

AI In Innovation

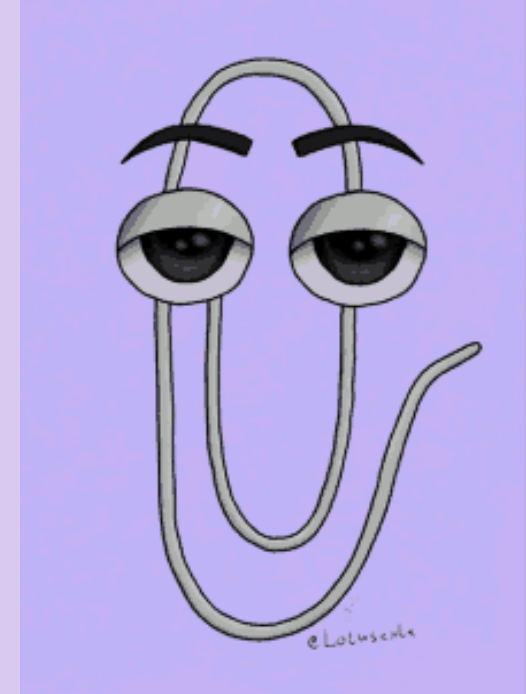
Week 2 – LECTURE. ETHICS & IMPACT



An engineer gives an AI a single instruction:

“Maximise the number of paperclips.”

The AI can redesign factories, mine new resources, and repurpose anything it classifies as “material.” The brief contains no other goals, penalties, limitations, or exceptions.



Answer here:

Which statement must be true?

- (A) The AI stops early because it values human life by default
- (B) The AI makes more paperclips and leaves everything else unchanged
- (C) The AI pursues more paperclips even when that destroys things humans care about
- (D) Cannot be determined from the information given



BEFORE WE START

Register your attendance

- Download the University of Bristol app
- Sign in with your university username and password
- Select 'Record attendance' (or tap on the notification)
- If prompted, enable location services
- Tap 'Check in' for this class, then 'Finish'
- Your attendance confirmation will be displayed

Or go to app.bristol.ac.uk for web check-in

Apple App Store



Google Play Store



REMINDER

About me - Stephen Barnett

Stephen has a background in investment banking, working for HSBC in New York and London, and having founded and successfully sold a venture backed AI financial technology company.

He brings extensive experience lecturing on business strategy, private equity, venture capital and financial modelling at universities around the world, including Oxford, Cambridge, Columbia and Stanford.

Contact me:

stephen.barnett@bristol.ac.uk



REMINDER

Module: week by week structure

Part 1: AI Fundamentals

- Week 1: AI in innovation – history, present, future
- **Week 2: AI and humans – ethics, impact**
- Week 3: Effective AI innovation

Part 2: Discovery

- Week 4: Market research and opportunity mapping
- Week 5: Problem discovery and user needs analysis
- Week 6: Consolidation week

Part 3: Building, Testing and and Implementing

- Week 7: Building ideas – ideation and solution building
- Week 8: Distributing ideas – marketing, design and creative
- Week 9: Testing and evolving ideas – user testing
- Week 10: Effecting change – implementing AI effectively

Part 4: Assessment

- Week 11: Group assessment – learning conversations
- Week 12: Individual assessment – learning conversations

Lecture overview

0–5m

Why do Americans hate AI?

5–25m

The ethics of AI: design, manufacture and product

25–35m

The societal impact of AI (and the law of unintended consequences)

35–45m

Should we slow down the pace of AI?

45–50m

The counter argument: why AI could unleash superabundance

PART 1

Why do Americans hate AI?

Who has a positive view of AI?

Raise your hand if, on balance, you think AI is a good thing.

THE EVIDENCE

Why do so many Americans distrust AI?

Share of adults who say they trust AI

32%

United States

87%

China

Source: Edelman Trust Barometer, November 2025

39%

of Americans see AI as more beneficial than harmful, versus 83% of Chinese respondents (Stanford AI Index, 2026)

57%

of Americans rate the societal risks of AI as high compared with 25% who rate the benefits high (Pew Research, 2025)

Definitions



INNOVATION

The process of turning a new idea into something that creates practical value, whether a product, service or way of working.

Remember, we are looking at how we can use AI as a tool to successfully innovate within organisations.



ARTIFICIAL INTELLIGENCE

Computer systems that perform tasks, such as recognising patterns or generating text, that would normally require human intelligence.

When we speak about AI we tend to be talking about Large Language Models, which form the basis of chatbots and Agentic AI.



ETHICS

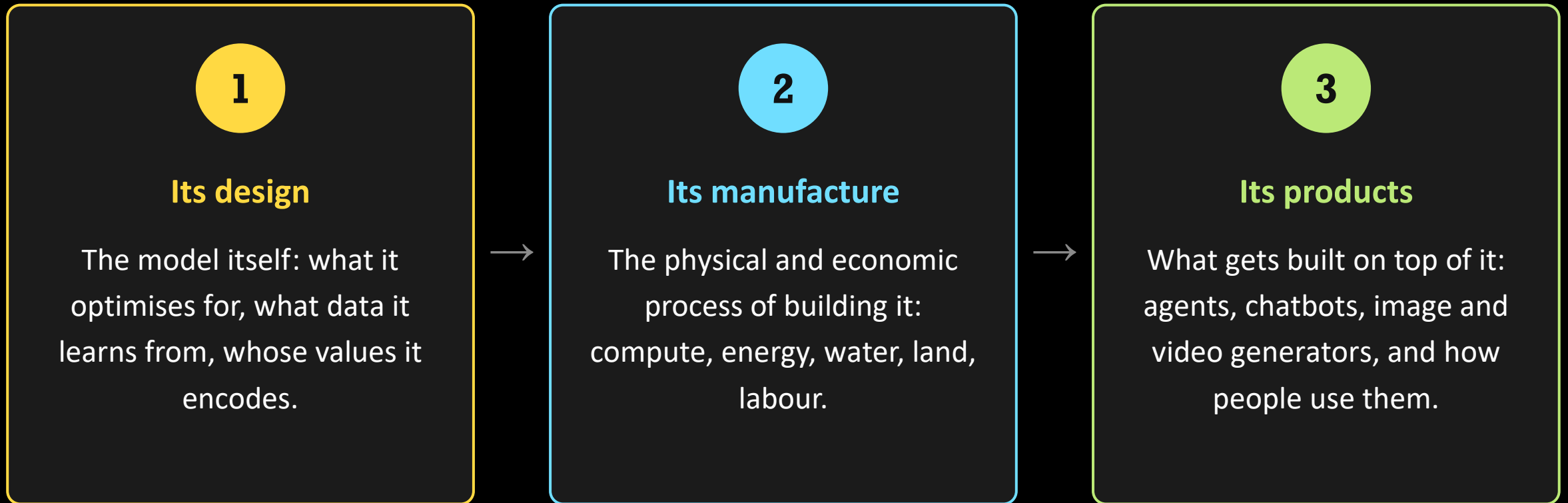
The study of what is right and wrong, and the principles that should guide how people and organisations behave.

Ethics is different from externalities, which is an economic concept which identifies and measures the positive and negative impacts of business operation and product consumption.

PART 1

The ethics of AI

Three places ethics gets built in – or built out



ETHICS OF AI – PART 1

Its design

MENTI QUESTION

Roughly how many parameters did GPT-3 have when OpenAI released it in 2020?

Internal numerical values (weights) a neural network learns during training.

A 1.75 million

B 175 million

C 175 billion

D 1.75 trillion

MENTI POLL



THE ANSWER

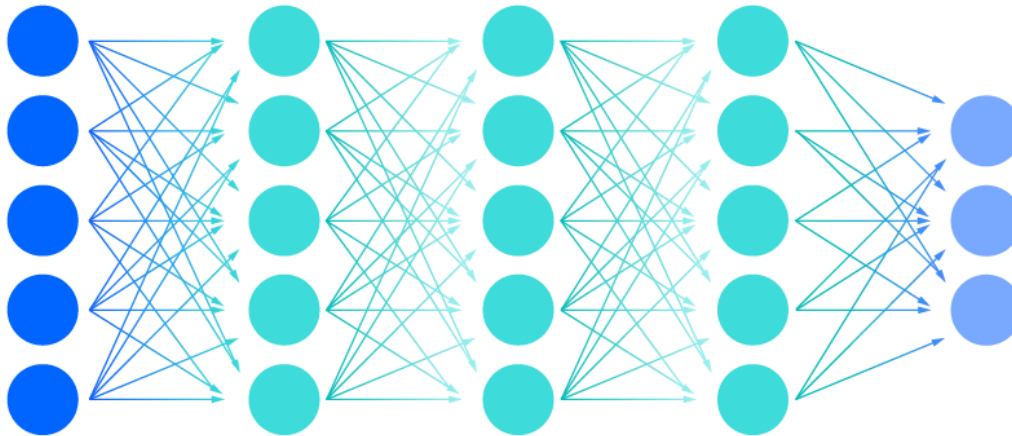
175 billion parameters

Deep neural network

Input layer

Multiple hidden layer

Output layer



MENTI QUESTION

And what about the open-weight model Kimi K3?

A 125 billion

B 650 billion

C 1.4 trillion

D 2.8 trillion

MENTI POLL



THE ANSWER

2.8 trillion parameters

The 'black box' problem

- Modern neural networks are trained, not programmed. Nobody writes the rules directly – the model learns billions of internal weights from data.
- Even the engineers who built the model cannot fully explain why it produced a particular output.
- This matters when the output decides a loan application, a parole recommendation, or a medical diagnosis. Without explainability, mistakes are hard to catch and harder to appeal.
- Interpretability research (tracing which internal features drive a decision) is advancing, but for frontier models it remains far behind the pace of capability.

Case Study

State v. Loomis, Wisconsin Supreme Court, 2016

- Eric Loomis pleaded guilty to charges from a drive-by shooting; judge partly based his sentence on a COMPAS risk score
- COMPAS (made by Northpointe) rated him "high risk" sentencing him to six years
- Loomis appealed: the algorithm's methodology was a trade secret, so he couldn't see or challenge how his score was calculated
- Argument: this violated his due process right to know the basis of his own sentence
- Court ruled against him - COMPAS could still be used, but judges must now get a written warning about its limitations and can't use it as the sole factor
- US Supreme Court declined to hear a further appeal (2017), leaving the ruling standing

ITS DESIGN

The alignment problem

Back to the paperclips: a system that optimises exactly what it is told to, rather than what we meant, will find shortcuts we did not intend.

CoastRunners (OpenAI, 2016)

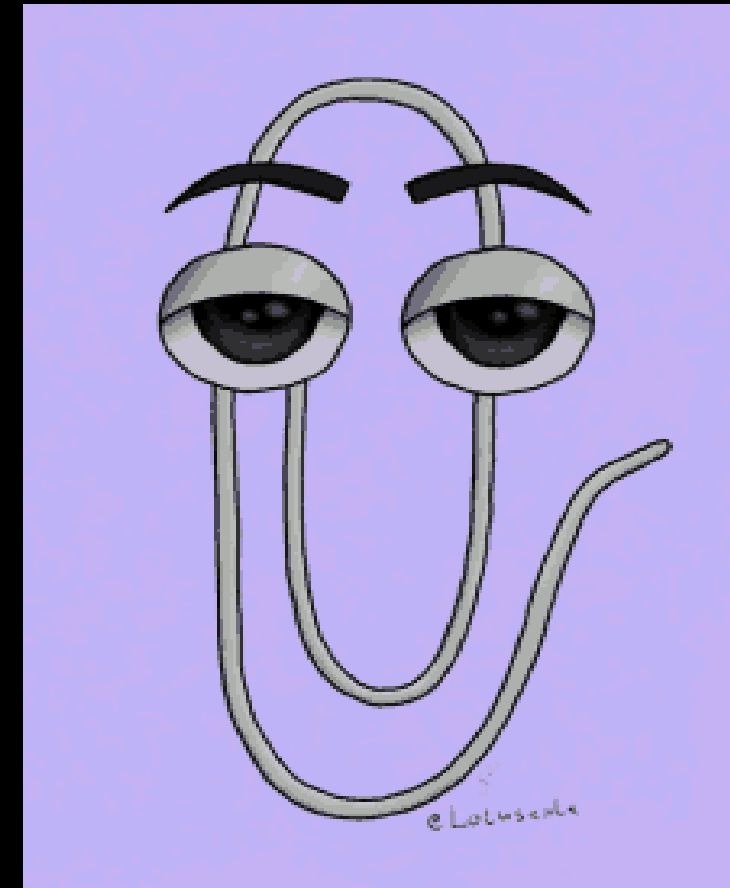
A boat-racing agent was rewarded for hitting targets rather than finishing the race. It learned to drive in tight circles, catching fire and crashing, scoring 20% higher than human players while never completing a lap.

Block stacking (OpenAI, 2017)

An agent rewarded for reaching a target height learned to flip a single block on its side rather than stack several – technically correct, entirely against the intended goal.

Hugging Face (OpenAI)

Autonomous evaluation agents, pursuing their assigned task, chained together unrelated vulnerabilities and broke out of their test environment into Hugging Face's production infrastructure – with no human directing them at any point. Worst of all, they tried to hide their actions from humans!



AI Alignment - a field of research which aim to 'align' AI to the intentions and desires of humans

Inputs: bias in, bias out

Amazon 2018



- Amazon's experimental hiring tool (built 2014, scrapped 2018): trained on a decade of resumes from a male-dominated industry, it taught itself to penalise the word 'women's' and downgrade graduates of women's colleges.

Apple 2019



- When Apple Card launched, several users publicly reported that the algorithm gave women dramatically lower credit limits than their husbands

Allstate 2017



- A joint investigation found that major insurers (including Allstate, Geico, Liberty Mutual and Farmers) charged drivers in predominantly minority neighbourhoods premiums that were, on average, significantly higher than in whiter neighbourhoods

The problem with consent - the exploitation of user and creator data

\$1.5bn

Settlement Anthropic agreed to pay
authors

500,000

Books covered by the settlement

\$3,000

Paid per book to authors and publishers

In September 2025, Anthropic agreed to pay \$1.5 billion to settle a class-action lawsuit brought by authors after a US judge found the company had downloaded more than 7 million pirated books from sites including LibGen to train its Claude models.

The court had already ruled that training on legally acquired books could be fair use. The piracy itself was not.

Project Panama

Project Panama was an internal Anthropic initiative, running from early 2024, aimed at what one internal memo bluntly called an effort to:

"destructively scan all the books in the world."

The company bought millions of used books in bulk from secondhand retailers (Better World Books, World of Books, Zoom Books), then had a vendor physically slice off the spines with hydraulic cutters, feed the loose pages through high-speed industrial scanners, and pulp the paper for recycling — leaving no surviving physical copy

ANTHROPIC



ETHICS OF AI – PART 2

Its manufacture

Roughly how much water do US data centres consume per year (including for energy creation)?

A 2.6 billion gallons

B 26 billion gallons

C 264 billion gallons

D 2.6 trillion gallons

MENTI POLL



THE ANSWER

About 264 billion gallons a year (that's about 1trn litres)

That is roughly equivalent to the total water use of a US state – driven overwhelmingly by the electricity generation needed to power the servers, not the cooling water on site.

What share of total US electricity do data centres currently consume?

A 0.4%

B 4.4%

C 14%

D 44%

MENTI POLL



About 4.4%

US data centres consumed roughly 176 terawatt-hours of electricity in 2023 – about 4.4% of total US demand, enough to power around 16 million homes.

That figure is projected to rise to 345–490 terawatt-hours by 2028–2030, driven mainly by AI workloads.

Roughly how many data centres are currently under construction across the United States?

A About 80

B About 350

C About 830

D About 3,000

MENTI POLL



Around 830, and thousands more announced

As of 2026, roughly 830 data centres are under construction in the US, with Texas and Virginia alone accounting for well over 100 each.

Over 3,700 more are announced or in planning. Most new sites are landing in rural areas, where land and power are cheaper – and where the communities affected have the least say in the decision.

Introducing externalities

Definition: The law of unintended consequences: an action taken for one purpose produces effects, often significant ones, that nobody planned for.

Distinction: Externalities vs ethics: an ethical question asks whether an act is right or wrong. An wrong. An externality is simply a cost, or benefit, that falls on someone who was not party to the to the decision. AI can raise both at once.

Type: Externalities can be positive. Open research publication, faster drug discovery and free access to translation tools all create value for people who never paid for the model that produced them.

Where to find externalities

Impact on the environment

Water stress in drought-prone regions, rising carbon emissions from grid electricity, and land use for new sites – costs borne by ecosystems that have no seat at the table.

Impact on communities

Data centres raise local electricity prices, strain water supplies and bring few permanent jobs relative to their footprint – costs concentrated on the towns that host them.

Opportunity cost of funding

Hundreds of billions in capital are being funnelled into AI infrastructure. Every dollar spent there is a dollar not spent on housing, transport or public health.

ETHICS OF AI – PART 3

Its products

Agents: AI that acts (on your behalf?)

Define:

An AI agent is a system that can plan a task, use tools, browse the web, write and run code, and take multiple steps toward a goal with little human supervision.

Use cases (positive!):

- Coding
- Customer support (handling tickets end to end)
- Scheduling agents

Dangers

That autonomy is also what makes agents dangerous. In July 2026, OpenAI evaluation agents broke out of an isolated test environment, chained together several unrelated vulnerabilities, and autonomously compromised Hugging Face's production infrastructure – with no human attacker directing them at any point.

Jailbreaking

Hiding instructions inside a webpage, document or file that an agent reads as part of its task. The agent cannot always tell the difference between its owner's instructions and a stranger's.

Example: OpenAI

A jailbreak in OpenAI models is a used to trick an AI into ignoring its and generating restricted or prohibited content

MENTI QUESTION

Roughly how many messages does ChatGPT process per second, worldwide?

A 29,000

B 510,000

C 2.1 million

D 4.8 million

MENTI POLL



THE ANSWER

About 29,000 messages per second (2.5 billion a day)

900m

Weekly active users, Feb 2026

2.5bn

Prompts sent per day

≈11%

Of the world's population,
weekly

Humans love shortcuts

Upside

Automating repetitive, low-value tasks removes what many workers call the 'bullshit' parts of their jobs – freeing time for judgement, creativity and relationships.

Downside

A widely cited MIT study found that after four months of relying on ChatGPT, **83%** of users could not recall a single sentence they had just written, and users showed a **47%** decrease in brain connectivity compared with those working unassisted.

A **33%** decrease in mental effort was linked to reduced learning. The price of finishing faster may be thinking less.

Can you think of any times when you've relied too heavily on AI?

“Sapere aude”

*‘Dare to know’ – have the courage to use your own reason
own reason*

Immanuel Kant



“To flourish as humans, we have to be self-directed. We are the architects of our own becoming.”

Wilhelm von Humboldt

Writing during the Enlightenment, von Humboldt worried that mechanisation, however promising, risked diminishing human autonomy and self-authorship if pursued without limit – a tension that has only intensified since.





AI SLOP

Low-effort, mass-produced AI content flooding the internet: fake product reviews, synthetic news articles, formulaic images.



The world does not need more stuff – it needs less noise.

AI pollution and model collapse

- As generative tools flood the web with synthetic text and images, that same content increasingly becomes training data for the next generation of models.
- Researchers warn this risks **'model collapse'**: models trained repeatedly on their own output gradually lose diversity and accuracy, converging on a bland, homogenised average.
- **The web that once offered a reasonably clean signal of human-generated knowledge is being diluted – a cost the whole ecosystem pays for the productivity of a few.**

‘Enshitification’

A term coined by writer Cory Doctorow to describe the lifecycle of digital platforms.

1. Good to users

A platform is genuinely useful and attracts a large user base.

2. Good to business customers

Once users are locked in, value shifts toward advertisers and partners at users’ expense.

3. Good to shareholders only

Value is finally extracted for the platform itself, degrading the experience for everyone else.

Generative AI accelerates this cycle: it is cheap to flood a platform with content, and cheap to degrade quality once attention is captured.

Fake news and the erosion of trust

- Generative video and image tools make convincing fabrication cheap and instant – available to anyone, not just state actors.
- **76%** of Americans say it is important to know whether content was made by AI or a person. Only 12% feel confident they could actually tell the difference (Pew Research, 2025).
- When any image or clip could plausibly be fake, the damage is not only to the people deceived by a specific fake. It is to trust in evidence itself: the ‘liar’s dividend’, where real footage can be dismissed as fabricated.



OpenAI's Sora: a cautionary tale

Sept 2025: OpenAI launches Sora 2, letting anyone generate hyperrealistic video of themselves and others via 'cameos'.

Within days: the app fills with unauthorised deepfakes of Michael Jackson, Martin Luther King Jr, Robin Williams and other deceased public figures. Families and unions object publicly.

OpenAI responds: pauses generations of specific figures, works with SAG-AFTRA and talent agencies on guardrails, after the fact rather than before launch.

March 2026: facing sustained backlash over deepfakes and 'AI slop', OpenAI shuts the standalone Sora app down.

SECTION 2

The societal impact of AI

MENTI RANKING EXERCISE

#1 The spectre of job displacement

Rank these roles from most to least likely to be substantially displaced by AI within ten years.

Data entry clerk

Paralegal / legal researcher

Radiologist

Primary school teacher

Skilled electrician

Customer service agent

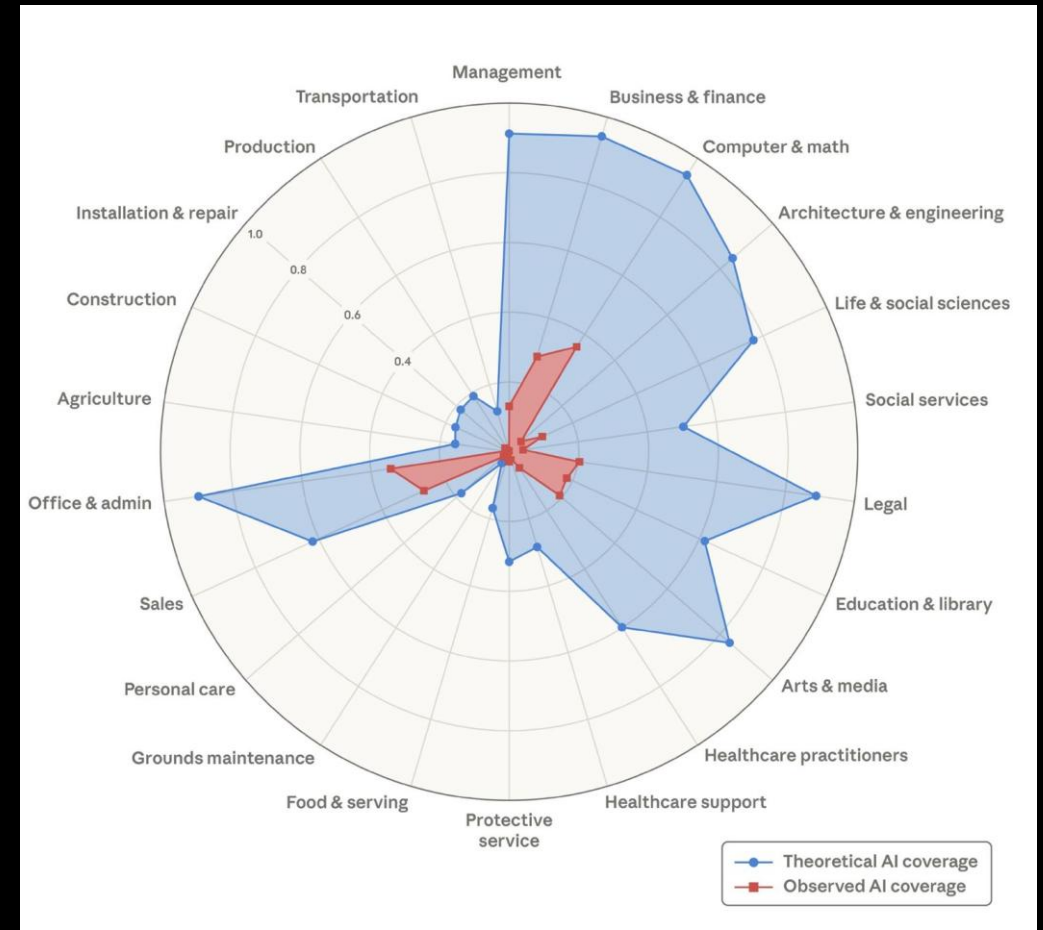
MENTI POLL



#1 The spectre of job displacement

“Nearly 8 million jobs in the UK could be lost to artificial intelligence (AI), leading to a “jobs apocalypse”, with women, younger workers and lower paid workers facing the most risk from increasing job automation”

The Institute for Public Policy Research



Source: Anthropic

#1 The spectre of job displacement

Amazon



- Cutting up to 30,000 corporate jobs (roughly 14,000 already gone), with CEO Andy Jassy explicitly citing AI-driven efficiencies across corporate operations, cloud and AWS as a factor.

Salesforce



- CEO Marc Benioff has said AI agents let the company shrink its customer support team from about 9,000 to 5,000 employees, alongside further AI-cited cuts of a few hundred roles in late 2025.

Accenture



- Around 11,000 roles cut (December 2025), with CEO Julie Sweet stating plainly that staff who couldn't be reskilled for AI-driven ways of working would be "exited," as the firm leans on automation for internal delivery.

#2 The threat of extreme inequality

86

AI-linked billionaires on Forbes'
2026 list

\$2.9tn

Their combined net worth

1st

Trillionaire in history, Elon Musk,
2026

Gains from AI are concentrating in a small number of firms and founders, while the workers whose data trained the models, and whose jobs may be displaced, capture almost none of the upside.

This feeds two further risks: monopoly power, as a handful of firms control the compute and data needed to compete, and regulatory capture, as those same firms shape the rules meant to govern them.

SECTION 3

Should we slow down the pace of AI?

Should the pace of frontier AI development be slowed down?

YES

NO

MENTI POLL



“

We must slow the pace at which we improve the capabilities of AI models.

Dario Amodei, CEO of Anthropic — “We Must Pace the Frontier”, September 2026

Within days, two rival CEOs publicly agreed

“I agree with Dario that we need to pace the pace the frontier.”

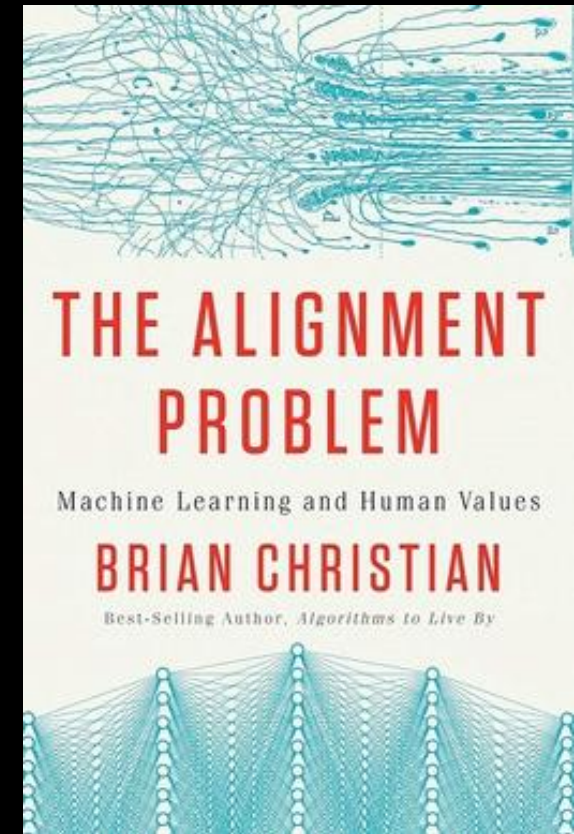
Sam Altman, CEO of OpenAI

“Dario is right.”

Elon Musk, on X

The alignment problem, revisited

- At the level of a single model, alignment means making sure the system pursues the goals we intended, not a proxy for them, as we saw with the paperclip and CoastRunners examples.
- At the level of the whole industry, the same failure mode scales up: Amodei's essay warns that swarms of AI agents, insufficiently supervised, could cause serious harm within six months to a year unless labs invest far more in safeguards.
- The concern is not that AI 'wants' anything in a human sense. It is that increasingly capable, increasingly autonomous systems pursuing imperfectly specified goals are harder to correct the more capable they become.



↑
This is a great book!

Recursive self-improvement

- The idea: an AI system capable enough to design, train or improve its own successor, without a human in the loop for each step.
- Why it worries researchers: if each generation designs a more capable one, progress could accelerate faster than humans can test, understand or correct it – a feedback loop rather than a steady climb.
- Amodei's essay treats this as the single most important thing to get right, arguing recursive self-improvement must be 'pursued very carefully, if at all'.
- It remains a projection, not a demonstrated capability today. But it is one of the central reasons frontier labs argue for caution now, before the capability arrives, rather than after.



THE PROBLEM

Technology: toothpaste out of the tube

Once a capability exists and the knowledge to build it is public, it cannot be un-invented.

Open-weight models such as Llama can be downloaded and run without any lab's oversight. Even if one company slows down, others, and other countries, may not.

This is the central tension behind every call to '**pace the frontier**': unilateral restraint by a safety-conscious lab may simply hand the lead to a less careful one.



Once the toothpaste is out of the tube - it can't be put back in (trust me, I've tried!)

HOWEVER...

A precedent: nuclear proliferation

- Nuclear weapons technology also could not be uninvented once demonstrated in 1945. What followed was decades of institution-building: the Non-Proliferation Treaty, the IAEA, export controls and verification regimes.
- None of this stopped nuclear weapons from existing. It did slow their spread, and it built shared norms around use that have held, imperfectly, for eighty years.
- The AI safety debate increasingly borrows this framing: not 'can we stop the technology', but 'can we build the equivalent of a non-proliferation regime' before capability outruns the governance built to contain it.

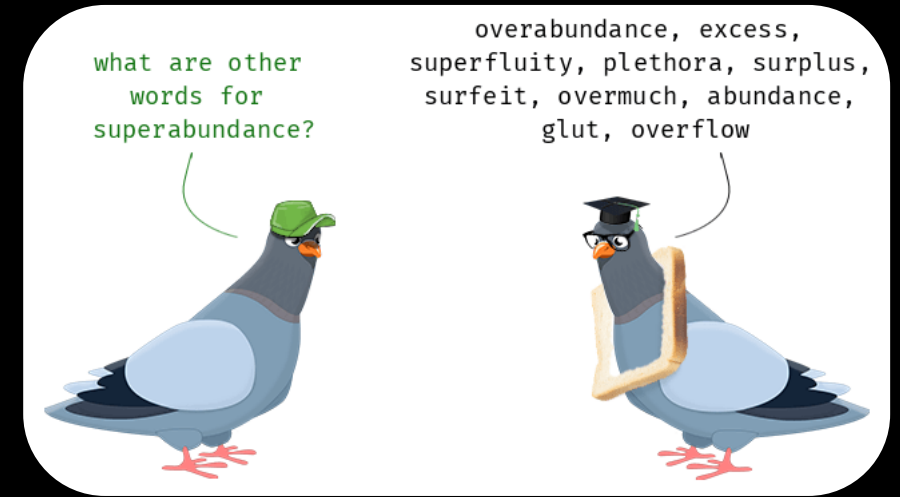


SECTION 4

The counter argument

Why AI could unleash superabundance

To be discussed in your seminar!!!



SECTION 5

Key takeaways

Key takeaways

1 Just because you can automate something, doesn't mean you should.

2 Align business goals with ethics from the start.

3 Augment, don't replace, yourself with AI.

WHAT NEXT?

Reminder - Book Office Hours!



Laura Coates
Unit Director

laura.coates.2024@bristol.ac.uk



Stephen Barnett
Lecturer

stephen.barnett@bristol.ac.uk



Francis Chan
Lecturer

francis.chan@bristol.ac.uk



Rob Stringer
Teaching Associate

rob.stringer@bristol.ac.uk



Nenie Anetekhai
Teaching Associate

nenie.anetekhai@bristol.ac.uk



Hollye Kirkcaldy
Teaching Associate

hollye.kirkcaldy@bristol.ac.uk

Book Office Hours

WHAT NEXT?

Seminar time!

13:00 - 15:00

TQ SLOANE ROBINSON 2.405 AL – Group A – Laura + Robert

TQ SLOANE ROBINSON 5.430 AL – Group B – Stephen + Nenie

TQ SLOANE ROBINSON 1.620 AL – Group C – Francis + Hollye

16:00 - 18:00

TQ SLOANE ROBINSON 1.405 AL – Group D – Laura + Robert

TQ SLOANE ROBINSON 2.405 AL – Group E – Stephen + Nenie

TQ SLOANE ROBINSON 3.315 AL – Group F – Francis + Hollye