Finance and AI

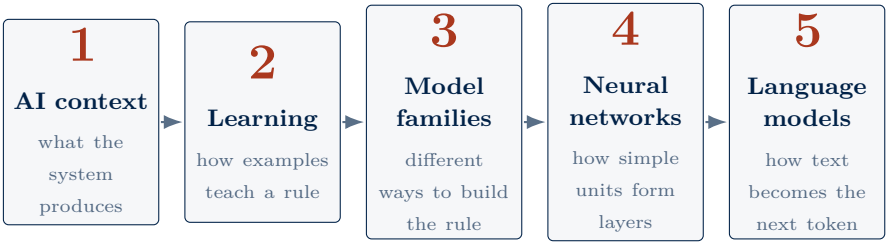
How modern models learn, represent, predict, and generate

Fatih Kansoy

Worcester College, University of Oxford

Day 3

A visual map of today's lecture



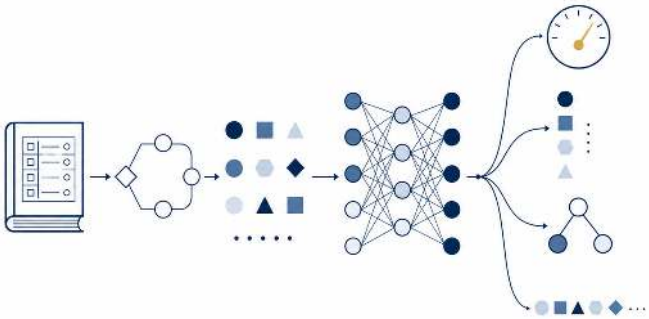
The aim is explanation, not mathematical derivation.

By the end, you should be able to explain one neuron calculation and the path from text to tokens, context, attention, and the next token.

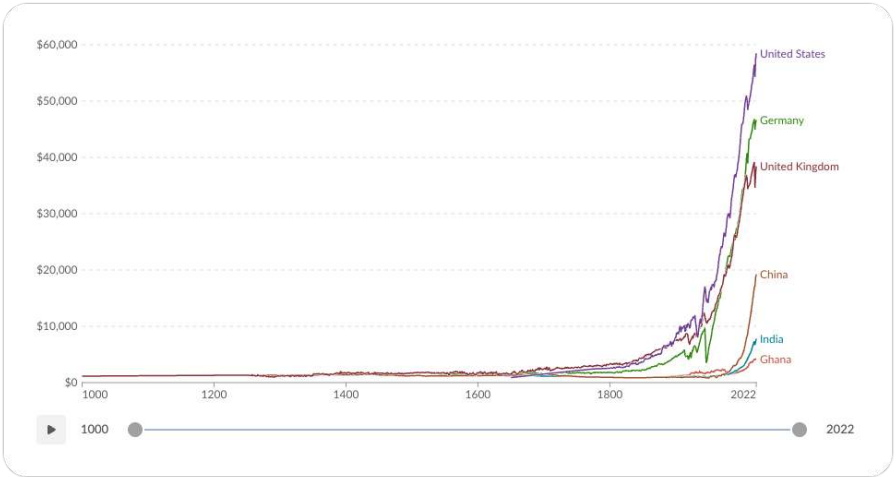
Orientation

Artificial intelligence: history and context

What an AI system produces, how its methods changed, and why this course begins with learning from data.



What happened and what is happening?

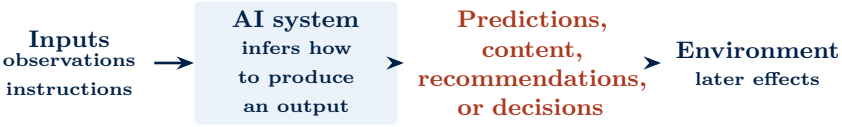


Source: Our World in Data.

AI is defined by what a system produces

For an explicit or implicit objective, an AI system **infers from inputs** how to generate outputs that can affect a physical or virtual environment.

Objective = criterion guiding an output; model = learned component; decision = use of an output.



AI is the broader system category. This course focuses on cases where a model inside the system is learned from data.

Source: OECD, updated definition of an AI system (2024).

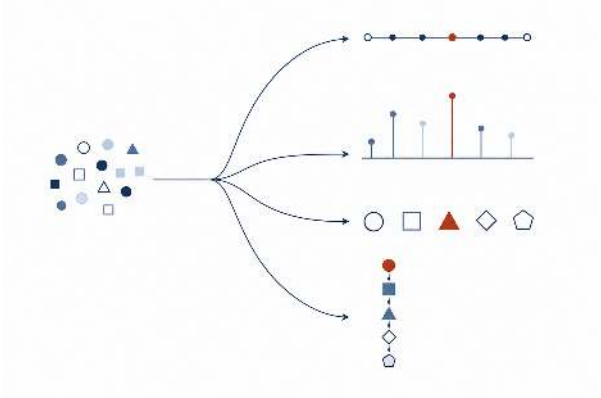
Definition and scope of artificial intelligence

> **Artificial intelligence**: building systems that perform tasks which normally require human intelligence, such as perceiving, reasoning, deciding, or producing language.

- **Narrow AI** performs one task envelope well: translation, chess, credit scoring. Every deployed system today is narrow.
- **General AI** would mean competence across arbitrary tasks. No such system exists, and it is not this week's subject.
- The definition says nothing about method. The methods that now deliver these capabilities are **learned from data**.

When a headline says "AI", ask first: which task, and how is the system judged on it?

Name the output before choosing the method



Number
ETA, demand, sale price

Probability or risk score
fraud, churn, default

Class after a rule
spam, identity, defect

Ranking
search, products, films

The required output determines the target, loss, and evaluation evidence.

Target = outcome or structure to learn; loss = how errors count; class = label after a decision rule.

Generative systems produce content from learned probability distributions; language models often generate a token sequence one choice at a time.

The terms answer different questions

SYSTEM CATEGORY

AI

Systems that infer outputs for objectives.

HOW A MODEL IS LEARNED

Machine learning

estimates models from data.

Deep learning

is ML using multilayer neural networks.

WHAT IS PRODUCED

Generative AI

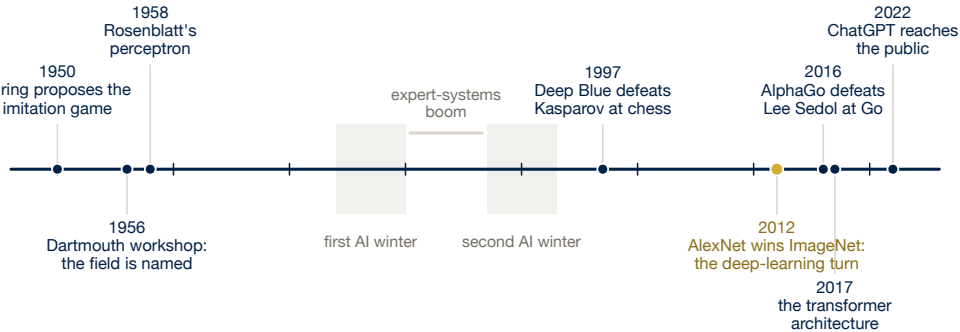
Systems intended to produce content: text, images, audio, code, and more.

Modern generative AI usually uses deep learning. But not every AI system learns from data, not every ML model is deep, and not every learned model generates content.

Source: OECD AI-system definition (2024); standard ML terminology.

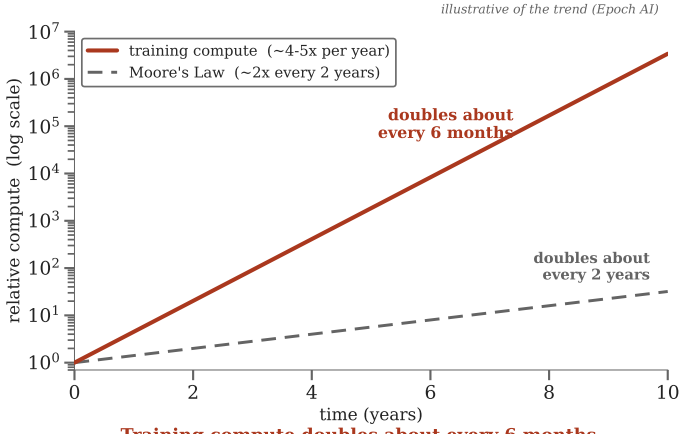
AI advanced in waves; methods accumulated

- In 1950 Alan Turing replaced "can machines think?" with a test of written answers; in 1956 a Dartmouth summer workshop named the field **artificial intelligence**.
- Twice, promises outran results and funding collapsed: the **AI winters**.
- The learning era after 2012 is the part this course examines.



Source: Turing (1950); Dartmouth proposal (1955/56); Computer History Museum (perceptron, expert systems, AI winters); IBM Research (Deep Blue); Krizhevsky et al. (2012); Silver et al., *Nature* (2016); Vaswani et al. (2017); OpenAI (2022).

Recent progress came from three reinforcing forces



Recorded examples
Larger datasets, including curated and labelled benchmarks.

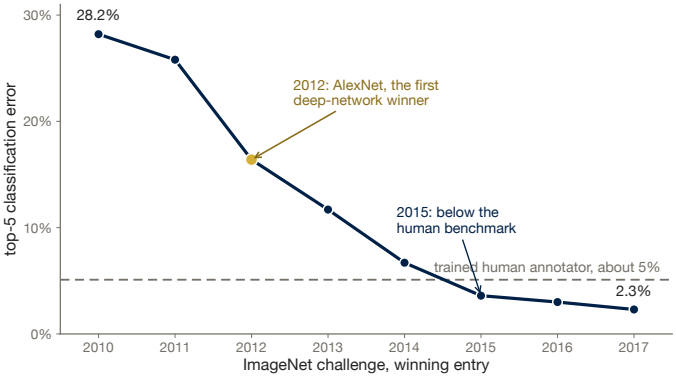
Compute
More experiments and larger training runs.

Learning methods
Architecture = permitted model structure; optimisation = procedure that adjusts it.

Epoch estimates that training compute for *selected notable models* rose about 4.5× per year from 2010. This is a historical sample trend, not a technological law; more compute does not guarantee usefulness. Scale still requires a defined test.

Source: Epoch AI, Trends in Artificial Intelligence (notable-model sample; accessed 2026).

A breakthrough is meaningful only inside a defined test



ImageNet made one claim comparable by fixing:

1. a labelled image task;
2. a benchmark: shared task, data, metric, and held-out cases;
3. a top-5 error metric: an error when the correct label is absent from the five highest-scored labels;
4. images withheld from fitting.

AlexNet's 2012 result was a large advance on this **defined classification problem**—not evidence of general intelligence.

Source: Russakovsky et al. (2015); official ILSVRC results archive (2016–17); AlexNet in Krizhevsky, Sutskever, and Hinton (2012). The plotted benchmark series and paper-specific test comparisons should not be interchanged.

We begin where outcomes are attached to historical cases

Setting	What is observed	What is learned	Course
Supervised	inputs x and outcomes y	an input-to-outcome rule	Days 1–3
Unsupervised	inputs x only	structure: groups or directions	Day 4
Reinforcement	states or observations, actions, and resulting rewards	a policy for acting	outside scope

x = recorded inputs; y = later outcome; policy = rule for acting; reward = consequence used to judge an action.

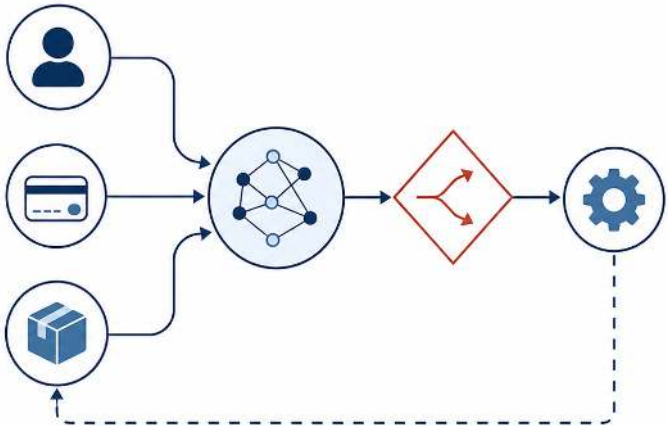
Self-supervision constructs training targets from the data themselves—for example, predicting a hidden or next token. It is a training strategy, not a separate business objective.

Next: how model outputs enter business decisions, and what evidence can establish value.

Orientation

Business applications of AI

A model becomes useful only when its output changes a decision and the resulting value—and harm—can be measured.



Machine learning applications by business function

Marketing
who will leave,
and who to target

Finance
fraud, and
credit risk

Operations
demand forecasts,
and maintenance

Customer
recommendations,
support triage

HR
screening,
handle with care

Risk
early warning
of trouble

This is not one department's tool. It is everyone's.

Netflix recommendation systems

Netflix's own engineers report that recommendations influence about **80%** of the hours watched, and estimate that personalisation saves **more than $1bn a year**, mostly through customers who stay.

Note who is doing the reporting. These are the company's own figures, and no outside party has audited them.

Source: Gomez-Uribe & Hunt, ACM TMIS 2016; Netflix's own ∼2015 estimate.

∼80%

of hours streamed are shaped by what the model recommends

A good recommender is not a feature. It is the business.

Real-time payment fraud detection

Global payment-card fraud losses reached about **$33.4bn** in 2024. Every card payment is scored for risk in **real time**, far faster than any human review could manage.

This is machine learning you never see, running on every card payment, everywhere, all day.

Source: The Nilson Report, card-fraud losses worldwide, 2024 (card fraud specifically).

$33.4bn

global card-fraud losses in a single year, 2024

At this speed and scale, the model is the only thing fast enough.

Email spam, phishing, and malware filtering

Google states that Gmail blocks **more than 99.9%** of spam, phishing and malware, filtering on the order of **15 billion** emails a day.

You never notice the ones it stops. That is the point.

Source: Google Safety Center (a vendor self-report, not an independent audit).

>99.9%

of spam, phishing and malware blocked before you see it

The best ML often shows up as nothing going wrong.

Generative AI use and coding productivity

The newest wave writes and codes. ChatGPT reports about **800 million** weekly users. In **one controlled study**, about 95 developers doing *a single* coding task finished **55% faster** with an AI assistant.

A striking result, but one narrow task run by the vendor. The same rules of evidence still apply.

Source: OpenAI (Oct 2025); GitHub Copilot study, 2022 (∼95 developers, one greenfield task, wide confidence interval).

55%

faster on *one* coding task, in a single vendor study

The newest wave, and the same discipline applies.

Estimated economic impact of generative AI

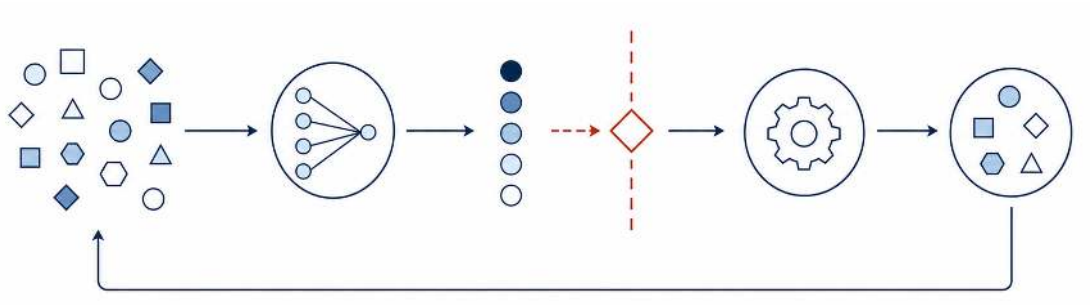
$2.6–4.4 trillion

a year that generative AI *could* add to the world economy, across dozens of use cases: a **projection**, not a measured figure

Source: McKinsey Global Institute, "The economic potential of generative AI," 2023. A modelled estimate, sensitive to adoption; other houses give different numbers.

A real opportunity, and an estimate, not a promise.

A model creates value only through a decision chain

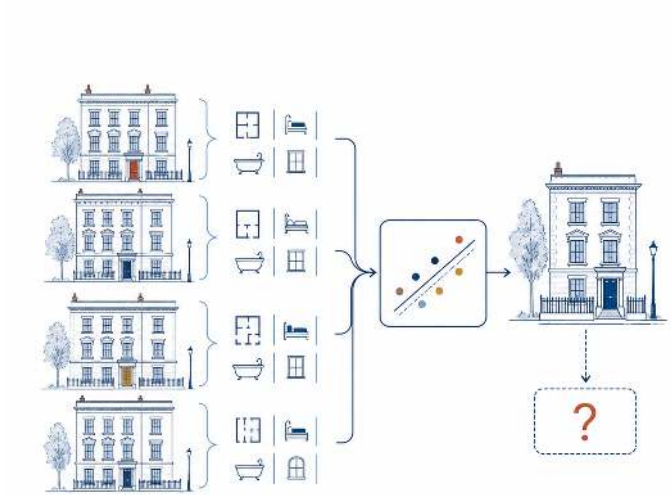


Part 1 now makes this chain testable: what is predicted, from which historical cases, at what moment, and for whose use?

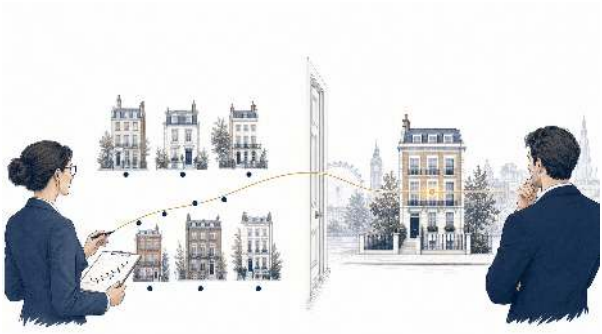
Part 1

A model learns from examples

A useful learning problem says what information is available, what answer is wanted, and how the rule will be checked on new cases.



A model learns a rule from completed examples



Historical examples

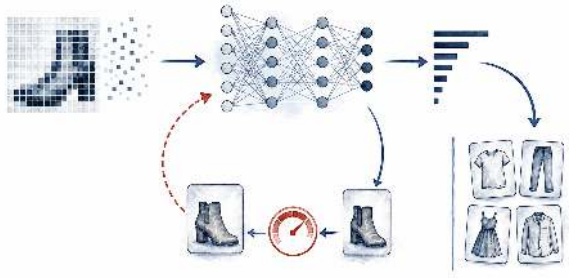
Recorded inputs and known outcomes teach the relationship.

A new case

Only the inputs are available; the learned rule proposes an output.

The target helps to learn and test the rule. It is not available when the new prediction is made.

Training changes the model; prediction uses it



Forward path

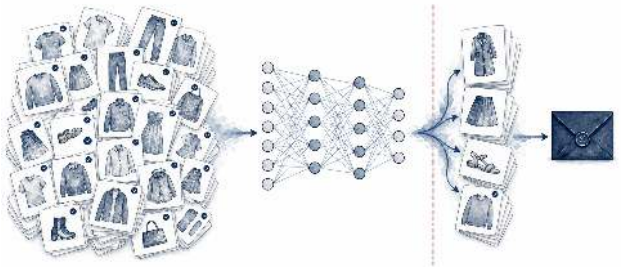
A new input passes through fixed learned weights to produce scores and an output.

Learning path

Known answers expose error; the feedback loop adjusts the weights and repeats.

Using a trained model does not normally retrain it.

The real test is a case the model did not study



Seen while learning

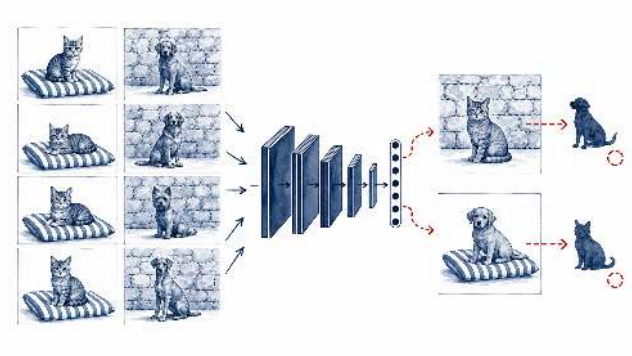
The model can improve by correcting mistakes on these examples.

Held back until evaluation

These cases reveal whether the rule works beyond the material it studied.

Testing is about new evidence, not repeating the practice questions.

A model can learn the wrong pattern for the right answers



During training

Cats happened to appear on cushions; dogs happened to appear by walls.

On a new background

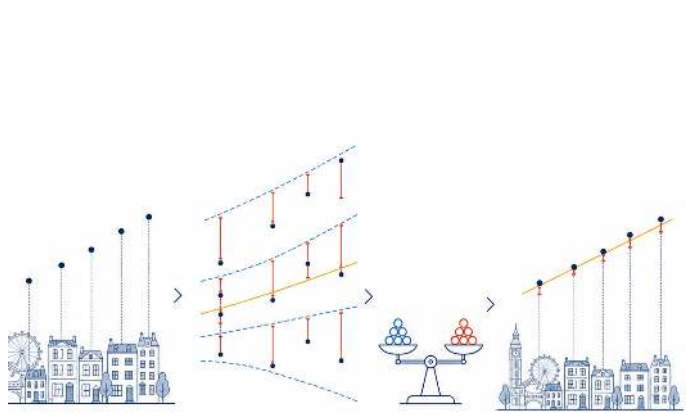
The shortcut fails because the model learned scenery instead of the animal.

Overfitting includes memorising examples and relying on patterns that do not travel to the setting where the model will be used.

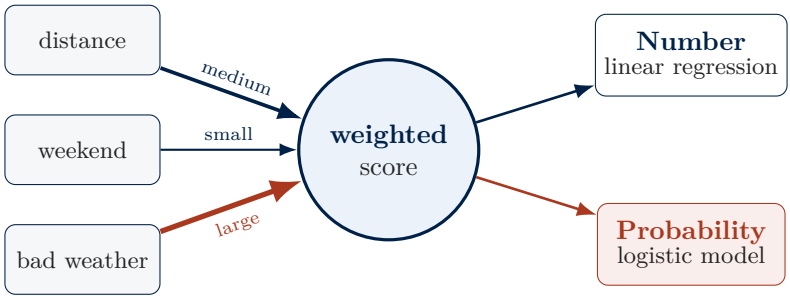
Part 2

Different models build different rules

Some add evidence into one score; some ask questions; some combine many rules; some search for groups or unusual cases.

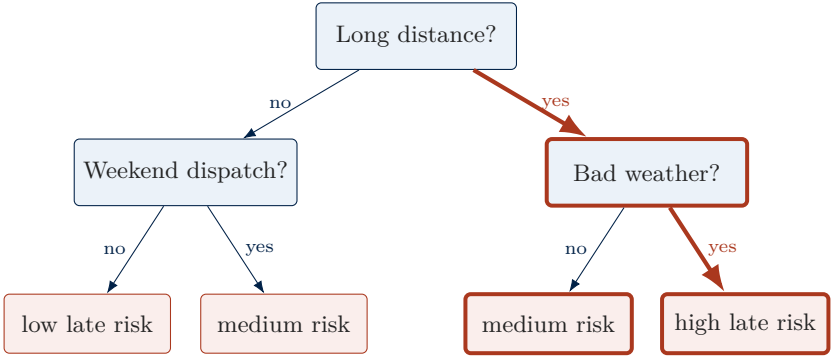


A weighted score is the simplest useful baseline



The same additive recipe is applied to every case. That makes it transparent, fast, and a valuable benchmark.

A decision tree follows one path of questions

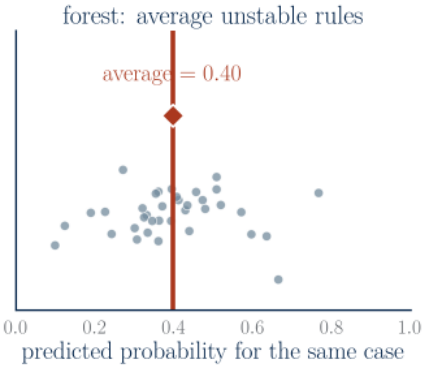


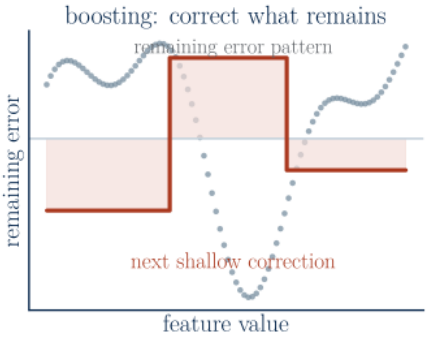
Mechanism
Each answer sends the case to a smaller comparison group.

Output
The final leaf supplies a number, probability, or class.

A tree is easy to trace, but a small change in the data can produce a different tree.

Tree ensembles improve a rule in two different ways





Random forest

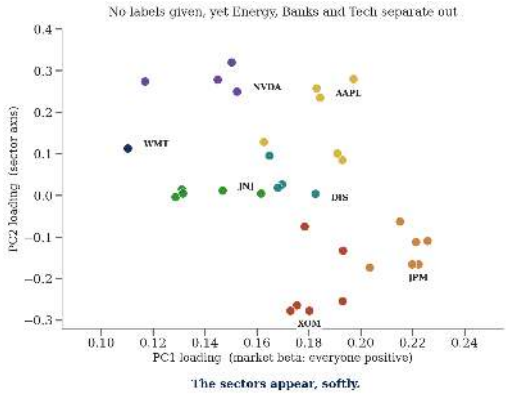
Build varied trees and average or vote, reducing the instability of one tree.

Gradient boosting

Add shallow trees in stages, each correcting part of the error still left.

Combining trees often improves prediction, while making any single case harder to explain.

Groups and unusual cases can emerge without supplied labels



Clustering

Nearby cases form groups under the measurements we chose.

Anomaly detection

A case far from its usual neighbours is flagged for investigation.

A flag is not automatically fraud, error, or danger; it is a reason to look more closely.

Different models are different ways to inspect the same case

Weighted score
Add evidence into a number or probability.

Tree or ensemble
Ask questions and combine many branching rules.

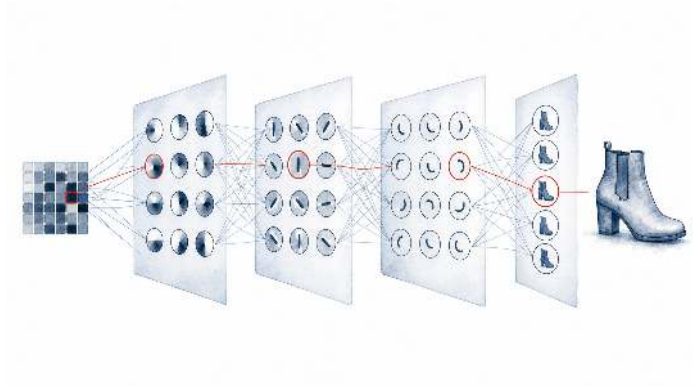
Groups or flags
Compare cases by similarity rather than a supplied target.

Start from the question and the useful output; then let evidence justify extra complexity.

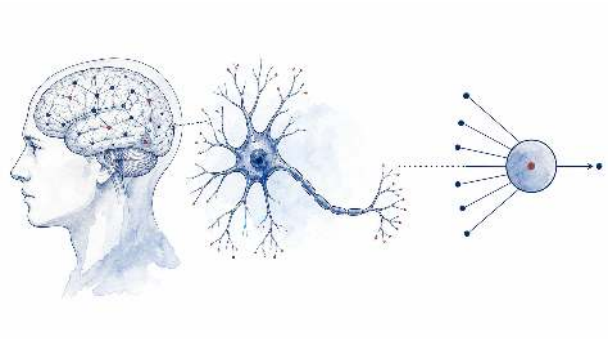
Part 3

Simple units, complex behaviour

A neuron performs a small calculation. Layers repeat that calculation many times and build increasingly useful internal representations.



The brain inspired the name—not the mechanism



Biological neuron
Receives signals from many other cells.

Artificial neuron
Combines numbers and produces another number.

The limit
A real neuron and brain are vastly more complex.

The analogy motivates the picture; the mathematics defines the model.

The model receives a numerical representation of the case

From case to fields

A real property has far more detail than a model can receive. We choose and encode recorded features such as size, distance, and age.

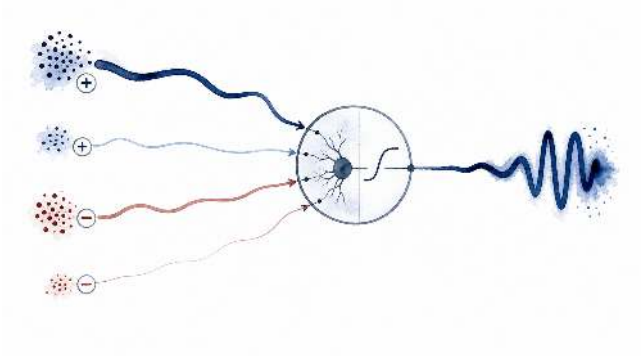
Input vector

$$\mathbf{x} = [\,85,\ 2,\ 12, \ldots\,]$$

The order and meaning of the entries must be consistent across cases.

Pixels, text pieces, and accounting fields all need a numerical representation before a neural network can process them.

A neuron weighs evidence for and against



Three learned roles

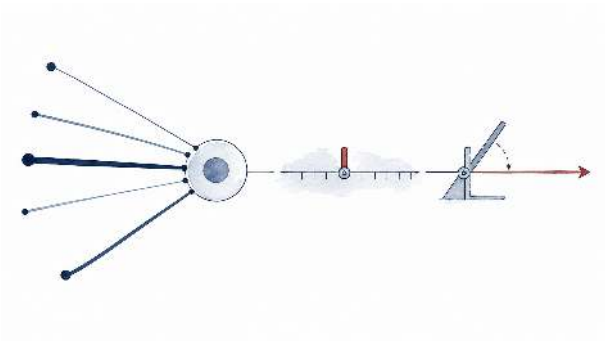
- A **positive weight** supports the signal.
- A **negative weight** pushes against it.
- The **bias** supplies a starting point.

$$z = w_1 x_1 + w_2 x_2 + b$$

z is the combined evidence before the unit decides what to pass on.

Training learns the weights and bias from examples; a person does not set each one by hand.

One tiny calculation: multiply, add, then apply ReLU



A numerical example

Inputs $\mathbf{x} = [2, 1]$, weights $\mathbf{w} = [+1, -1]$, and bias $b = 0$:

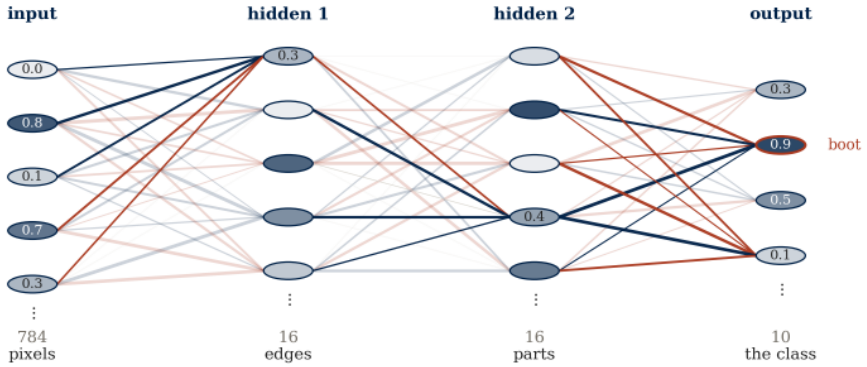
$$\begin{aligned} z &= (+1)(2) + (-1)(1) + 0 \\ &= 1 \end{aligned}$$

$$a = \mathrm{ReLU}(z) = \max(0, z) = 1$$

If $z < 0$, ReLU outputs zero. If $z > 0$, it passes the positive value forward.

A deep network repeats this small calculation across many units and layers.

Layers turn one numerical representation into another



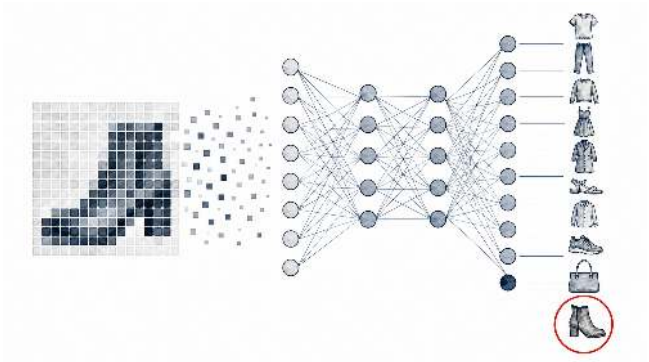
Input layer
Receives the encoded case: here, 784 pixel values.

Hidden layers
Learned weights build intermediate signals and patterns.

Output layer
Produces scores for the possible answers.

Different weights make units respond to different patterns; each layer passes its outputs to the next.

A forward pass turns an input into output scores



Forward pass

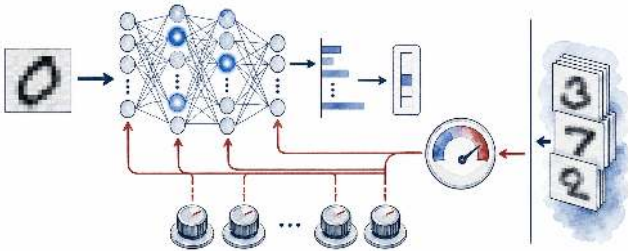
The input moves through the current learned weights and becomes a set of scores.

No learning yet

Producing this output does not by itself change any weight.

The highest score can support an answer, but it is not a guarantee that the answer is correct.

Training uses error to adjust the weights

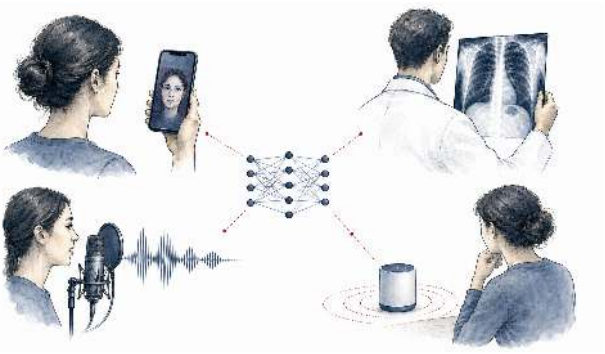


predict $\longrightarrow$ compare with the known answer $\longrightarrow$ adjust $\longrightarrow$ repeat

Backpropagation assigns responsibility for the error; an optimiser makes small changes to the relevant weights.

After training, the weights are frozen; unseen cases reveal whether the learned pattern travels.

The architecture should respect the structure of the data



Images
Reuse local detectors across positions.

Ordered data
Preserve order and context in text, sound, or time series.

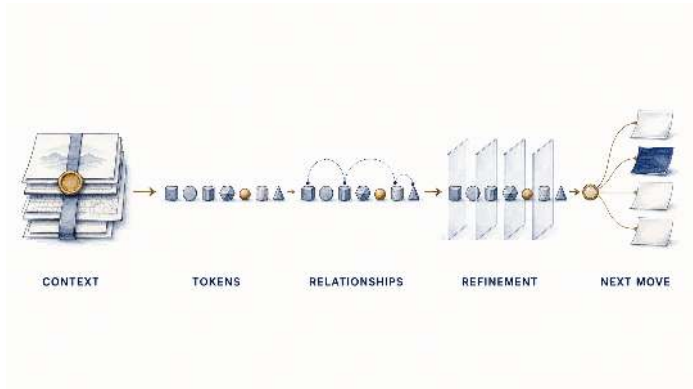
Connected data
Use links among people, firms, accounts, or assets.

These are specialised ways to organise the same ingredients: numbers, weights, layers, and outputs.

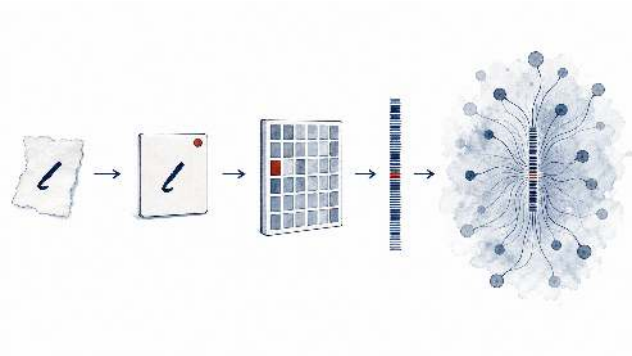
Part 4

How language models continue text

Text becomes tokens and vectors; attention mixes contextual clues; Transformer layers refine them; the model proposes the next token.



Text becomes tokens, token IDs, and vectors



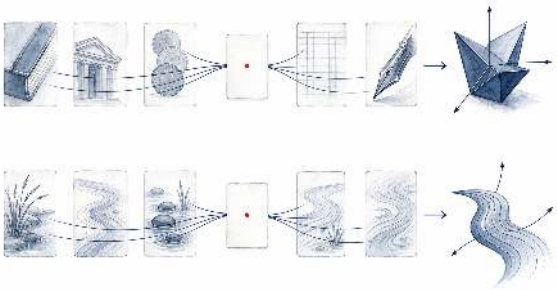
Token
A model-specific text piece: a word, part of a word, or punctuation.

Token ID
An index that identifies the piece in the model's vocabulary.

Starting vector
Learned numbers give the first layer something it can process.

The model does not receive a sentence as human-readable meaning; it receives an ordered sequence of numerical representations.

Tokens become vectors; context gives them meaning



Starting embedding

A learned vector gives the token an initial numerical representation.

Contextual representation

Surrounding tokens and position refine `bank` differently in `bank loan` and `river bank`.

Meaning is not stored in one permanent token vector; it is refined through context.

Attention lets a token gather the clues it needs

Example

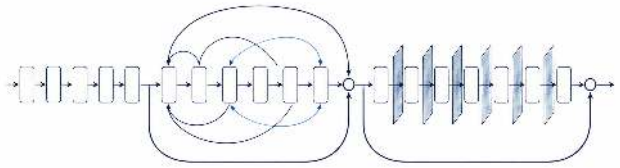
> `The bank approved the loan after reviewing the borrower.`

To refine `bank`, the model can give more influence to `approved`, `loan`, and `borrower` than to less relevant positions.

Attention is relevance-weighted information mixing.

An attention weight measures influence in this calculation; it is not factual confidence and not an explanation by itself.

A Transformer repeatedly mixes and transforms information



1. Attention

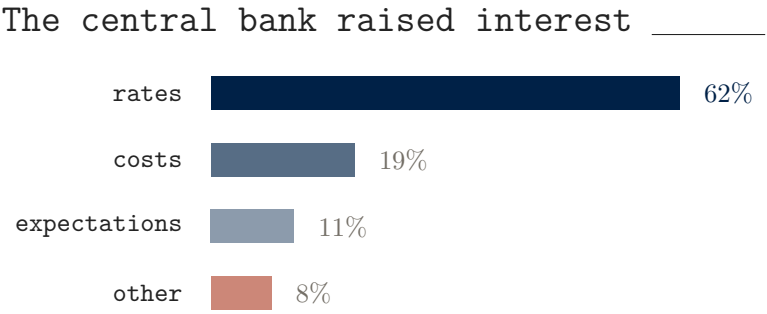
Each position gathers relevant clues from the available context.

2. Small neural network

Each updated position is transformed before the next block repeats the process.

Repeated blocks refine what every token represents in this particular context.

The model chooses among possible next tokens



What the percentages mean

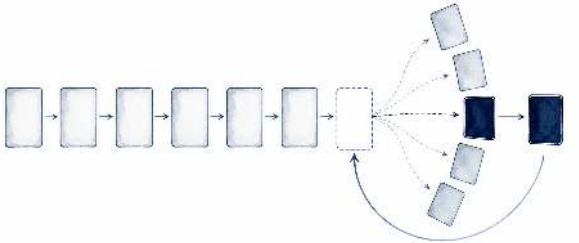
Plausible continuations under the current context.

What they do not mean

The probability that a whole statement is factually true.

Probabilities are illustrative and do not come from a deployed model.

Generation appends one token and repeats

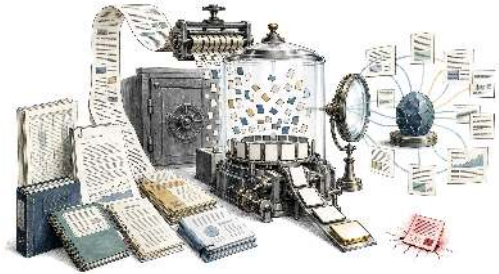


read the context ⟶ score possible next tokens ⟶ append one ⟶ repeat

Each selected token becomes part of the context for the next step.

A fluent paragraph is built through many local next-token choices.

Capability is learned in stages; a prompt is temporary context



1. Pretraining
Practise continuation across very large text collections.

2. Post-training
Examples and feedback shape instruction following, safety, and tool use.

3. A prompt
Supplies temporary context for one request; it does not normally change the learned weights.

Pretraining, post-training, and prompting happen at different stages and do different jobs.

The host builds the request before each model call

THE APP ASSEMBLES A REQUEST PACKET

```
{
 "instructions": ["Follow safety and privacy rules.",
                  "Use tools for current facts."],
 "user_question": "What is the weather in Oxford?",
 "tool_result": "Oxford weather; observed at 09:00.",
 "allowed_tools": ["get_weather"]
}
```

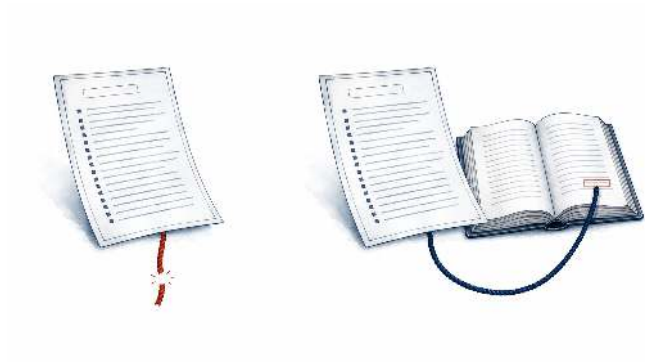
The model receives the whole packet—not only the user's sentence.

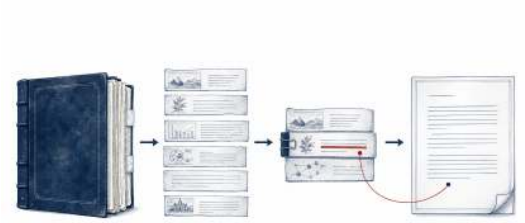
The app—also called the host—controls instructions, evidence, allowed tools, state, and permissions. An agent is this governed system around the model.

Part 5

A capable model can still fail in use

Accuracy is not the same as truth, robustness, privacy, security, or value. The surrounding evidence and decision process remain essential.





Retrieval can add current evidence

1. retrieve candidate passages;
2. place selected passages in the context;
3. generate an answer and check each important claim.

A citation is a pointer, not proof. The cited span must actually support the claim.

Retrieval reduces missing-context risk; it does not make the model or the source automatically correct.

Risk can enter through data, change, access, or misuse

Historical bias

Past decisions may reflect past policy and unequal treatment.

Changing regimes

A learned pattern may fail after a crisis, rate shift, or rule change.

Privacy

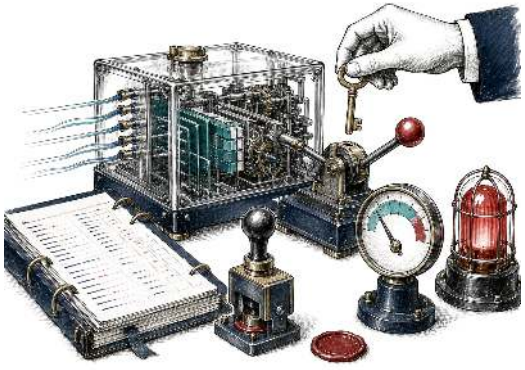
Prompts, files, and logs can expose private information.

Security

Untrusted content may manipulate the model, a connected tool, or its user.

Ask not only how accurate the model is, but under which data, regime, access, and controls.

Complexity must earn its place—and its controls



Compare

Test a complex candidate against a simple baseline on the same unseen cases.

Count the full cost

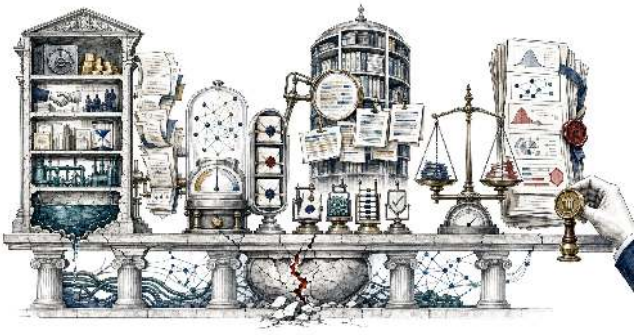
Include data, computing, monitoring, explanation, and human review.

Keep checking

Ownership, access, monitoring, and intervention continue after deployment.

The best model is the simplest one adequate for the decision and the evidence available.

Next, we will see how these models are used in finance



How much?
Prices, cash flow, and risk.

Will it happen?
Default, fraud, and distress.

Groups or flags?
Customers, regimes, and transactions.

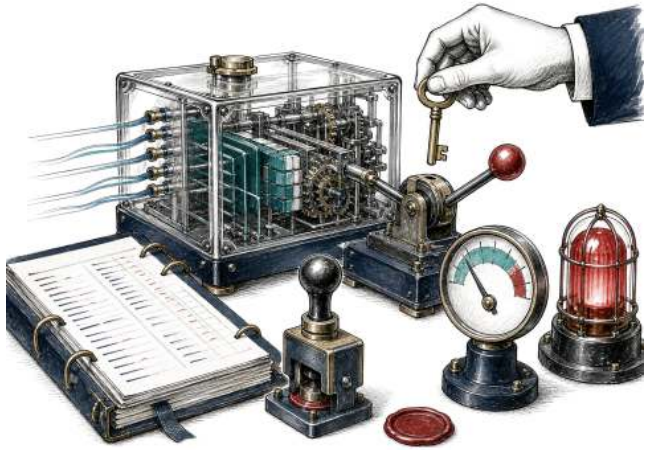
What does text say?
Filings, news, and calls.

For every application: what is predicted, what evidence supports it, which errors matter, and who acts on the output?

Part 6

Risks in AI-assisted decisions

When scores are pointed at people, the same learning machinery carries duties of fairness, evidence, appeal, and accountability.



Consequences of errors in AI-assisted decisions

Everything this week scored things: prices, rates, response records. The same machinery is now routinely pointed at people, at scale:

- About 99% of Fortune 500 companies filter job applicants through automated tracking systems; 88% of employers say qualified candidates get screened out before a human looks (Harvard Business School, 2021).
- In McKinsey's 2024 survey, 65% of organisations were already using generative AI in at least one business function.
- A mispriced flat costs pounds; a misjudged person can lose a loan, a job, or years of their life. The 712 missed subscribers we tolerated were a budget line; on people, the same error rates become questions of justice.

Nothing in the mathematics changes. Every duty does. This part is a casebook: real systems, real people, and the lesson each case leaves behind.

Historical bias in model targets

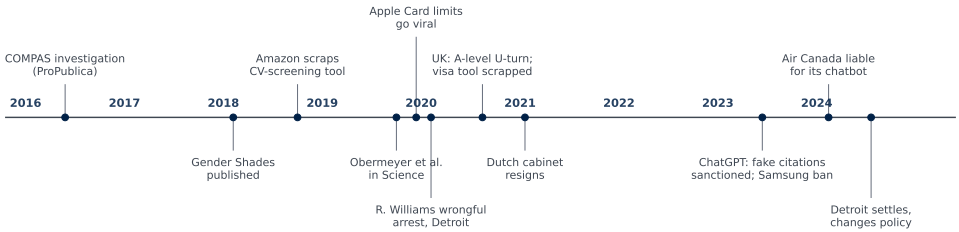
The week's quiet lesson, stated plainly: a model learns the target it is given, and the target carries the past.

- Our own target was "said yes in 2008 to 2010", not "should be called": the model inherited the old campaigns' choices.
- Day 1's recruiting case, in full: Amazon trained an experimental CV screener on a decade of its own hires; it learned the past's imbalance, marked down CVs containing "women's", and was scrapped (Reuters, 2018).
- Day 1's healthcare case: "cost" stood in for "need" in a system applied to about 200 million people a year; fixing the proxy would nearly triple the share of black patients flagged for extra care, 17.7% to 46.5% (Obermeyer et al., *Science*, 2019).

No malice is required anywhere in the pipeline. The target does the damage on its own.

Documented failures of automated decision systems

None of what follows is hypothetical. Every event on this line has a named person, a court ruling, a regulator's report, or a resignation attached:

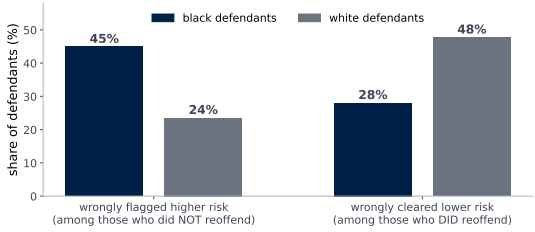


- The pattern repeats across justice, hiring, credit, education, immigration, welfare and customer service, under different laws.
- The next slides take six apart, each with the same question: what exactly failed, and what would have caught it?

Each event is sourced on its case slide and in the Sources pages.

Fairness trade-offs in the COMPAS score

COMPAS: *Correctional Offender Management Profiling for Alternative Sanctions*. It estimates reoffending risk. ProPublica compared 7,000 Florida scores with two-year outcomes:

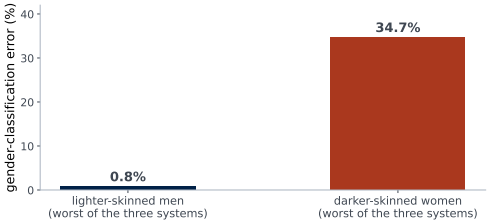


- Overall accuracy was similar by race; the *mistakes* were not: they fell on opposite groups, in opposite directions.
- The vendor's defence was also true: a given score meant the same reoffending rate in both groups. Both properties cannot hold at once when base rates differ, so fairness is a *choice*, and someone must own it.

ProPublica, "Machine Bias" and methodology (2016); impossibility: Kleinberg et al. (2016), Chouldechova (2017).

Subgroup error rates in face analysis

In 2018 Buolamwini and Gebru tested three commercial face-analysis systems (Microsoft, IBM, Face++), each with a high advertised overall accuracy, on a balanced set of faces. By subgroup:



- Error rates ran from under 1% for lighter-skinned men to 34.7% for darker-skinned women, a gap invisible in the single headline number.
- The audit worked: vendors cut the gaps within months, and in June 2020 IBM withdrew from general-purpose face recognition altogether.

Buolamwini and Gebru, *Gender Shades*, PMLR 81 (2018). Test one confusion matrix per group: this week's tool, applied.

Wrongful arrests from face-recognition matches

What happened	January 2020: Detroit police ran a grainy CCTV still through face recognition; it returned Robert Williams' licence photo. He was arrested in front of his family and held about 30 hours. He had nothing to do with the crime.
Not a one-off	In 2023 Porcha Woodruff, eight months pregnant, was wrongly arrested the same way. At least three such wrongful arrests are documented in Detroit alone.
The outcome	The ACLU sued; in 2024 Detroit paid $300,000 and adopted one of the strictest US police policies: a face match alone can never justify an arrest, only a lead to corroborate.

- Three numbers were confused: the vendor's lab accuracy, the subgroup error rate (previous slide), and what the output *is*: a lead, not an identification.

A score treated as proof: the exact failure the Dutch affair repeats at national scale.

The 2020 A-level grading model

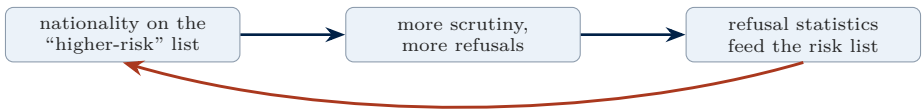
The task	Covid cancelled England's 2020 exams. Ofqual's model adjusted teacher-assessed grades to keep national results in line with history, leaning on each school's past performance.
What happened	About 39% of teacher-assessed A-level grades were pulled down. Small cohorts, often private schools, largely kept teacher grades; strong pupils in large state schools were capped by their school's history. Street protests followed.
The outcome	On 17 August 2020 the government reversed course and restored teacher grades; the chief regulator resigned within days; the Prime Minister blamed a "mutant algorithm".

- The model did roughly what it was designed to do. The design goal, hold the national distribution steady, was itself the injustice: a pupil was graded by her school's past, not her own work.

Accurate on average is not acceptable case by case. For a policy maker, the algorithm *is* the policy.

Nationality and feedback loops in visa screening

The system	From 2015 the UK Home Office streamed every visa application red, amber or green. Nationality was an input; a secret list of "higher-risk" nationalities pushed applicants into the red stream.
The outcome	Challenged as discriminatory by JCWI and Foxglove, the Home Office scrapped the tool on 7 August 2020, days before the case would have been heard. Foxglove's director called it "speedy boarding for white people".



- The Dutch affair's two ingredients, nationality as an input and a self-reinforcing loop, caught earlier because someone sued. External scrutiny, not internal accuracy metrics, stopped the harm.

Transparency and the Apple Card investigation

What happened	In November 2019 complaints went viral, including from Apple's co-founder, that wives received far lower credit limits than husbands with shared finances. New York's financial regulator investigated Goldman Sachs, the issuer.
The finding	After reviewing about 400,000 applications, the regulator found *no unlawful discrimination*: similar credit profiles got similar limits, and the card scored individual, not household, credit histories.
The critique	The same report faulted the programme anyway: customers could not get an intelligible reason for their limit, and "the algorithm decided" satisfied no one.

- Two lessons in one case: a viral anecdote is not an audit; and an audit is not an excuse for opacity. An unexplainable decision is a business failure even when it is lawful.
- The questions stay open at scale: across two million 2019 US mortgage applications, lenders were about 80% more likely to deny black applicants than similar white ones, 17 factors held constant (The Markup, 2021; an estimate, not a ruling).

Organisational liability for generative AI outputs

Three documented failures from the generative-AI era, each with a price attached:

Case	What happened	The bill
Mata v. Avianca (US court, 2023)	lawyers filed six case citations ChatGPT had invented; asked whether they were real, ChatGPT said yes	$5,000 sanction, careers dented
Moffatt v. Air Canada (2024)	the airline's chatbot invented a refund policy; the airline argued the bot was "a separate legal entity" and lost	CAD $812, and the precedent
Samsung (2023)	engineers pasted confidential source code into ChatGPT to debug it; the material could not be recalled	a company-wide ban

- A fluent answer is not a sourced answer; asking the model to check itself is not verification. Your organisation owns what its AI says, and what staff type into a public tool leaves the building.

Dutch childcare benefits scandal: case overview

The Netherlands, 2005 to 2019. The tax authority screened childcare benefit claims with a self-learning risk model, and treated high risk scores as grounds for action.

- Around 26,000 parents were wrongly accused of fraudulent claims, and many were ordered to repay years of benefits in full, immediately.
- Among the inputs was the claimant's nationality. Families with dual citizenship were flagged disproportionately.
- The consequences ran through debt, lost homes and broken families; reporting has linked more than a thousand children's foster placements to the affair, with some estimates higher.
- In January 2021 the Dutch government resigned over the scandal.

Parliamentary inquiry *Unprecedented Injustice* (2020); Amnesty International, *Xenophobic Machines* (2021).

Dutch childcare benefits scandal: system failures

Set the scandal against this week's own checklist:

The week's tool	What it would have asked
error rates by group (Day 1)	who is being falsely flagged, and is any group carrying the mistakes?
asymmetric costs (Day 2)	a wrongly accused family is not a €4 mistake; what threshold do the real costs imply?
drift and monitoring (Day 3)	is the model still valid on today's claimants?
audit before scale (today)	was the score ever checked against verified outcomes before it was acted on at scale?

- The deepest failure sat outside the model: **a risk score was treated as proof**, and families had to prove their innocence against it.

Common failure patterns and safeguards

The pattern	Seen in	The defence, from this week
a proxy target carries the past	Amazon hiring; Obermeyer; the bank's own target	interrogate the target column before fitting
an average hides the subgroups	COMPAS; Gender Shades	one confusion matrix per group
a score treated as proof	Williams arrest; the Dutch affair	a score is a lead; humans must stay able to say no
secret criteria, feedback loops	visa streaming; the Dutch affair	document inputs; audit against verified outcomes
right on average, wrong per person	the A-level algorithm	individual appeal routes; legitimacy is not a metric
fluent fabrication, owned words	Avianca brief; Air Canada bot	verify against sources; you own your AI's statements

Every defence in the right column is a tool this course already taught. The failures were failures of use, not of exotic mathematics.

Safeguards for high-stakes AI systems

- **Document** the target, the data's origin, the split, the scores, and the costs behind every threshold: this week's habits, written down.
- **Test by group**, not only on average: one confusion matrix per group the decision could harm.
- **Keep a human path**: a score triggers review by someone empowered to disagree, and the subject can appeal.
- **Monitor drift**: Day 3 showed models age; on people the ageing is silent until it is a scandal.
- **Know the law**: the EU AI Act classifies systems that assess creditworthiness, make employment decisions, or decide access to essential public services as high risk, with duties close to this list.

Predicting risk about a person is not evidence of wrongdoing by that person. Build every process as if that sentence will be tested in court, because it will be.

Checklist for evaluating AI-assisted decisions

The casebook compresses into questions anyone can ask, whichever seat you sit in:

As a…	Ask…
consumer	Can I get a reason for this decision, and is there a route of appeal to a human? (Apple Card; A-levels)
buyer or builder	What was the target, who chose it, and what history does it carry? What are the error rates *by group*, on held-out data, not the headline average? (Amazon; Obermeyer; Gender Shades)
user of AI output	Is this answer sourced, or only fluent? Who is accountable when it is wrong, and what am I typing into it? (Avianca; Air Canada; Samsung)
policy maker	Is the score used as a lead or as proof? Are the criteria inspectable, is there a feedback loop, and who audits outcomes at scale? (Williams; visa tool; the Dutch affair)

The questions cost nothing and would have caught every case in this part. Asking them is what "responsible AI" means in practice.