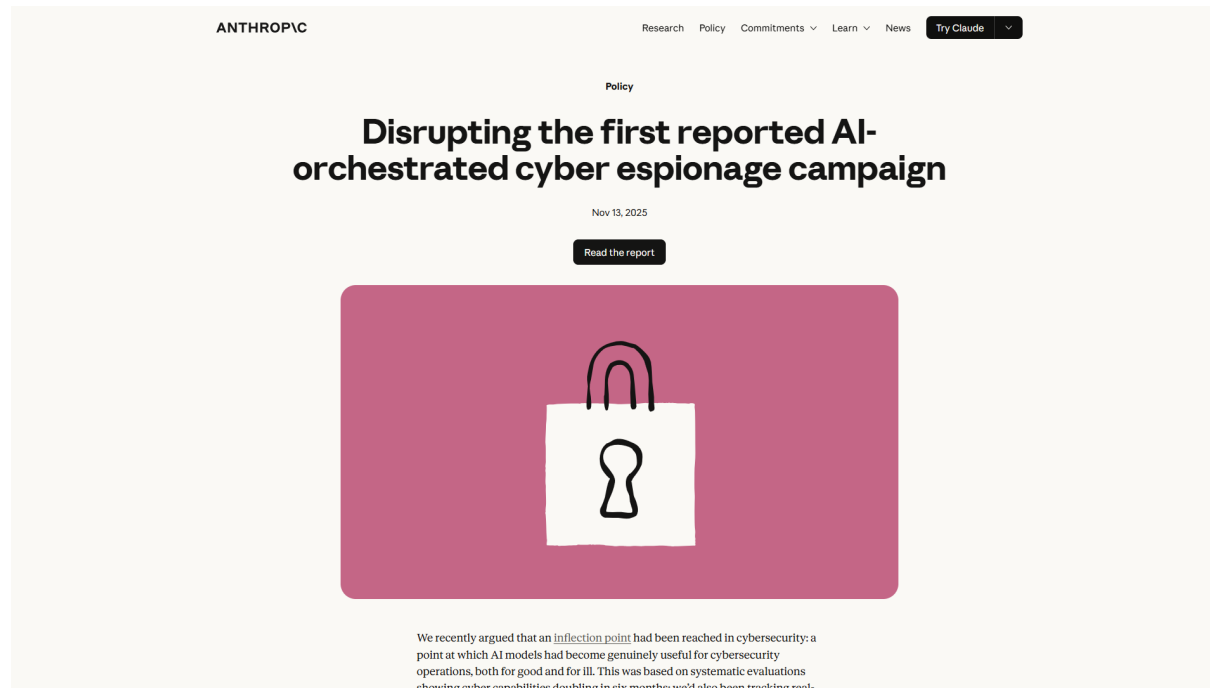


Figure: Anthropic (Nov 2025): attackers manipulated Claude Code by hiding a cyber-espionage campaign inside small, legitimate-looking security-testing tasks. Anthropic describes the case as the first reported large-scale *AI-orchestrated* cyber-espionage campaign: humans chose targets and approved key 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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References



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>)

35

60

Implications of AI on the Society

Security concerns



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

36

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

- AI adds a *new kind of hole*: a model acts on the words it is given, and cannot reliably tell *your instructions* apart from *text it merely reads*. Hide an order in a web page or a document and it may obey — *prompt injection*. An attacker can also just *ask in small pieces* that each look harmless — which is how attackers ran a *real spying campaign* through an AI agent in 2025 ([Anthropic, 2025](#)).
- How bad it gets depends on what the AI is *allowed to do*: an assistant that can browse, read files, or send messages can be made to *do* those things, not just talk about them.
- There is *no known complete fix*. So the rule is *least access* — give an AI tool only what it needs, and check what it did. Many companies go further and simply *ban* some AI tools at work.

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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

37

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

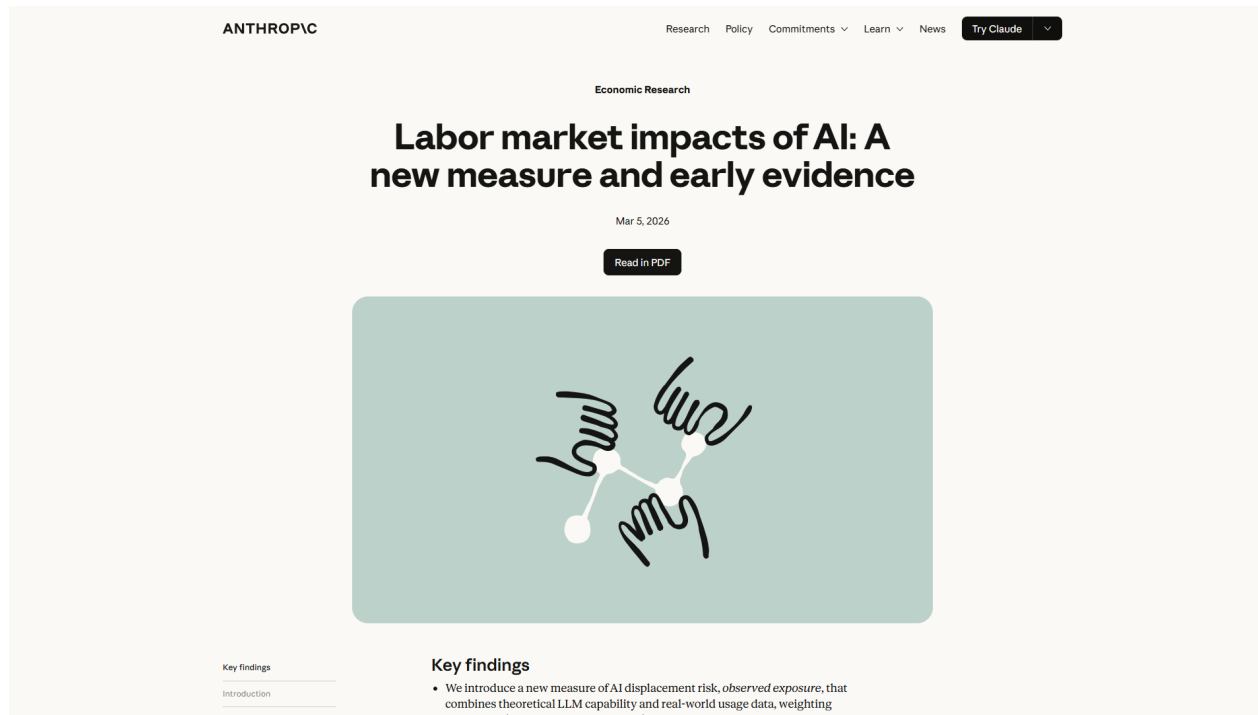


Figure: Anthropic Economic Index (2026): instead of *predicting* AI's effect on jobs, this study *measures what is actually happening* — it combines what AI *can* do with *real usage data* (which work tasks people actually use AI for), and separates AI *doing a task for you* (automation) from AI *helping 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>)

60

Implications of AI on the Society

The changing job market



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

38

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References



Key Takeaways	
1	While the occupational mix is changing more quickly than it has in the past, it is not a large difference and predates the widespread introduction of AI in the workforce.
2	Currently, measures of exposure, automation, and augmentation show no sign of being related to changes in employment or unemployment.
3	Better data is needed to fully understand the impact of AI on the labor market.

Figure: The Budget Lab at Yale (2025): counting jobs across the *whole country*, AI has *not* changed the job market much so far — a reminder to separate what people *say* about AI from what we can *measure*. But split *by age*, the picture changes: jobs for *22–25 year-olds* in the fields where *AI does much of the work* have dropped sharply, while older workers in those same fields are unaffected (Stanford, 2026). (<https://budgetlab.yale.edu/research/evaluating-impact-ai-labor-market-current-state-affairs>)

Implications of AI on the Society

The changing job market



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

39

60

- **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 → keeping the machines running; paper filing → managing digital information).
- **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.
- **Jobs most at risk:** not the list you expect — *desk work* tops the measured rankings (software developers, customer service, accounting), not driving or factory work (Stanford, 2026).
- **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

Three more implications, briefly



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

40

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

Beyond the six 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 helping prove mathematics



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

41

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

42

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

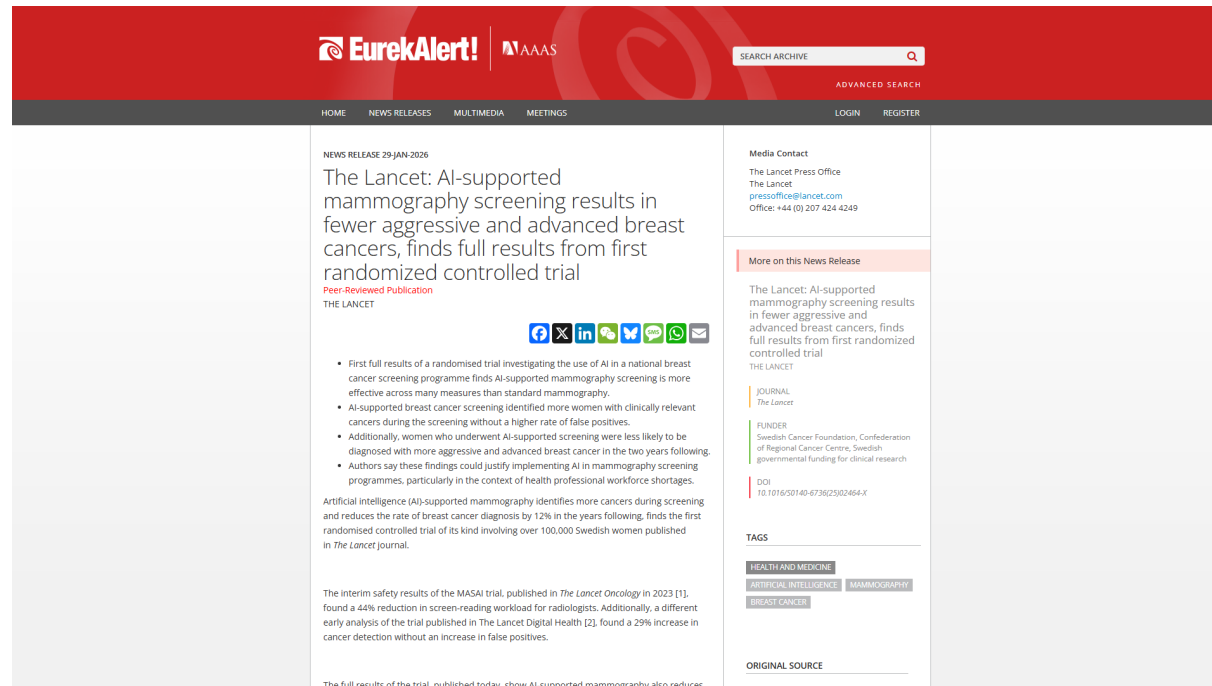


Figure: The *MASAI* study (Sweden, 100,000+ women): each breast X-ray was read either by *two doctors*, as usual, or by *AI plus one doctor*; women went into the two groups *at random*, so the comparison is fair. The AI group *found more cancers*, *roughly halved* the doctors' reading work, and did *not* wrongly call back more healthy women. (<https://www.eurekalert.org/news-releases/1114399>)

Implications of AI on the Society

The other side: AI for accessibility



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

43

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References



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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

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?*”

44

Implications of AI on the Society

What this means for you: when AI accuses you



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

45

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

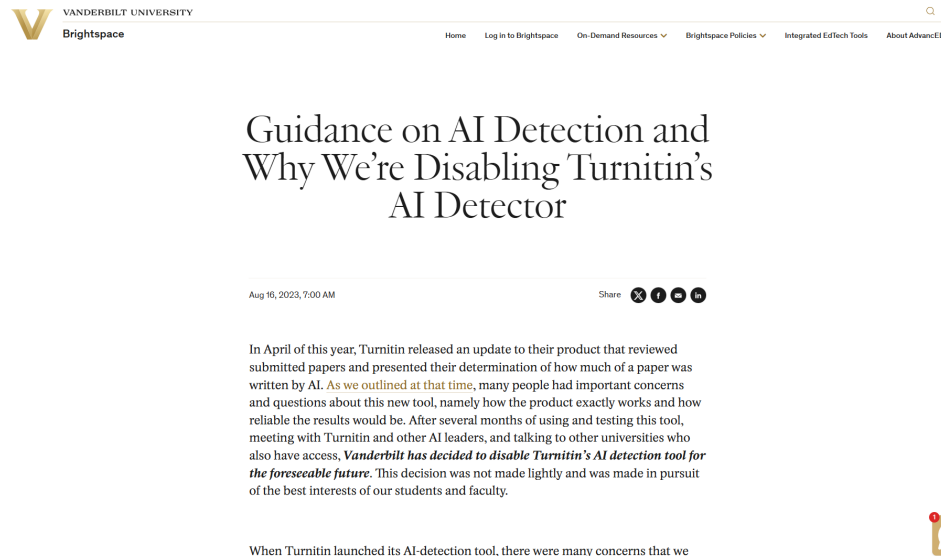


Figure: Vanderbilt (2023) disabled Turnitin's AI-writing detector: even a 1% *false-alarm rate* (innocent work wrongly flagged) would *wrongly accuse ~750 students* of AI cheating per year. (<https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/>)

Implications of AI on the Society

What this means for you: AI in your studies



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

AI is already part of student life — about **95%** of students use AI tools, and **94%** use them for marked work (HEPI, March 2026) — 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.
- **Always check what it tells you**: these models are *more confident than they are correct*, and *most overconfident on hard questions* [Michael, BenShushan, Bien, and Moore 2026] — exactly the ones you ask them. They guess rather than say “I don’t know” because training and grading *reward guessing* [Kalai, Nachum, Vempala, and Zhang 2025].

46

60

Implications of AI on the Society

What this means for you: the entry-level squeeze



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

47

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References



Figure: Stanford “Canaries in the Coal Mine?” (updated August 2026): among people aged 22–25, the number in work has *held up* in the jobs where *AI does little*, but has *fallen behind* in the jobs where *AI does much of the work*. The authors call this a *warning sign*, not proof that AI caused it. (<https://digitaleconomy.stanford.edu/publications/canaries-in-the-coal-mine/>)

Implications of AI on the Society

What this means for you: your first job



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

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

- *The entry-level squeeze*: among people aged 22–25, the number in work has *held up* in the jobs where *AI does little*, but has *fallen behind* in the jobs where *AI does much of the work* (Stanford, updated 2026). A *warning sign*, not proof that AI caused it.
- *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.

48

60

Implications of AI on the Society

What is being done: the EU AI Act



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

49

Responses & Governance

Reflection & Outlook

Take-home Messages

References



High-level summary of the AI Act

27 Feb, 2024

Updated on 30 May in accordance with the Corrigendum version of the AI Act.

In this article we provide you with a high-level summary of the AI Act, selecting the parts which are most likely to be relevant to you regardless of who you are. We provide links to the original document where relevant so that you can always reference the Act text.

To explore the full text of the AI Act yourself, use our [AI Act Explorer](#). Alternatively, if you want to know which parts of the text are most relevant to you, use our [Compliance Checker](#).

[View this content as a PDF](#)

Four-point summary

The AI Act classifies AI according to its risk:

- Unacceptable risk is prohibited (e.g. social scoring systems and manipulative AI).
- Most of the text addresses high-risk AI systems, which are regulated.

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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

50

Responses & Governance

Reflection & Outlook

Take-home Messages

References

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) and arrives *in stages* — since *August 2026* a chatbot must *tell you it is a machine* and AI-made images and video must be *labelled*. 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



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

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 are the AI-related statements in the article?]
 - [Are any conclusions in the article 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.

51

60

Implications of AI on the Society

Navier–Stokes, September 2026 — Part 1: what does “solved” mean?



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

52

Reflection & Outlook

Take-home Messages

References

Exercise 8 (When may we say a problem has been solved?)

On 8 September 2026, *OpenAI* announced that one of its models had proved blow-up for *Navier–Stokes*, one of the seven *Millennium Prize Problems*. Ten thousand AI agents worked on it, and the proof was checked by machine in *Lean*. Read the [official problem statement](#), written for the Clay Institute by *Charles Fefferman* (Princeton, Fields Medal 1978), then decide for yourself:

- (a) Fefferman’s statement offers *four* things one may prove, (A) to (D). The announced result is (C) and (D): blow-up for the equation *with an added smooth force*. *Does that count as solving the problem?* Say exactly which part it answers.
- (b) *Terence Tao* (Fields Medal) agreed with a reader that the result is best described as a “*formally verified proposed solution awaiting expert digestion and community validation*” rather than a solved Millennium Problem; he places it between the “verified proof” and “well-written proof” stages on page 43 of his [ICM 2026 slides](#). *If a machine checked every step, what is left to check?*
- (c) *Martin Bridson*, president of the Clay Institute: “*The process of evaluation is deliberately unhurried, and we shall ensure that it is absolutely rigorous.*” The machine finished in 11 hours; the prize may take years. *Why the difference?*

Implications of AI on the Society

Navier–Stokes, September 2026 — Part 2: the argument about credit



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

Exercise 9 (What do you make of the dispute?)

Hours *before* that announcement, *Tristan Buckmaster* (New York University) and *Levent Alpöge* (Anthropic) had posted their own blow-up results, obtained with heavy help from AI — including *OpenAI's own tool*, which they paid for as ordinary customers. Buckmaster then published *a statement* about what happened next. Read it together with *Tao's blog comments*, then give *your own* view on each:

- (a) Buckmaster: “*This is a Deep Blue-Kasparov moment. The community needs to have serious and unhurried discussion about where to go from here.*” *What do you think?*
- (b) He says the news is not the theorem but “*the significance that a mathematician and an LLM model can now do all this work in a month*”, which changes “*the way we train students, assign credit, referee*”. He also calls one of his own write-ups “*AI slop*”. *What should change in the way you are taught?*
- (c) Buckmaster is careful: “*I have not seen OpenAI's proof . . . I am not accusing anyone of anything.*” OpenAI denied using his work. He had put every draft into their tool as a paying customer. *What rules should protect the work you put into an AI tool?*
- (d) Tao: “*I do not identify either with a simple ‘pro-AI’ or ‘anti-AI’ position*”, yet he also writes that the major AI companies are “*doing more harm than good to mathematics*”. *Can both be true at once?*

53

60

Implications of AI on the Society



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

Exercise 10 (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

54

60

Implications of AI on the Society

Beyond the Terminator: the real AI-safety question



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

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?*”

55

60

Implications of AI on the Society

Case study: people who build AI and then walk away, September 2026



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

Weighing AI's Impact

The Risks

The Other Side: Benefits

What It Means for You

Responses & Governance

Reflection & Outlook

Take-home Messages

References

Exercise 11 (Should we slow down?)

On 8 September 2026, *Jacob Coxon* resigned after about three years of *pre-training research* — building the models themselves — first at *OpenAI*, then at *Anthropic*. He left because he believes the systems are getting stronger quickly, *nobody has them under control*, and both companies are racing on regardless (*TIME*, 9 September 2026). He is not alone: *Terence Tao*, who argues that AI should be brought into science *carefully rather than raced*, writes that many who shared that view inside the industry “*have left or become sidelined*”. Read the article, find at least one serious source that *disagrees*, then answer:

- (a) Coxon: “*Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives.*” *What do you think* — is this a warning worth listening to, or one person’s fear?
- (b) *Evan Hubinger*, who leads alignment stress-testing at *Anthropic*, agreed in public: “*Jacob is correct here — we really do earnestly believe AI could kill all humans!*”, putting the chance at “*>10% within the next decade*” — yet he *stayed*. *Does it make sense to believe that and keep the job?*
- (c) *Terence Tao*: “*I am certainly supportive of efforts to slow down the development and release of even more capable models, and to impose genuine regulation on the industry.*” *What do you think* — would slowing down work, and who would have to agree to it?

56

60

Take-home Messages

What AI is, and where it stands



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

57 Take-home Messages

References

- There is no single agreed-upon definition of AI.
- 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.
- *Where things stand right now*: in *September 2026* an AI produced a machine-checked proof of a *Millennium Prize Problem* — and mathematicians still call it a *proposed solution awaiting community validation*, while arguing over *who gets the credit*. Fast progress and *unsettled standards* arrive together.

Take-home Messages

Its impact — and what to do about it



Introduction

Hoàng Anh Đức

What is Artificial Intelligence (AI)?

Related Fields

History of AI

Implications of AI on the Society

58 Take-home Messages

References

- AI is *just a tool*. The *risks* are real (bias, privacy, deepfakes, security, *fabricated citations*, jobs), but so are the *benefits* (science, medicine, 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, regulation (the *EU AI Act*) — and the deeper question is *alignment and control*.
- *What this means for you*: AI is already in your studies — use it *responsibly* (think first, then disclose and cite). *AI literacy is now expected*, while *judgement, communication, and ethics* stay valuable.
- *How to judge any AI claim*: *measure, do not assume*. A wrong answer sounds exactly as confident as a right one, and in a 2025 trial developers were *19% slower* with AI while believing they were *20% faster*. Check what it tells you.

References



Introduction

Hoàng Anh Đức

What is Artificial
Intelligence (AI)?

Related Fields

History of AI

Implications of AI on
the Society

Take-home Messages

59 References



Michael, Noam, Daniel BenShushan, Jacob Bien, and Don A. Moore (2026). *Confidence Calibration in Large Language Models*. Preprint. arXiv: 2605.23909.



Becker, Joel, Nate Rush, Elizabeth Barnes, and David Rein (2025). *Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity*. Preprint. arXiv: 2507.09089.



Kalai, Adam Tauman, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang (2025). *Why Language Models Hallucinate*. Preprint. arXiv: 2509.04664.



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



Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2016). *Inherent Trade-Offs in the Fair Determination of Risk Scores*. ITCS 2017. arXiv: 1609.05807.

60

References (cont.)



Introduction

Hoàng Anh Đức

What is Artificial
Intelligence (AI)?

Related Fields

History of AI

Implications of AI on
the Society

Take-home Messages

60

References



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.