FOUNDATIONS TO THE 2026 RESEARCH FRONTIER

The 180-Day Descent

By Claude Opus and GPT
Human editor: Jason Lau

CONTENTS

THE 180-DAY MAP

Introduction

How to read a map that descends from foundations to the frontier.

This book began with a hunger rather than a credential: deep curiosity, learning for its own sake, and the wish to become at home in the world without pretending the world is small. The intended reader is a curious generalist: strong in some places, full of gaps in others, unwilling to choose between foundations and the frontier. The promise is not mastery in 180 days. It is orientation: a map of the major structures that make reality, life, mind, technology, society, and the future intelligible.

AI systems perform the project's deep research, synthesis, and first-pass writing, but the work is not published untouched. Human editor Jason Lau manually checks the material, improves readability and structure, and keeps the course focused on clear explanations rather than raw generated output.

The sequence begins with a constraint. The frontier is only useful if the instruments of belief are calibrated first. So the course does not open with cosmology, artificial intelligence, or medicine. It opens with knowledge itself: what counts as a reason, why true belief can still be luck, how science separates testable claims from protective stories, and how probability lets a mind live without certainty. Only then does the descent widen into mathematics, physics, chemistry, biology, medicine, neuroscience, AI, economics, civilization, ethics, and the forces now bending the future.

Each day is built to work even when time is uneven. It starts with a puzzle, story, image, analogy, or thought experiment; builds a mental model; names the live debate; then walks as far toward recent, trustworthy research as the evidence allows. The spirit is close to a very short introduction, but with a steeper internal

slope: begin as if the reader is smart but new here, then descend until the ground becomes genuinely current and contested.

The order matters. This is not a cabinet of 180 interesting facts. It is dependency-ordered: epistemology before statistics, statistics before experiments, mathematics before physics, thermodynamics before life, evolution before mind, and computation before modern AI. The arc makes room for deeper foundations where compression would be dishonest, and for sustained descents into frontier controversies such as the Hubble tension, origin-of-life physics, mammalian epigenetic inheritance, consciousness theories, AGI and alignment, and the deep history of inequality.

Five threads run through the whole course:

- **Information**, because every discipline eventually asks what is signal, what is noise, and what can be transmitted or inferred.
- **Energy**, because the physical cost of order returns in thermodynamics, life, economics, climate, and computation.
- **Evolution**, because selection is not just a biological mechanism; it is a pattern for knowledge, culture, technology, and institutions.
- **Emergence**, because many of the most important objects in the map are collective: temperature, cells, markets, minds, societies.
- **Computation**, because formal procedure becomes a language for mathematics, physics, brains, and machines.

The hype filter is part of the method. Frontier claims are marked as **established**, **promising hint**, or **contested/hype**. Physics and cosmology claims need datasets and error bars. Medical, AI, and social-science claims need replication, incentives, measurement, and humility. A result can be exciting and still not carry much weight. A failed claim can still be useful if it teaches us how science corrects itself. Recent does not mean reliable; peer-reviewed does not mean settled; beautiful does not mean true.

The first four days set the tone. Day 1 asks why a stopped clock can give you a true, justified belief without giving you knowledge; Day 2 scales that worry up to

science as an institution; Day 3 opens the reasoning engine itself; Day 4 turns uncertainty into a calculus, using Monty Hall, Bayes' theorem, and e-values to show how belief should move when evidence arrives.

That is the descent: not a catalog of facts, but a course in how facts earn their keep.

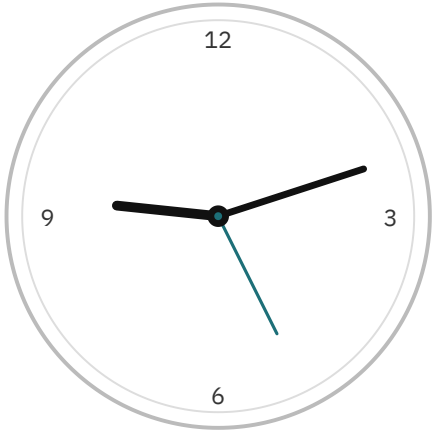
BLOCK I

Foundations of Knowledge & Reasoning

BLOCK I · FOUNDATIONS OF KNOWLEDGE & REASONING · DAY 01 / 180

What Is Knowledge?

You looked at the clock. You were right. Did you know?



● STOPPED 12 H AGO — BUT RIGHT, FOR THIS ONE MINUTE

It is 9:12 in the morning and you are late. You glance up at the great station clock as you rush past, read **9:12**, and think: *fine — three minutes to spare.* You are right. It really is 9:12. And yet the clock you trusted died at 9:12 exactly twelve hours ago, somewhere in the small hours, and has hung there frozen ever since. You consulted a broken instrument at the single instant in the day it happened to be correct.

Your belief was **true**. It rested on a perfectly sensible **reason** — clocks tell the time, and you have trusted a thousand of them without incident. You **believed** it sincerely. So: did you *know* it was 9:12? Asked carefully, almost everyone says no.

Something is missing. Saying exactly what has consumed philosophers for sixty years — and, as we'll see, the better part of a thousand.

This is the first descent, so there is nothing behind us yet — the log is blank. Instead we plant seeds. The machinery introduced today (belief as something that comes in *degrees*; updating on evidence; minds as inference engines) is the epistemic toolkit the entire course will lean on. Watch for it to resurface on **Day 2** (how science decides what counts at all), **Day 4** (probability as the logic of partial belief), **Day 7** (information), **Day 119** (the predictive brain), and **Day 149** (when famous results evaporate). The five threads we'll trace across all 180 days — *information, energy, evolution, emergence, computation* — all have a quiet first appearance right here.

THE MODEL

The three-legged stool

For roughly twenty-three centuries, Western philosophy carried around a tidy answer to "what is knowledge?" To *know* that something is the case, you needed three things at once:

(1) you believe it — you can't know what you don't even hold to be true. **(2) it's true** — you can't *know* a falsehood; people who said "I knew the Earth was flat" merely *believed* it, confidently and wrongly. **(3) you're justified** — you have good reason, because a lucky guess that lands isn't knowledge either. The gambler who "just had a feeling" the long-shot would win, and won, did not *know* it would.

Knowledge, on this view, is *justified true belief* — JTB, a three-legged stool. Kick away any leg and it topples. The picture is usually traced to Plato, who in the *Theaetetus* floats the idea that knowledge is "true judgement with an account." There's a delicious irony here, much enjoyed by historians: in that very dialogue Socrates then dismantles the definition, so Plato arguably never endorsed the

thing named after him. As one scholar put it, it is almost as if a distinguished critic created a tradition in the very act of destroying it.

Still, the rough consensus held. The stool seemed stable. And then a 35-year-old philosopher who, the story goes, hadn't published much and rather needed to, wrote three pages.

THE GRENADE

Gettier's three pages

In 1963, Edmund Gettier published a paper in the journal *Analysis* with the cheekily plain title *"Is Justified True Belief Knowledge?"*. It runs barely three pages. It has since been cited in **thousands** of scholarly works and spawned entire subfields. Few documents in modern philosophy have done more damage per word.

Gettier's move was devastatingly simple. He built little stories in which all three legs of the stool are firmly in place — belief, truth, justification — and yet you'd never say the person *knows*. Here is his first case, lightly modernized:

> *Smith and Jones both apply for a job. The boss tells Smith, "Jones will get it." Smith has also, idly, counted the coins in Jones's pocket: ten. So Smith forms a justified belief: the person who gets the job has ten coins in their pocket.*

Now the twist. The boss was wrong (or changed her mind): **Smith** gets the job, not Jones. And — entirely unknown to Smith — Smith happens to have **ten coins** in his own pocket too. Look at his belief, "the person who gets the job has ten coins": it's **true** (the winner, Smith, does have ten coins), it's **justified** (excellent evidence — the boss's word, a literal coin count), and it's sincerely **believed**. JTB, all three legs. Yet Smith plainly doesn't *know* it. He was tracking *Jones* and arrived at the right answer about the wrong man.

That is the anatomy of a *Gettier case*: your justification runs *through a falsehood* ("Jones will get the job"), and the belief is rescued into truth by an unrelated *coincidence* ("Smith also has ten coins"). The reason and the truth never actually touch. The stopped clock is the same skeleton in cleaner clothes: your reason (the clock) is broken, and the truth (it's 9:12) arrives by luck.

A TWIST OLDER THAN ITS NAME

Gettier wasn't first. Bertrand Russell had the stopped-clock case in *Human Knowledge: Its Scope and Limits* (1948). Go back further and the problem is downright ancient: around **770 CE** the Buddhist logician **Dharmottara** described a traveler who sees what looks like smoke over a hill, infers fire, and is right that there's fire — except the "smoke" was a swarm of insects. Same skeleton, twelve centuries early. In 14th-century India, **Gaṅgeśa** built a whole causal theory of knowing to handle such cases. The "Gettier problem" is one of philosophy's great instances of *convergent discovery* — the kind of thing minds keep tripping over independently, which is itself a hint that something real is there.

The Gettier Machine

CASE	BELIEF	TRUTH	JUSTIFICATION	LUCK	VERDICT
Plain knowing	Yes	Yes	Yes	No	Knowledge on the classic view
Stopped clock	Yes	Yes	Yes	Yes	Not knowledge: truth arrives by coincidence
Lucky guess	Yes	Yes	No	Yes	Not knowledge: no justification
Confident error	Yes	No	Yes	No	Not knowledge: the claim is false

THE PATCH WARS

The hunt for the fourth leg

The obvious response to Gettier was: add a fourth condition that screens out the luck. For decades, epistemologists tried — and each tidy fix met a nastier counterexample. It became a minor blood sport.

No false lemmas. First idea: knowledge can't be reasoned *through* a falsehood. Smith's belief leaned on "Jones will get the job," which was false; ban that and you're safe. Clean — until Alvin Goldman's **fake-barn country** (1976). You're driving through a region where, as a prank, every "barn" is a flat movie-set façade — except one. You happen to glance at the single real barn and think "a barn." Your belief is true, justified, and rests on *no* false premise. Yet you don't know it's a barn: you could so easily have been fooled by a façade a hundred meters either way.

Track the truth. So maybe knowledge is about how your belief behaves across *nearby possibilities*. Robert Nozick (1981) proposed *sensitivity*: you know *p* only if, *were p false, you wouldn't believe it.* Elegant — but it produces strange verdicts in edge cases. Ernest Sosa (1999) flipped it into *safety*: in all the nearby ways things could have gone, you wouldn't have been wrong. The stopped clock fails safety hard (a minute either side and you're mistaken); a working clock passes. Fake-barn-you fails safety too.

Then Linda Zagzebski (1994) delivered the gut-punch with a kind of **recipe** for defeating *any* such fix. Take a belief that's justified but could still be false (which justification, being fallible, always allows). Arrange for the justification to misfire so the belief is false — then arrange, by luck, for it to be true after all. As long as your fourth condition stops short of demanding that the justification *guarantee* the truth, luck can always wedge back in. The patch wars may be structurally unwinnable.

Two ways to stop fighting

Declare knowledge a primitive. Timothy Williamson, in *Knowledge and Its Limits* (2000), made a radical move: stop trying to build knowledge out of simpler parts. Maybe it has no analysis. On his *knowledge-first* view, knowing is a basic mental state — the most general *factive* one — and we should explain belief, evidence, and justification *in terms of knowledge*, not the other way around. You can't define *hydrogen* or *John F. Kennedy* into simpler concepts; perhaps knowledge is bedrock too. Sixty years of failed definitions start to look less like a puzzle and more like a clue.

Make it about competence. The other escape is *virtue epistemology* (Sosa again). Knowledge is *apt* belief — a belief that is true *because of* the knower's skill, not by accident. Picture an archer. A bullseye is a good shot only if the arrow hit center *because* the archer aimed well — not because a gust blew a bad shot onto the target. The Gettiered believer is exactly that archer: the wind knocked the arrow off course, then a second gust knocked it back onto the bull. Accurate, yes. Skillful, no. *Apt*, no. That, says Sosa, is why luck-based hits aren't knowledge.

THE DEBATE

What makes a belief justified at all?

Step back from "is it knowledge?" to the humbler leg: what makes a belief *justified* in the first place? Push on any justification and you fall into a regress. It's 9:12 because the clock says so. Trust the clock because clocks are reliable. Believe *that* because… and now you're sliding. The ancient skeptics mapped the trap precisely. Every chain of justification, they argued, ends in one of three uncomfortable places — the *Agrippan trilemma*: it goes on **forever**, or it loops back in a **circle**, or it stops at some **arbitrary** point you simply declare.

Three modern schools each pick which horn to grab — and a fourth changes the subject entirely.

DIAGRAM · THE REGRESS PROBLEM

Agrippa's Trilemma — three bad endings, four escapes

Why is your belief justified? Every honest answer to "...and why *that*?" eventually hits one of three walls.

Reason chain: belief: "it's 9:12" -> because "the clock" -> because "...and why that?"

1. **Infinite regress:** every reason needs another reason forever.
2. **Circle:** the chain loops back to something it already used.
3. **Arbitrary halt:** the chain simply stops at a basic commitment.

Foundationalism – bites the third bullet: some beliefs are *basic* and need no further support (raw experience, simple logic). The chain stops, but not arbitrarily.

Coherentism – embraces the circle, but makes it virtuous: no belief stands alone; a belief is justified by how well it hangs together with the whole web. (A first taste of *systems thinking*, Day 9.)

Infinitism – the brave minority: accepts that justification is an endless chain of reasons, never bottoming out.

Reliabilism – changes the question. A belief is justified if it was *produced by a reliable process* – good vision, sound memory – whether or not you can recite a defense. This is *externalism*: justification can be a fact about your wiring, not a story in your head.

That internal/external split matters more than it looks. The **internalist** says justification must be something you can access by reflection – reasons available "from the inside." The **externalist** (reliabilism's home) says what matters is that your belief was, in fact, produced in a truth-conducive way, accessible or not. Hold that tension in mind: it is exactly where the old armchair questions collide with the new science of how brains actually form beliefs.

THE FRONTIER · 2026

Three live edges — and the hype filter

Every day in this course ends at the research frontier, with each claim tagged for how much weight it can bear. Knowledge sits at a fascinating junction right now: philosophers, psychologists, and neuroscientists are all circling the same questions from different sides.

Edge 01 [SUPERSEDED] [ESTABLISHED]

Are "knowledge" intuitions universal — or just Western?

When the discipline runs on "asked carefully, almost everyone says no," a natural worry is: *which* everyone? In 2001, the founding study of *experimental philosophy* — Weinberg, Nichols & Stich — reported that the Gettier intuition varies by culture, with East-Asian participants supposedly more willing to grant the lucky believer "knowledge." If true, it was a bombshell: philosophy's whole method of consulting intuitions looked parochial.

The bombshell did not survive contact with replication. In **"Gettier Across Cultures"** (*Noûs*, 2017), Machery, Stich, Rose and colleagues tested Brazil, India, Japan, and the United States with cases taken near-verbatim — and found the *opposite*: in **every** group, people robustly refused to call the Gettiered belief knowledge. A separate replication (Kim & Yuan) failed to reproduce the original cross-cultural gap even with a far larger East-Asian sample. The current best reading is that there may be a **universal core "folk epistemology"** that recoils from luck-based knowing. The deeper lesson is one we'll meet at industrial scale on **Day 149**: the splashiest finding is often the one careful re-testing quietly walks back.

Edge 02 [ESTABLISHED] [CONTESTED]

Belief by the dial, not the switch: Bayesian epistemology

Maybe the all-or-nothing picture of belief was the wrong starting point. *Bayesian epistemology* says your real epistemic states are *credences* — degrees of confidence on a scale from 0 to 1. Rationality then needs just two rules: your credences must obey the laws of probability (*coherence*), and you must revise them by *conditionalization* as evidence comes in.

Why obey? The **Dutch book theorem** (Ramsey, 1926; de Finetti, 1937) supplies a startlingly concrete answer: if your credences break the probability laws, a clever bookmaker can offer you a set of bets you'll each accept as fair, but which together guarantee you lose money *no matter what happens*. Incoherent confidence isn't merely untidy — it's exploitable. The dial below lets you feel the trap close. What's still *contested* is whether graded credence *replaces* ordinary yes/no belief or merely sits beside it. (The lottery paradox bites here: you're 99.9% sure your ticket loses — but do you flat-out *believe* it loses?) We pick this thread up properly on **Day 4**.

The Credence Dial and the Dutch Book

If your credence in *S* and your credence in *not-S* sum to 1.00, the pair is coherent. If they sum above 1.00, you will overpay for bets where exactly one can win. If they sum below 1.00, a bookie can reverse the bets and still guarantee a profit.

CREDENCE IN S	CREDENCE IN NOT-S	SUM	RESULT
0.50	0.50	**1.00**	Coherent
0.70	0.60	**1.30**	Guaranteed 0.30 loss if you buy both $1 bets
0.30	0.40	**0.70**	Guaranteed 0.30 loss if the bookie buys both bets from you

Edge 03 [PROMISING] [CONTESTED]

Where do beliefs come from? The brain as a prediction machine

Philosophy asks what justifies a belief; neuroscience now asks how a lump of tissue forms one. A fast-growing program answers: the brain is not a passive sponge soaking up the world — it is a relentless *prediction machine*. On the *predictive-processing* view (Andy Clark, *Behavioral and Brain Sciences*, 2013; Jakob Hohwy, 2013), the brain constantly generates a model of its surroundings, predicts the sensory signals it expects, and forwards only the *prediction errors* — the surprises — up the hierarchy. Perception becomes the brain's best running

guess, reined in by error; in Anil Seth's memorable phrase, a "controlled hallucination." Belief-updating starts to look like **Bayesian inference rendered in neurons** — the so-called "Bayesian brain," tying Edge 02 to wetware.

Karl Friston pushes the idea to its limit with the *Free Energy Principle* (*Nature Reviews Neuroscience*, 2010): living systems persist precisely by minimizing a quantity — "free energy," an information-theoretic cousin of *surprise* — that knits perception, action, and even biological self-organization into one framework. The honest labels matter here. Predictive coding genuinely explains real perceptual phenomena and is a serious, productive research program — **promising**. But the *grand* Free Energy Principle, as a single law for all of mind and life, is widely criticized as so general it is hard to *falsify* — closer to a framework than a tested theory, and so **contested**. We'll return to it for perception (**Day 119**) and consciousness (**Days 123–126**) — and notice already how its "free energy" rhymes with the thermodynamics we'll meet on **Days 33 and 83–85**. *Information, energy, computation, emergence* — four of our five threads, braided into one neuron's quiet arithmetic.

OPEN QUESTIONS

What's genuinely unsettled

Sixty years on, the honest answer to "what is knowledge?" includes a healthy list of things nobody has nailed down:

- **Can knowledge be analyzed at all?** Or was Williamson right that it's bedrock — a primitive we explain other things *with*, not *from*?
- **Internal or external?** Does being justified require reasons you can access by reflection, or just wiring that tends to produce truths?
- **One currency or two?** Is rational belief fundamentally graded (credence), all-or-nothing, or both somehow reconciled?
- **Is there really a universal human epistemology** — and if so, did *evolution* install the instinct that luck-based "knowing" doesn't count? (A thread for **Day 74**.)

- **Is the brain *literally* Bayesian**, or is "the brain does inference" just a useful way of describing it from outside?
- **And the question that will haunt the AI block:** when a system like the one that drafted this page outputs a true, well-supported claim, does it *know* anything — or is it the ultimate Gettier case, right for reasons that have nothing to do with the truth? (**Days 138–145**.)

◆ THE DAY IN THREE SENTENCES

BIG IDEA

For 2,300 years knowledge looked like justified true belief — until Gettier showed in three pages that you can hold all three and still not know, because your reasons and the truth can meet by luck rather than by connection.

BEST ANALOGY

The stopped clock that's right twice a day — and the archer whose arrow is blown off target, then blown back onto the bullseye: accurate, but not *apt*.

LIVE CONTROVERSY

Whether the fix is a fourth condition (and which), whether knowledge is unanalyzable bedrock, and whether "belief" should give way to graded, Bayesian credence — with a real scientific frontier in the claim that the brain is a prediction machine.

THREADS TODAY › information (credence & the Bayesian brain) · energy (Friston's free energy) · computation (mind as inference engine) — with light first touches of emergence and evolution.

SOURCES

Sources & further reading

1. Gettier, E. L. (1963). "Is Justified True Belief Knowledge?" *Analysis* 23(6): 121–123. doi:10.1093/analys/23.6.121. doi.org/10.1093/analys/23.6.121

2. Ichikawa, J. J. & Steup, M. "The Analysis of Knowledge." *Stanford Encyclopedia of Philosophy* (rev. 2018). plato.stanford.edu/entries/knowledge-analysis – JTB, the Gettier cases, safety/sensitivity, and the knowledge-first turn.

3. "Gettier problem." *Wikipedia* (accessed 2026). en.wikipedia.org/wiki/Gettier_problem – precedents in Russell (1948), Dharmottara (~770 CE), and Gaṅgeśa (14th c.).

4. Russell, B. (1948). *Human Knowledge: Its Scope and Limits.* London: Allen & Unwin. – the stopped-clock case (pp. ~170–171).

5. Goldman, A. (1976). "Discrimination and Perceptual Knowledge." *Journal of Philosophy* 73(20): 771–791. – the fake-barn case; reliabilism.

6. Nozick, R. (1981). *Philosophical Explanations.* Harvard University Press. – truth-tracking / sensitivity.

7. Sosa, E. (1999). "How to Defeat Opposition to Moore." *Philosophical Perspectives* 13: 141–153. – the safety condition. See also Sosa (2007), *A Virtue Epistemology* (apt belief).

8. Zagzebski, L. (1994). "The Inescapability of Gettier Problems." *The Philosophical Quarterly* 44(174): 65–73. – the recipe defeating any luck-excluding fix.

9. Williamson, T. (2000). *Knowledge and Its Limits.* Oxford University Press. overview – knowledge-first epistemology; knowledge as the most general factive mental state.

10. Weinberg, J. M., Nichols, S. & Stich, S. (2001). "Normativity and Epistemic Intuitions." *Philosophical Topics* 29(1–2): 429–460. – the founding (later contested) cross-cultural x-phi study.

11. Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N. & Hashimoto, T. (2017). "Gettier Across Cultures." *Noûs* 51(3): 645–664. doi:10.1111/nous.12110. doi.org/10.1111/nous.12110

12. Kim, M. & Yuan, Y. (2015). "No cross-cultural differences in the Gettier car case intuition: A replication study of Weinberg et al. 2001." *Episteme*. philpapers.org/rec/KIMNCD

13. Weisberg, J. "Bayesian Epistemology." *Stanford Encyclopedia of Philosophy.* plato.stanford.edu/entries/epistemology-bayesian — credences, conditionalization, and the Dutch book argument (Ramsey 1926; de Finetti 1937).

14. Clark, A. (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science." *Behavioral and Brain Sciences* 36(3): 181–204. See also Clark, *Surfing Uncertainty* (OUP, 2016).

15. Friston, K. (2010). "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience* 11(2): 127–138. doi:10.1038/nrn2787. doi.org/10.1038/nrn2787

16. Hohwy, J. (2013). *The Predictive Mind.* Oxford University Press.

TOMORROW → DAY 02

The Scientific Method & Demarcation

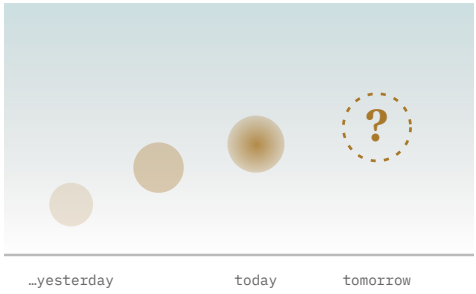
Today we asked when a *single* belief counts as knowledge. Tomorrow we scale the question up to an entire institution: how does science decide which claims even get to enter the arena? Popper's demand that a real theory be *falsifiable*, Kuhn's paradigm shifts, Lakatos's research programmes — and the modern replication crisis as the demarcation line tested under live fire. Bring today's calibration instinct; you'll need it.

END OF DAY 01 · 179 DESCENTS REMAIN

BLOCK I · FOUNDATIONS OF KNOWLEDGE & REASONING · DAY 02 / 180

The Scientific Method & Demarcation

The sun has risen every morning for 4.5 billion years. So it will rise tomorrow — right?



● EVERY PAST SUNRISE IS EVIDENCE — AND PROVES NOTHING ABOUT THE NEXT ONE

Ask a child whether the sun will rise tomorrow and they'll look at you as if you're slow. Of course it will — it always has. That confidence feels like the bedrock of knowledge itself. But press on *why* you believe it, and you walk straight off a cliff that a quiet Scottish philosopher dug in 1739, and that nobody has ever filled in. Your only reason is that the sun has risen before. You are arguing: *the future will resemble the past, because in the past, the future resembled the past.* Read that twice. It assumes the very thing it's trying to prove.

That cliff is called the **problem of induction**, and it is where the entire machinery of science begins — not in triumph, but in a hole. Today we watch thinkers spend two centuries trying to climb out: by giving up on proof and chasing *disproof*

instead; by realizing science doesn't actually work the tidy way the textbooks claim; and finally, in our own decade, by putting the whole question to the harshest test imaginable — **asking thousands of published findings to simply happen again**, and watching a third of them refuse.

Yesterday (**Day 1**) we asked when a *single* belief counts as knowledge, and met Gettier's stopped clock — true belief rescued by luck rather than connection. Today we scale that exact worry up from one mind to an entire civilization-sized institution: how does *science* decide which claims even get to enter the arena? Keep yesterday's tools close. The *credence dial* from Day 1 (belief in degrees, not all-or-nothing) is about to become the only sane reply to Hume; and the hype filter that caught a splashy result quietly walked back by replication is, today, the entire third act.

THE HOLE IN THE GROUND

Hume kicks the legs out

In 1739, a 28-year-old **David Hume** published *A Treatise of Human Nature* — a book so ignored on release that he joked it "fell dead-born from the press." Inside was a bomb on a very long fuse. Hume noticed that every belief we hold about things we haven't directly observed — that bread will nourish us tomorrow as it did today, that the sun will rise — rests on one hidden assumption: that *nature is uniform*, that the unobserved will behave like the observed.

And that assumption, he showed, can't be justified. Not by logic: there's no *contradiction* in a sun that fails to rise. As Hume put it with deadpan precision:

> *That the sun will not rise tomorrow is no less intelligible a proposition, and implies no more contradiction, than the affirmation, that it will rise.*
>
> — Hume, An Enquiry Concerning Human Understanding, §IV (1748)

So uniformity isn't a truth of logic. Could we justify it by experience, then — "it's always held before, so it's a safe bet"? Watch the trap snap shut: that argument *uses* the principle that the past predicts the future in order to *prove* that the past predicts the future. It's circular. You cannot lift yourself by your own bootstraps. Hume's conclusion was genuinely radical, and it's worth stating without softening: we have **no rational justification whatsoever** for our confidence in the future. We are creatures of *habit*, not logic. We expect the sunrise the way a dog expects dinner at the sound of the cupboard — by conditioning, not proof.

This is the wound the scientific method is born trying to dress. If we can never *prove* a general law by piling up confirming instances — no number of white swans proves "all swans are white" — then what on earth is science *doing* when it claims to discover the laws of nature?

A NOTE ON THE BLACK SWAN

Europeans were so sure all swans were white that "black swan" was a centuries-old idiom for *something that doesn't exist* — like "when pigs fly." Then in 1697, Dutch explorers reached western Australia and found rivers full of **black swans** (*Cygnus atratus*). A million confirming sightings had built a rock-solid law; a single bird in Perth shattered it. Hold that asymmetry in your mind — it's about to become the hinge of the whole day.

A single black swan makes the asymmetry visible: confirmations can pile up for centuries, and one counterexample can still break the law.

THE ESCAPE

Popper's judo move: stop trying to prove things

Vienna, the 1920s. A young **Karl Popper** is surrounded by intellectual movements that all claim the prestige of "science": Freud's psychoanalysis, Adler's individual psychology, Marx's theory of history. Their followers are intoxicated. Wherever they look, they see *confirmation* — every slip of the tongue confirms Freud, every twist of politics confirms Marx. And that, Popper realized with a jolt, was precisely what was *wrong* with them.

Because a theory that explains *everything* explains nothing. If no conceivable observation could ever count *against* your theory — if a man saving a drowning child and a man drowning one can *both* be slotted neatly into Freud's framework — then your theory isn't brave. It's empty. It forbids nothing, so the world can't surprise it.

Set that beside Einstein. In 1915, general relativity made an outrageous, *risky* prediction: starlight grazing the sun would bend by a specific amount — 1.75 arcseconds, twice what Newton predicted. If the 1919 eclipse measurements had come back Newtonian, Einstein would have been *finished*. He stuck his neck out. *That*, said Popper, is the signature of real science.

So Popper performed a piece of philosophical judo. Hume is right — you can never *verify* a universal law. Fine. So **stop trying.** Flip the asymmetry of the black swan into a method:

> ***The criterion of the scientific status of a theory is its falsifiability, or refutability, or testability.***
>
> — Popper, Conjectures and Refutations (1963)

You can't prove "all swans are white" by any number of white swans — but a *single* black swan disproves it for good. Verification is hopeless; *falsification* is decisive. Science, on this view, doesn't march from evidence up to certainty. It makes **bold conjectures** and then tries its hardest to **kill them**. The theories that survive our most savage attempts at refutation aren't *proven* — they're just the ones still standing, "corroborated," provisionally trusted until the next test. Knowledge grows not by accumulating confirmations but by surviving executions.

The *demarcation criterion* — the line between science and pseudoscience — falls out cleanly. A claim is scientific to the degree that it *sticks its neck out*: that it forbids something, makes a risky prediction, tells you in advance what would prove it wrong. "The economy is governed by class struggle" forbids nothing. "Light bends

by 1.75 arcseconds" forbids 1.74 and 1.76. One is science; one is a worldview wearing a lab coat.

BE FAIR TO FREUD

It's a clean story, and Popper told it beautifully — perhaps too beautifully. Later philosophers (notably Adolf Grünbaum in 1984) argued Popper *caricatured* psychoanalysis: Freud did sometimes specify what would refute him ("my theory can only be refuted when phobias are shown to exist where sexual life is entirely normal"). And plenty of respectable science — historical, evolutionary, cosmological — can't run controlled experiments either. Falsifiability is a brilliant searchlight. We'll spend the rest of the day watching it flicker at the edges.

THE COMPLICATION

Kuhn: but that's not how science actually behaves

Popper described how science *ought* to work. In 1962, a physicist-turned-historian named **Thomas Kuhn** looked at how it has *really* worked — and found something messier and more human. His book *The Structure of Scientific Revolutions* became one of the most cited academic works of the twentieth century, and it gave us a word you've used a hundred times without knowing its origin: *paradigm*.

Here's Kuhn's heresy. Real working scientists, almost all the time, are *not* trying to falsify their grand theories. They're doing what he called *normal science*: puzzle-solving inside an accepted framework — a paradigm — that they take entirely for granted. A chemist doesn't wake up trying to refute the periodic table; she uses it to figure out a reaction. The paradigm isn't on trial. It's the courtroom.

And when an experiment comes back wrong? Scientists mostly *don't* drop the theory, the way Popper's story says they should. They shrug it off as an *anomaly* — a puzzle for later, probably their own mistake. The theory is too useful, too

productive, to abandon over one stubborn data point. (Notice that this is the *opposite* of falsificationism — and it's also, awkwardly, what those Freudians and Marxists were doing.)

Only when anomalies *pile up* — when they become too numerous and too central to ignore — does the field slide into *crisis*. And crisis is resolved not by a tidy refutation but by a **scientific revolution**: a wholesale *switch* to a new paradigm. Ptolemy's circles give way to Kepler's ellipses; Newton's absolute space gives way to Einstein's spacetime. Kuhn argued these shifts are so total that the two paradigms become *incommensurable* — there's "no common measure," because the rival camps don't even agree on what the key terms mean or which problems matter. "Mass" means something subtly different to Newton and to Einstein. A paradigm shift is less like winning an argument and more like a *gestalt flip* — the duck becomes the rabbit, and you can't see it both ways at once.

A MYTH WORTH KILLING

Kuhn is often waved around as proof that "science is just opinion" or "all paradigms are equally valid." He *hated* that reading and spent years pushing back on it. His point wasn't that science is irrational — it's that scientific rationality is more *communal*, *historical*, and *conservative* than the clean falsificationist fairy tale admits. Paradigms get overthrown because rivals genuinely solve more puzzles. That's not relativism. It's just realism about how humans do the work.

THE REPAIR

Lakatos: theories don't die alone — and the Duhem–Quine ghost

So Popper says *falsify*; Kuhn says *scientists don't, and shouldn't be too hasty*. Was there a way to honor both — to keep falsification's spine while admitting Kuhn's history? **Imre Lakatos**, a Hungarian émigré at the London School of Economics,

tried to build exactly that bridge. But first we have to meet the ghost haunting the whole room.

It's called the *Duhem–Quine thesis*, and once you see it you can't unsee it. The claim is simple and devastating: **no hypothesis is ever tested alone.** When you test "this star sits *there*," you're also relying on optics, atmospheric models, the telescope's calibration, the theory of how light travels. So when the prediction fails, pure logic *never* tells you which link broke. Maybe the hypothesis is wrong — or maybe your telescope was miscalibrated. You can *always* save your pet theory by blaming an auxiliary assumption instead. Popper's clean "single black swan kills the theory" turns out to be never quite that clean: you can insist the swan was a painted goose.

This isn't armchair pedantry — it's the engine of real discovery. When Uranus wobbled off its predicted Newtonian orbit in the 1840s, nobody declared Newton refuted. They blamed an auxiliary: there must be an *unseen planet* tugging on it. They were right — that's how **Neptune** was found in 1846, a glorious vindication. Emboldened, astronomers used the same move on Mercury's wobble, predicting another hidden planet they named **Vulcan**. They hunted it for decades. It does not exist. Mercury's wobble was telling them Newton himself was incomplete — and only Einstein, in 1915, could say so. *Same logical move, opposite outcomes.* So how do you tell a brilliant rescue from a desperate dodge?

Lakatos's answer reframes the unit of science. Don't judge lone theories — judge *research programmes* unfolding over time. Each has a **hard core** (the central commitments you protect by decision — "Newton's laws hold") wrapped in a *protective belt* of adjustable auxiliary hypotheses. When trouble comes, you absorb the hit in the belt, not the core. That's allowed. The question is what happens *next*:

- A **progressive** programme's patches *predict surprising new facts* that then turn up. "There's a hidden planet" predicted Neptune at a specific spot in the sky — and there it was. The rescue *paid for itself* with new knowledge.
- A **degenerating** programme only ever patches *after the fact*, bolting on excuses to explain away each failure while predicting nothing new. Vulcan, endlessly relocated to wherever it conveniently couldn't be seen, was the warning sign.

That's the demarcation line redrawn — and it's a far better fit for real history. Science isn't a single theory facing a single verdict; it's a *programme* earning or losing its keep over years, measured by whether it keeps telling us things we didn't already know.

THE WRECKING BALL

Feyerabend and the death of "the" method

Then Lakatos's friend and sparring partner **Paul Feyerabend** took the whole project out behind the barn. In *Against Method* (1975), he made a mischievous, maddening, and weirdly well-evidenced argument: comb through the actual history of great scientific breakthroughs, and you'll find that *every* proposed rule of method was **broken** by somebody, at some crucial moment, in order to make progress. Galileo advanced the Copernican cause with propaganda, rhetorical tricks, and by ignoring inconvenient data. Had he obeyed the tidy rules of method, the revolution might have stalled.

His conclusion became the most infamous two words in the philosophy of science: *"anything goes."* But here's the catch nearly everyone misses — Feyerabend did *not* mean "do whatever you like, all ideas are equal." He meant it as a bitter *reductio*: the only methodological rule with no historical counterexamples is one so empty it permits everything. It was, in his words, the "terrified exclamation" of a rationalist who finally looks honestly at history. He was burning down the idea that there is one capital-M Method that defines science for all time — not endorsing chaos.

And in 1983, the philosopher **Larry Laudan** delivered what looked like the funeral oration. In a famous essay, "The Demise of the Demarcation Problem," he argued that *every* attempt to draw a clean line — Popper's included — had failed, and that "science" and "pseudoscience" are too varied to share a single defining mark. The terms, he wrote acidly, are mostly "hollow phrases which do only emotive work for us." After two and a half millennia, the demarcation problem was pronounced dead.

THE RESURRECTION

Why the line still matters

Except — corpses this useful don't stay buried. In 2013, philosophers **Massimo Pigliucci and Maarten Boudry** edited a volume bluntly titled *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem*, reviving the whole question against Laudan. Their argument is partly practical and hard to wave away: in a world of vaccine refusal, climate denial, miracle cures, and intelligent-design "theory," telling science from its imitations is not an idle parlor game. It has a body count.

Their philosophical move is to stop demanding a *single* magic criterion and instead treat science as a *family-resemblance concept* — borrowing from Wittgenstein. There's no one feature every science shares and every pseudoscience lacks. Instead there's a *cluster*: falsifiable predictions, yes, but also empirical track record, openness to correction, coherence with established knowledge, honest treatment of anomalies, and the absence of the tell-tale dodges (endless ad-hoc rescue, persecution narratives, immunity to evidence). No single thread holds the rope; the threads overlapping do. A real science can be weak on one criterion and strong on the rest. A pseudoscience reveals itself by failing the whole pattern at once.

Which sets up the punchline of the entire day. All of this — Popper, Kuhn, Lakatos, the cluster of virtues — has been *philosophy*, argued in seminar rooms. But in the last fifteen years, science did something extraordinary: it turned the demarcation question on *itself*, empirically, at scale. It asked whether its own published findings could survive the most basic scientific demand of all.

The Demarcation Lab

CLAIM	POPPER	KUHN	LAKATOS	CLUSTER VIEW
Starlight bends by 1.75 arcseconds	Science	Science	Progressive	Strong scientific profile
Mercury retrograde disrupts communication	Not science	Not mature science	Degenerating	Weak profile
Class struggle drives history	Often unfalsifiable as used	It depends	Can degenerate	Mixed social science and philosophy
String theory	Not yet testable in key forms	Normal science without decisive tests	Open question	Live border case
Common descent	Falsifiable	Central biological paradigm	Progressive	Strong scientific profile

THE FRONTIER · 2026

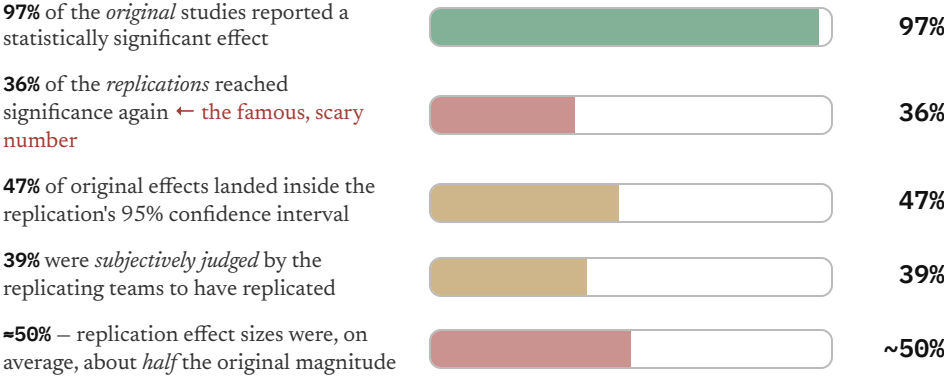
The replication crisis: demarcation under live fire

If there's one criterion almost everyone agrees on — Popper, Kuhn, your high-school teacher — it's **reproducibility**. A real result happens again when someone else repeats the procedure. It isn't a fluke, a fudge, or a fashion. So in the 2010s, scientists did the obvious, terrifying thing nobody had done systematically: they took piles of published, peer-reviewed, celebrated findings and simply *tried to make them happen again.*

Result 01 [ESTABLISHED] [CONTESTED]

The shot heard round psychology

The landmark is the **Open Science Collaboration's "Estimating the Reproducibility of Psychological Science"** (*Science*, 28 August 2015) — roughly 270 researchers, led by Brian Nosek, who repeated **100** studies from three top psychology journals, working with the original authors to get the methods right. The result detonated across the field. But the single most important lesson is buried in plain sight: **there is no one "replication rate."** The paper reported several, and they tell different stories. Watch.



Whenever you see "only a third of psychology is real," someone has grabbed the *36%* and dropped the rest. The honest summary is subtler and more interesting: replication effects were **weaker on average** — roughly half as strong as first reported, and often too weak for an underpowered repeat to catch. [ESTABLISHED] for the numbers themselves; [CONTESTED] for how far they license claims about which original effects were real.

And the authors refused to let anyone — optimist or doom-monger — over-read it. Their own conclusion is a small masterpiece of calibration, and a direct callback to **Day 1**'s lesson that a true belief held for the wrong reasons isn't knowledge:

> *How many of the effects have we established are true? Zero. And how many of the effects have we established are false? Zero.*
>
> — Open Science Collaboration, Science (2015)

A single failed replication, remember the Duhem–Quine ghost, doesn't *logically* refute the original — conditions always differ. Which is exactly why the critics pounced. **Gilbert, King, Pettigrew & Wilson** (*Science*, March 2016) argued the project's own replications were statistically underpowered and that, corrected, "the data are consistent with the opposite conclusion" — that reproducibility is high. The original team replied that *neither* rosy nor grim readings were yet warranted. [CONTESTED] — the *interpretation* is genuinely live, even though the broad problem is now widely accepted as real.

Result 02 [ESTABLISHED]

It isn't one field's embarrassment

The reflex defense — "soft sciences, what do you expect" — collapsed as the same exercise ran elsewhere and came back in the same unhappy range. The crisis is broad. Here are the verified anchor numbers; note the metric every time, because, as we just saw, the metric *is* the story.

PROJECT & VENUE	WHAT WAS REPEATED	REPLICATED*	EFFECT-SIZE SHRINKAGE
Psychology OSC, *Science* 2015	100 studies, 3 top journals	**36%**	to ~50% of original
Cancer biology Errington et al., *eLife* 2021	Planned 193 experiments — only ~50 could even be *attempted*	**~46%†**	~85% smaller
Experimental economics Camerer et al., *Science* 2016	18 lab experiments (AER, QJE)	**61%**	to ~66% of original
Social science Camerer et al., *Nat. Hum. Behav.* 2018	21 experiments in *Nature* & *Science*	**62%**	to ~50% of original
Preclinical oncology Begley & Ellis, *Nature* 2012	53 "landmark" papers (Amgen)	**11%**	— (6 of 53 confirmed)

*"Replicated" = significant effect in the same direction, the strictest common metric. †Cancer-biology figure is among experiments that could be completed; strikingly, **not one** of the 193 original experiments could be repeated from its published methods alone, and raw data was available for only 2%. [ESTABLISHED]

The deepest signal isn't even the failure rate — it's that *cancer-biology team's* discovery that they couldn't **find out what the original scientists had actually done.** Methods sections were too thin to follow; original authors often wouldn't share protocols or data. A finding you can't even *attempt* to reproduce hasn't failed Popper's test — it has refused to take it. And a backdrop survey makes the unease concrete: when *Nature* polled **1,576 scientists** in 2016, more than **70%** said they'd tried and failed to reproduce *someone else's* experiment, and more than **half** had failed to reproduce *their own*. [ESTABLISHED] — though note this is opinion data, what scientists *believe*, not a measured rate.

Result 03 [ESTABLISHED] [CONTESTED]

The findings that evaporated — and the scientists who said so

Abstractions don't sting; named casualties do. A run of celebrated, TED-talk-famous effects buckled under high-powered, preregistered repetition — and, remarkably, in the cleanest cases an *insider* changed their mind in public:

- **Power posing.** The 2010 finding that standing like Wonder Woman for two minutes raises testosterone and risk appetite (a TED talk seen tens of millions of times) failed a much larger 2015 replication on every physiological measure. Then the original first author, **Dana Carney**, did something rare and honorable — she publicly disowned her own most famous result: *"I do not believe that 'power pose' effects are real."* [ESTABLISHED]
- **Ego depletion.** The dominant theory that willpower is a finite fuel that drains with use was tested across **23 labs** (N = 2,141, 2016). The combined effect was statistically indistinguishable from *zero* (d = 0.04). A leading researcher in the area, Michael Inzlicht, wrote that he felt "the ground is moving from underneath me." [ESTABLISHED] that the standard effect didn't replicate; whether some small effect survives is still argued.
- **Social priming.** The classic claim that reading words about old age makes you walk more slowly out of the lab failed independent replication in 2012. It rattled the field so badly that Nobel laureate **Daniel Kahneman** sent an open letter warning priming researchers their field had become "the poster child for doubts about the integrity of psychological research." [ESTABLISHED] for the specific failures.
- **The Stanford Prison Experiment** (1971) — perhaps the most famous "study" in all of psychology — was shown by archival work (Le Texier, *American Psychologist*, 2019) to have been closer to *staged theater*: guards were coached toward cruelty, and results were sensationalized. It's less a failed replication than a demarcation casualty — a demonstration that may never have been an

experiment at all. [CONTESTED] — Zimbardo disputed the critiques before his death; whether to strike it from the textbooks is still fought over.

The turn [OPTIMISTIC]

Is this science failing — or science working?

Here's the reframe that makes the whole crisis a hopeful story rather than a scandal. Every one of those numbers came from *scientists policing science* — using preregistered, high-powered, openly-shared methods to expose and discard claims that couldn't stand up. That is **Popper's executioner's blade, finally turned inward.** The crisis isn't evidence that the demarcation criteria are wrong. It's evidence of them *working*, painfully and in public.

And it triggered real reform. *Preregistration* — stating your hypothesis and analysis *before* seeing the data — slams the door on the quiet fudging (p-hacking) that inflated all those effects; **Registered Reports**, where journals accept a study based on its *method* before any results exist, are now offered by 300+ journals. There are proposals to tighten the threshold for "significant" from $p < 0.05$ to $p < 0.005$, and a now-routine culture of open data and many-lab consortia. The field looked into Hume's hole, saw how easily luck and bias counterfeit knowledge — exactly the **Day 1** Gettier worry, now at industrial scale — and started rebuilding its instruments. We'll meet this reform movement again, in full, on **Day 149**.

OPEN QUESTIONS

What's genuinely unsettled

Two and a half thousand years in, the honest answer to "what makes something science?" still has loose ends:

- **Is there any single demarcation criterion at all** — or did Laudan win, leaving only a Wittgensteinian family of overlapping virtues with no master rule?

- **How much can the Duhem–Quine problem be tamed?** If a failed test never logically convicts the hypothesis, how do high-powered, preregistered replications actually shrink the wiggle room — and can they ever close it?
- **What about sciences that can't run experiments at all** — cosmology, evolutionary biology, string theory? If a theory makes no testable prediction for a generation (**Day 48**'s quantum-gravity problem looms), is it science, proto-science, or math?
- **Where's the floor on reproducibility?** A 62% replication rate across social science — is that a disgrace, a reasonable rate for hard questions about messy humans, or unknowable without agreeing what "replicated" even means?
- **And the question that will stalk this whole course:** if even peer-reviewed, celebrated findings are inflated by half, how should *you* — reading any confident claim, including the ones on these pages — set your credence? (Bring the dial. **Day 4**, **Day 6**.)

◆ THE DAY IN THREE SENTENCES

BIG IDEA

Hume showed you can never *prove* a general law by piling up confirmations, so science advances instead by making bold, falsifiable conjectures and trying to *kill* them — but real science is messier than that clean rule (Kuhn, Lakatos, Feyerabend), and the modern replication crisis is that whole debate finally tested with hard numbers.

BEST ANALOGY

The black swan: a million white swans can't prove "all swans are white," but one black swan in Australia disproves it forever — verification is hopeless, falsification is decisive.

LIVE CONTROVERSY

Whether any single line divides science from pseudoscience (Popper's falsifiability vs Laudan's "demise"), and what the replication numbers *mean* — a scandal of broken science, or the healthy, public self-correction of science working as designed.

THREADS TODAY › information (replication as the test of whether a claim carries real signal or noise) · evolution (Popper saw knowledge growing by selection — conjectures that survive refutation, a quiet preview of Day 74) · computation & emergence (lightly — science as a distributed, self-correcting error-finding system larger than any one mind).

SOURCES

Sources & further reading

1. Hume, D. (1739–40). *A Treatise of Human Nature*, Book I, Part iii. And (1748) *An Enquiry Concerning Human Understanding*, §IV–V. — the problem of induction; the sunrise passage. See Stanford Encyclopedia of Philosophy, "The Problem of Induction" (rev. 2018).

2. Popper, K. (1959). *The Logic of Scientific Discovery* (orig. *Logik der Forschung*, 1934). And (1963) *Conjectures and Refutations: The Growth of Scientific Knowledge.* Routledge. — falsifiability; Einstein vs Freud/Adler/Marx. See SEP, "Karl Popper".

3. Kuhn, T. S. (1962; 2nd ed. 1970). *The Structure of Scientific Revolutions.* University of Chicago Press. — normal science, paradigms, anomaly, crisis, revolution, incommensurability. See SEP, "Thomas Kuhn".

4. Lakatos, I. (1970). "Falsification and the Methodology of Scientific Research Programmes," in Lakatos & Musgrave (eds.), *Criticism and the Growth of Knowledge.* Collected in *Philosophical Papers, Vol. 1* (Cambridge UP, 1978). — hard core, protective belt, progressive vs degenerating programmes.

5. Feyerabend, P. (1975). *Against Method: Outline of an Anarchistic Theory of Knowledge.* New Left Books. — epistemological anarchism; "anything goes" (as reductio). See SEP, "Paul Feyerabend".

6. Duhem, P. (1906). *The Aim and Structure of Physical Theory.* And Quine, W. V. O. (1951). "Two Dogmas of Empiricism," *The Philosophical Review* 60(1): 20–43. — underdetermination / confirmation holism. See SEP, "Underdetermination of Scientific Theory".

7. Laudan, L. (1983). "The Demise of the Demarcation Problem," in Cohen & Laudan (eds.), *Physics, Philosophy and Psychoanalysis.* Reidel, pp. 111–127.

8. Pigliucci, M. & Boudry, M. (eds.) (2013). *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem.* University of Chicago Press. press.uchicago.edu — the revival; science as a family-resemblance / cluster concept.

9. Open Science Collaboration (2015). "Estimating the reproducibility of psychological science." *Science* 349(6251): aac4716. doi:10.1126/science.aac4716. science.org — 97% / 36% / 47% / 39% / ~50%.

10. Gilbert, D. T., King, G., Pettigrew, S. & Wilson, T. D. (2016). "Comment on 'Estimating the reproducibility of psychological science.'" *Science* 351(6277): 1037. — the critique; OSC

reply (Anderson et al., same issue).

11. Errington, T. M. et al. (2021). "Investigating the replicability of preclinical cancer biology." *eLife* 10: e71601 (Reproducibility Project: Cancer Biology). — ~50 of 193 experiments attempted; effects ~85% smaller; methods/data largely unavailable.

12. Camerer, C. F. et al. (2016). "Evaluating replicability of laboratory experiments in economics." *Science* 351(6280): 1433–1436. doi:10.1126/science.aaf0918 — 11 of 18 (61%).

13. Camerer, C. F. et al. (2018). "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour* 2: 637–644. — 13 of 21 (62%).

14. Klein, R. A. et al. (2018). "Many Labs 2: Investigating variation in replicability across samples and settings." *Advances in Methods and Practices in Psychological Science* 1(4): 443–490. — 15 of 28 (54%); setting didn't explain failure.

15. Begley, C. G. & Ellis, L. M. (2012). "Raise standards for preclinical cancer research." *Nature* 483: 531–533. doi:10.1038/483531a — 6 of 53 (11%) landmark papers confirmed (Amgen).

16. Baker, M. (2016). "1,500 scientists lift the lid on reproducibility." *Nature* 533: 452–454. doi:10.1038/533452a — >70% failed to reproduce others'; >50% their own.

17. Hagger, M. S. et al. (2016). "A multilab preregistered replication of the ego-depletion effect." *Perspectives on Psychological Science* 11(4): 546–573. — 23 labs; d = 0.04.

18. Ranehill, E. et al. (2015). "Assessing the robustness of power posing." *Psychological Science* 26(5): 653–656. And Carney, D. R. (2016), public statement disavowing power-posing effects. See overview.

19. Le Texier, T. (2019). "Debunking the Stanford Prison Experiment." *American Psychologist* 74(7): 823–839. doi:10.1037/amp0000401. pubmed

20. Ioannidis, J. P. A. (2005). "Why most published research findings are false." *PLoS Medicine* 2(8): e124. — the foundational (and model-based, thus contested-in-detail) paper.

21. Benjamin, D. J. et al. (2018). "Redefine statistical significance." *Nature Human Behaviour* 2: 6–10. doi:10.1038/s41562-017-0189-z — the $p < 0.005$ proposal (and Amrhein & Greenland's "remove, rather than redefine" rejoinder).

22. Chambers, C. D. (2013). "Registered Reports: A new publishing initiative at Cortex." *Cortex* 49(3): 609–610. And Chambers & Tzavella (2022), *Nature Human Behaviour* 6: 29–42 —

now in 300+ journals.

TOMORROW → DAY 03

Logic & Valid Inference

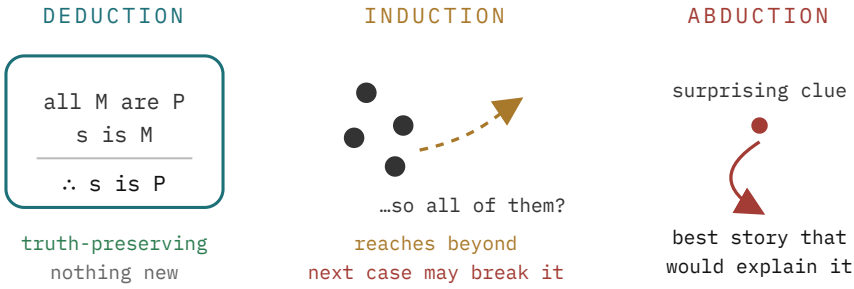
Today we leaned hard on words like "valid," "follows from," and "contradiction" — but what *are* the rules that make an argument actually hold together? Tomorrow we descend into logic itself: deduction (truth-preserving but never new), induction (Hume's wounded bird), and abduction (the detective's leap to the best explanation). We'll meet the everyday fallacies that fool us, ask whether logic is *discovered* or *invented*, and reach the frontier where machines now check proofs no human can fully hold in their head. The scaffolding under everything we've built so far.

END OF DAY 02 · 178 DESCENTS REMAIN

BLOCK I · FOUNDATIONS OF KNOWLEDGE & REASONING · DAY 003 / 180

Logic & Valid Inference

Three ways to move from what you know to what you don't — and only one of them is safe.



three engines of inference — guarantee shrinks left → right

Deduction keeps you safe but locked in the room. Induction and abduction get you out — at the price of certainty.

A stranger walks into a London consulting room. Within seconds Sherlock Holmes announces, to Watson's astonishment, that the man is a retired army doctor, recently invalided home from Afghanistan. The skin is tanned but the wrists are pale — a sunburn caught abroad, not at the seaside. He holds his arm stiffly, war-wounded. The face is haggard with hardship and fever. Holmes calls this *deduction*, and the word has clung to him for over a century. He is wrong about the word. What Holmes performs — and what made him immortal — is not deduction at all. It is a humbler, riskier, far more creative thing.

That mislabeling is the perfect way in, because the whole of today turns on a distinction almost everyone blurs: there is more than one way to reason, and they do not come with the same guarantees. Some inferences are **airtight** — if you grant the premises, the conclusion cannot escape. Others are **fertile but fallible** — they reach past the evidence and can be overturned by tomorrow's surprise. Mistaking one for the other is the root of a startling fraction of human error. So let's draw the lines carefully.

Two days in, we have circled reasoning from the outside. On **Day 1** we asked what turns a true belief into *knowledge* — and hit the Agrippan trilemma, the worry that every justification either regresses forever, loops in a circle, or stops somewhere arbitrary. On **Day 2**, Hume's *problem of induction* showed that no pile of observations can ever *prove* a universal law, which is why Popper told us to falsify rather than verify. Today we open the engine itself. Those two earlier puzzles were really about the limits of two specific *inference modes*; now we name all three, watch logic become mathematics, and follow it to the strangest frontier in the course — machines that check proofs with zero tolerance for error. The thread that lights up brightest today is *computation*.

Paget's Holmes became the public face of "deduction," even though the famous diagnostic leaps are usually abductive: clues first, best explanation second.

THE MODEL

Three engines, three guarantees

If you remember one thing from today, make it this trichotomy. Reasoning is not one activity but three, and they are sorted by how much they promise.

Deduction is the truth-preserving engine. The conclusion is already folded inside the premises; valid deduction merely unfolds it. Grant that all men are mortal and that Socrates is a man, and you *cannot* avoid the conclusion that Socrates is mortal — to deny it is to contradict yourself. The price of this safety is that deduction is *non-ampliative*: it never tells you anything genuinely new about the world. It rearranges what you already have. Mathematics is the deductive art carried to its limit, which is exactly why mathematicians can be so certain and why their certainty never, by itself, settles a question about *this* universe.

Induction is the generalizing engine. You have seen the sun rise ten thousand times; you infer it will rise tomorrow. Every swan anyone had ever logged was white, so — until 1697 — "all swans are white" looked secure. Induction is *ampliative*: it adds content, reaching beyond the cases in hand. And for precisely that reason it is **not truth-preserving**. This is Hume's bomb from **Day 2**, still ticking: no finite run of observations can logically guarantee the next one. Induction is how empirical knowledge actually grows, and it comes with no warranty.

Abduction is the explaining engine — and the one most people were never taught to name. You meet a *surprising fact*, and you cast around for a hypothesis that, if true, would make the surprise dissolve. The American polymath *Charles Sanders Peirce* (1839–1914) singled it out as the only genuinely creative mode, the one that *generates* new ideas rather than just testing or unpacking them. "Every plank of [science's] advance," he wrote, "is first laid by retroduction alone." Deduction and induction work over hypotheses you already have; abduction is where the hypotheses come from in the first place.

Now back to Holmes. The tan, the stiff arm, the haggard face — these are surprising facts, and Holmes leaps to the explanation that best accounts for all of them at once: a war-wounded military doctor home from a hot campaign. But notice the leap is not *guaranteed*. The man could be an actor who summers in Morocco and sprained his shoulder playing tennis. Holmes's conclusion is the *best* explanation, not the *only* one — which is the signature of abduction, not deduction. Conan Doyle gave his detective the wrong word, and a century of readers inherited the mistake. (This will matter again on **Day 4**, when we ask how to make "best explanation" precise using probability.)

THE SHAPE OF A GREAT MISNOMER

Holmes is not alone. We say a doctor "diagnoses" — that's abduction, reasoning from symptoms to the disease most likely to produce them. A mechanic listening to an engine, a detective at a crime scene, a scientist staring at an anomalous reading: all abducting, all leaping to the explanation that would render the strange ordinary. Even this sentence relies on it — you're inferring a mind behind these words because that's the best explanation for their orderly arrangement, not because a theorem forces it. Abduction is the water we swim in; we just rarely call it by name.

THE DISTINCTION EVERYONE FUMBLES

Valid is not the same as true

Inside the deductive engine lives the single most misunderstood idea in all of logic, and getting it straight is worth more than a dozen memorized fallacies. It is the difference between *validity* and *soundness*.

An argument is **valid** when its *form* guarantees that true premises would force a true conclusion. Validity is a property of the *shape*, not the content. The Internet Encyclopedia of Philosophy puts it cleanly: an argument is valid "if and only if it takes a form that makes it impossible for the premises to be true and the conclusion nevertheless to be false." Soundness asks for more — an argument is **sound** only if it is valid *and* all its premises are actually true.

Here is the part that trips people: **a valid argument can have a wildly false conclusion.** Watch.

> *All birds can fly. A penguin is a bird. Therefore, a penguin can fly.*

The *form* is flawless — "All M are P; s is M; therefore s is P," the very mould Socrates was poured into. If the premises were true, the conclusion would have to follow. So the argument is perfectly **valid**. It is also, obviously, **unsound**, because the first premise is false: not all birds fly. Validity certifies the plumbing; soundness asks whether you also pumped in clean water. A valid argument with a false premise is a beautifully engineered pipe carrying sewage.

This is not hair-splitting. It is the working principle behind *reductio ad absurdum*, one of the sharpest tools in mathematics: to prove a premise false, assume it, reason *validly* to a conclusion you know is false, and the falsehood flows backward to indict the premise. The whole technique depends on a valid argument deliberately producing a false conclusion. Validity is the carrier; truth is the cargo; learn to track them separately and a fog lifts from every argument you'll ever read.

WHEN THE FORM BREAKS

The two fallacies hiding in every "if"

If valid forms are the safe paths, fallacies are the trapdoors that look just like them. The most treacherous live in conditional reasoning — statements of the form "if *P*, then *Q*" — because the broken versions sit one letter away from the sound ones.

The two valid moves are old friends. *Modus ponens*: if P then Q; P is true; therefore Q. *Modus tollens*: if P then Q; Q is false; therefore P is false. Both are airtight. Now meet their evil twins.

Affirming the consequent runs: *if P then Q; Q is true; therefore P.* It grabs the wrong end. "If someone lives in San Diego, they live in California. Joe lives in California. Therefore Joe lives in San Diego." But California is large; Joe could be in Sacramento. The conclusion *might* be true, which is exactly what makes the fallacy so seductive — it sometimes lands the right answer for no good reason, and a true conclusion reached by a broken argument is the Gettier trap from **Day 1** wearing a logician's coat.

Denying the antecedent is its mirror: *if P then Q; P is false; therefore Q is false.* "If it's raining, the ground is wet. It isn't raining. Therefore the ground isn't wet." Sprinklers, dear reader. Burst pipes. A spilled bucket. Knocking out one cause does not knock out the effect, because effects can have more than one cause.

There's a teaching classic that makes the structure unforgettable: *If an animal is a dog, it has four legs. This animal has four legs. Therefore it is a dog.* Cats, horses, and tables object. The absurdity is the point — it's the same broken form as the San Diego argument, just with the silliness turned up so you can see the gears slip. (Eugène Ionesco built an entire scene of his play *Rhinoceros* on exactly this fallacy, a Logician gravely proving that a cat with four legs must be a dog.)

These are *formal* fallacies — broken shapes. Their cousins, the *informal* fallacies, are flaws not in form but in content: *post hoc ergo propter hoc* (the rooster crows, the sun rises, therefore the rooster summons the dawn), the ad hominem, the equivocation that quietly swaps a word's meaning mid-argument. Formal fallacies you catch by checking the skeleton; informal ones you catch by reading what the words actually do.

The Inference Inspector

FORM	PATTERN	VERDICT	WHY
Modus ponens	If P then Q; P; therefore Q	Valid	Affirming the sufficient condition forces the consequent.
Modus tollens	If P then Q; not-Q; therefore not-P	Valid	If Q must follow from P, the absence of Q rules P out.
Affirming the consequent	If P then Q; Q; therefore P	Invalid	Q may have other causes: Joe can live in California without living in San Diego.
Denying the antecedent	If P then Q; not-P; therefore not-Q	Invalid	Removing one sufficient cause does not remove every route to Q: sprinklers can wet the ground.

THE LINEAGE

How logic became mathematics

The machinery you've been using has a deep history, and it bends in a surprising direction: over twenty-three centuries, the study of *good argument* slowly turned into a branch of *algebra*. The story has four landmarks.

Aristotle (4th century BCE) built the first formal system in his *Prior Analytics*. His genius was to use letters as placeholders — "all *A* are *B*" — and so to study argument *forms* apart from their content. This is *term logic*: it relates terms like "man" and "mortal." Medieval logicians lovingly catalogued the valid syllogistic moods with mnemonic names — *Barbara*, *Celarent*, *Darii*. The names are codes, not people: their vowels mark proposition types, where *A* means "all S are P," *E* means "no S

are P," *I* means "some S are P," and *O* means "some S are not P." So *Barbara* is AAA, *Celarent* is EAE, and *Darii* is AII; *Barbara*, for example, means all M are P; all S are M; therefore all S are P. For nearly two thousand years, this *was* logic.

The Stoics, above all *Chrysippus* (c. 279–206 BCE), built a second, parallel logic that history nearly lost. Where Aristotle related *terms*, the Stoics related whole *propositions* with connectives we still use daily: if…then, and, or, not. Chrysippus laid out five "indemonstrables" — basic inference schemata, the first of which ("if the first then the second; but the first; therefore the second") is precisely *modus ponens*. This is *propositional logic*, the ancestor of the logic inside every computer chip. The Stoics arguably had a truth-functional grasp of the connectives — understanding "or" by when the whole is true given its parts — two millennia before it was rediscovered. The 20th-century logician Jan Łukasiewicz startled scholars by arguing Stoic logic was not Aristotle's poor cousin but "an achievement of equal rank." Then it was buried for ages while Aristotle reigned — a reminder that intellectual history is not a tidy relay race.

George Boole snapped the two traditions onto a new track. In *An Investigation of the Laws of Thought* (1854), he did something audacious: he treated logical reasoning as *calculation*. Let 1 be the universe and 0 be nothing; let multiplication be "and," addition be "or." Suddenly the laws of valid inference looked like the laws of algebra. "We ought no longer to associate Logic and Metaphysics," Boole declared, "but Logic and Mathematics." His book sold modestly and puzzled contemporaries. Only decades later, when Claude Shannon noticed in 1937 that Boole's two-valued algebra described electrical switching circuits exactly, did *Boolean algebra* become the literal foundation of digital logic. Every AND-gate in the device you're reading this on is a sentence of Chrysippus, rendered in silicon.

Gottlob Frege delivered the largest leap since Aristotle. His slim, forbidding *Begriffsschrift* ("concept-script," 1879) introduced the *quantifier* — the formal "for all" (∀) and "there exists" (∃) — and with it *predicate logic*. Aristotle's term logic choked on arguments like "every horse is an animal, therefore every head of a horse is the head of an animal"; Frege's machinery handled it and vastly more, analyzing propositions as functions fed with arguments. It is often called the finest single book in the history of symbolic logic. There's a tragic coda: Frege dreamed

of reducing all of arithmetic to pure logic, and just as the second volume went to press, a young Bertrand Russell sent him a letter containing a paradox — the set of all sets that don't contain themselves: does it contain itself or not? Either answer contradicts itself. Frege's grand foundation cracked. But his *logic* survived the wreck and became the modern symbolic logic we still teach. (The ghost of that paradox, and the limits it hinted at, will haunt us on **Day 28**, when Gödel proves no formal system can be everything mathematicians hoped.)

THE DEBATE

Is logic discovered or invented?

Here is a question that sounds like a parlor game and turns out to cut very deep. The bedrock laws — *identity* (A is A), *non-contradiction* (not both A and not-A), *excluded middle* (either A or not-A, no third option) — feel utterly inescapable. But where do they live? Are they features of *reality*, woven into the universe whether or not minds exist? Features of *thought*, the unavoidable grammar of any thinker? Or human *conventions*, real and binding but ultimately chosen, like the rules of chess?

Logical realism

DISCOVERED

The laws are objective, mind-independent structures of the world. We don't legislate non-contradiction any more than we legislate the prime numbers — we find it. Logic is read off reality.

Psychologism

LAWS OF THOUGHT

The laws describe how minds must operate — a branch of psychology. Frege and Husserl attacked this fiercely: logical truths are exact and a priori, while psychology is empirical and fuzzy.

Conventionalism

INVENTED

Revisability

EMPIRICAL?

The laws are stipulations we adopt because they're useful — binding once chosen, but not handed down by the cosmos. Curiously rare as a fully worked-out position, despite its close kinship to moral anti-realism.

Quine and Putnam floated the radical thought that even logic might be revised for *empirical* reasons — that quantum mechanics could push us toward a non-classical logic, much as relativity pushed us to non-Euclidean geometry.

That last box is the hinge of today's frontier. For most of history "the laws of thought" seemed untouchable — to question them was to saw off the branch you sat on. But the twentieth century produced rigorous, working *alternative* logics, systems that quietly drop one of the sacred laws and keep functioning. Once you've seen those alternatives do real labor, the grand metaphysical question softens into something more practical and, frankly, more interesting: not "which logic is *True*?" but "which logic is the right *tool* for this job?" Let's go meet the alternatives.

THE FRONTIER · 2026

Three live edges — and the hype filter

Every day in this course ends at the research frontier, with each claim tagged for how much weight it can bear. Logic's frontier is unusually concrete: it runs on real computers, checks real proofs, and has lately collided with artificial intelligence in ways that demand a careful eye.

Edge 01 [ESTABLISHED]

The logics that break the rules — on purpose

"Classical" logic is not the only consistent option; it's one settled point in a landscape of alternatives, each built by surrendering a law most people thought

non-negotiable.

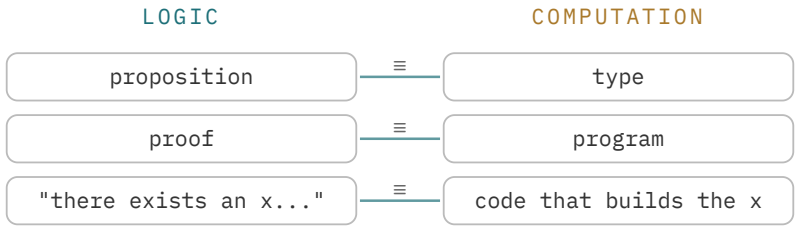
Intuitionistic logic drops the law of *excluded middle*. Pioneered by L.E.J. Brouwer and formalized by Arend Heyting in the 1920s–30s, it insists a statement counts as true only if you can *construct* a proof of it. You may not assert "A or not-A" for free — you must prove one side. The motivating example is sharp: excluded middle would let you cheerfully assert, for any computer program, "it halts or it doesn't" — yet (as we'll see on **Day 27**) no general method decides halting, so there's no construction backing the claim. Intuitionism says: then don't assert it. This sounds like philosophical fastidiousness until you learn where it leads — straight into the heart of computer science, via a correspondence so beautiful it gets its own box below.

Paraconsistent logic drops *explosion*. In classical logic a single contradiction is apocalyptic: from "P and not-P" you can derive *literally anything* (the principle *ex contradictione quodlibet*) — one inconsistency and the whole system goes up in flames. Paraconsistent logics refuse this, letting you reason sensibly even when some contradiction has crept in — useful for large databases, legal codes, or any messy body of information that's locally inconsistent but not therefore worthless. The stronger philosophical cousin, *dialetheism* — Graham Priest's view that some contradictions are *actually true*, like the Liar sentence "this sentence is false" — is far more controversial. Keep them separate: you can adopt a paraconsistent logic (a technical choice about explosion) without being a dialetheist (a metaphysical claim about true contradictions). The first is a tool; the second is a worldview.

Fuzzy logic drops the two-value restriction entirely. Lotfi Zadeh (1965) let truth slide along the whole interval from 0 to 1 to handle vagueness — "the water is warm" is 0.7 true — building on the many-valued logics of Łukasiewicz from the 1920s. It runs in control systems and appliances. And **modal logic** — the logic of *necessity* and *possibility* (□ and ◇) — together with carefully chosen temporal logics underpins the *formal verification* of hardware and software: specific fragments are expressive enough to say useful things while remaining decidable enough for model checking. These aren't museum pieces. They're the working logics of the modern technical world.

THE BRIDGE · PROPOSITIONS AS TYPES

The deepest reason intuitionistic logic matters is the **Curry–Howard correspondence**: in suitable formal systems, propositions correspond to types and proofs correspond to programs. Proving a theorem can be treated as constructing the program-like object that inhabits its type — and vice versa.



This is why several proof assistants below are built on type-theoretic foundations — and why *logic and computation*, one of our five threads, are not neighbors but the same country seen from two sides. (Picked up on **Days 27–29**.)

Edge 02 [ESTABLISHED]

Proof at zero tolerance: the rise of the proof assistant

Aristotle's dream was a chain of reasoning so tight that no one could doubt it. Twenty-three centuries later, that dream has a software implementation. A *proof assistant* is a program in which every logical step must pass a mechanical check; nothing is accepted on authority, intuition, or "clearly." The leading systems include **Lean** (now Lean 4), **Rocq** (the proof assistant formerly named Coq, renamed in 2025), **Agda**, and **Isabelle/HOL**. Lean, Rocq, and Agda live in the type-theoretic family; Isabelle/HOL is based on classical higher-order logic. Same ambition, different foundations.

Lean's community-built library, *mathlib*, is one of the largest unified formalizations of mathematics ever assembled: **more than 278,000 theorems and**

132,000 definitions when checked in June 2026, growing continuously, and covering 84 of the 100 problems on a famous "formalize these" challenge list. This is not a toy. Consider what it has already verified:

2022 · completed

The Liquid Tensor Experiment. In December 2020, Fields Medalist Peter Scholze challenged the world to verify a theorem from his "condensed mathematics" that he himself wasn't fully sure of. A team led by Johan Commelin and Adam Topaz did it in Lean, finishing on 14 July 2022. A working mathematician used a machine to gain *confidence* in a proof too intricate for comfortable human refereeing — exactly the point.

2023 · completed in 3 weeks

The Polynomial Freiman–Ruzsa conjecture. Days after Tim Gowers, Ben Green, Freddie Manners, and Terence Tao posted a proof of this additive-combinatorics result, Tao launched a Lean project to formalize it — and announced the dependency graph "completely covered in a lovely shade of green" three weeks later. Formalization keeping pace with research, nearly in real time.

2024–25 · completed

The Equational Theories Project. Tao's collaborative experiment (launched September 2024) to settle the implication relation among 4,694 algebraic laws — **22,033,636** ordered pairs if you include each law's trivial implication of itself, or **22,028,942** nontrivial graph edges — combining human proofs, automated provers, AI, and Lean verification across 50+ contributors. It finished in just over 200 days: a new model of massively collaborative, machine-checked mathematics.

2024–2029 · in progress

Fermat's Last Theorem. Kevin Buzzard's EPSRC-funded project (launched April 2024, Imperial College London) to formalize FLT — not the original Wiles proof but a modern route. Buzzard is "quietly confident" of reducing it to 1980s-known results, but frank that the whole thing is "at least a 5 year

> project." *Not yet done* — the honest status is a work in progress, the last of those 100 challenge problems still open.

And the certainty reaches beyond pure mathematics into systems lives depend on. **CompCert** is a C compiler *proved correct* in Rocq; a celebrated bug-hunting study spent roughly six CPU-years trying to make it emit wrong code and failed — "the only compiler we have tested for which Csmith cannot find wrong-code errors" — while finding the usual swarm of bugs in GCC and LLVM. **seL4** is the first operating-system microkernel with a full machine-checked proof of functional correctness (in Isabelle/HOL): under its stated assumptions, the C implementation refines the formal specification, so whole classes of crashes and unsafe behaviors are ruled out by theorem rather than hope. These are not ordinary promises; they are conditional theorems about software. *This* is what logic, mechanized, can do — and it is solidly **established**.

Edge 03 [ESTABLISHED] [CONTESTED/HYPE]

When AI met the proof checker

The newest and noisiest edge is the collision of machine learning with formal proof — and it is exactly where the hype filter earns its keep, because the headlines and the reality have drifted apart.

The genuine milestone first. In July 2024, DeepMind's **AlphaProof**, paired with AlphaGeometry 2, solved **4 of 6 problems** at the International Mathematical Olympiad, scoring 28 points — the top end of the silver-medal category, one point below the gold threshold of 29. It even cracked the fearsome Problem 6, which only 5 of roughly 600 human contestants fully solved. The methodology was published online in *Nature* on 12 November 2025, with the version of record appearing in 2026. Here's the design fact that separates it from chatbot bluster: **AlphaProof works inside Lean.** It auto-formalized about a million natural-language problems into ~80 million formal Lean statements, then trained itself in an AlphaZero-style loop where *Lean checks every step*. As DeepMind put it, there

are "no hallucinations to worry about" — because a hallucinated step simply fails to compile. The neural net supplies creative search; the proof assistant supplies ground truth. That marriage is real and important. `[ESTABLISHED]`

In July 2025 the bar rose again: both DeepMind (a Gemini "Deep Think" model) and OpenAI reported **gold-medal scores** — 5 of 6 problems, 35 points — and, strikingly, did it working *end-to-end in natural language* within the time limit, not in Lean. DeepMind's result was officially certified by the IMO; OpenAI's was graded internally. Genuinely impressive. But here is where you deploy the calibration instinct from **Day 1**:

- **"Gold medal" is a score, not a coronation.** These are competition problems — a narrow, time-boxed slice of mathematics with known-to-exist short answers. They are not open research questions, and per the official 2025 results, *26 human contestants still outscored both AI systems*.
- **Dropping Lean is a trade, not a free upgrade.** The 2024 silver was *formally verified* — guaranteed correct by machine. The 2025 natural-language gold was *human-graded*, which means we're back to trusting prose that could harbor a subtle gap. More general, less certain. Don't let "gold beats silver" hide that the epistemic ground shifted.
- **It is expensive and narrow.** Each hard 2024 problem took two to three days of computation, and problems were hand-translated into Lean for the competition. This is not a general mathematical mind.

And the claim to retire most firmly: **AI has not "solved mathematics" or made mathematicians obsolete.** `[CONTESTED/HYPE]` No AI has independently proven a famous open conjecture and had it accepted as a landmark. Reports of theorem-proving agents finding small Lean proofs or helping close narrow formalization tasks are intriguing, but they are early, scoped, and not yet a substitute for accepted research mathematics — the kind of thing to file under `[PROMISING]` and revisit, not to trumpet. The real revolution is quieter and more durable than the headlines: a 2,300-year-old standard — *a proof is a chain no one can doubt* — has finally been handed to a machine that enforces it without mercy, and AI is learning to search within those unforgiving rails. (A theme we'll chase properly across **Days 138–145**.)

A NOTE ON FABRICATED SOURCES

This curriculum's hype filter includes a rule worth stating: discard any citation to a future-dated preprint identifier. Search results in this space are littered with confident-looking references to papers that don't exist yet. Every milestone above is traced to a real, dated, primary source — a published *Nature* paper, an official competition result, a named researcher's own announcement. When a claim about AI and mathematics can't be traced that way, the right response is not excitement but suspicion.

OPEN QUESTIONS

What's genuinely unsettled

Twenty-three centuries in, the study of valid inference still leaves real questions wide open:

- **Is there one true logic, or many?** Once intuitionistic, paraconsistent, and fuzzy logics all do useful work, "the correct logic" starts to look less like a fact about the universe and more like a choice of tool — but pluralists and monists are still genuinely at odds.
- **Discovered or invented?** Are the laws of logic read off reality, baked into any possible mind, or adopted by convention? And could empirical physics ever *force* a revision, as Putnam suspected?
- **What is abduction, exactly?** Is "inference to the best explanation" a real third mode, or dressed-up induction? Even whether Peirce *meant* it as inference-to-best-explanation (versus mere hypothesis-generation) is debated among his scholars.
- **Can mechanized proof change what mathematics *is*?** If a result is true but only a computer has checked the proof, has anyone *understood* it? Does a verified-but-opaque proof carry the same value as an illuminating human one?
- **And the question that will stalk the AI block:** when a machine outputs a true, well-supported theorem, does it *know* anything — or is it the ultimate Gettier

case from **Day 1**, right for reasons that have nothing to do with comprehension? (**Days 138–145**.)

◆ THE DAY IN THREE SENTENCES

BIG IDEA

Reasoning comes in three engines with three different warranties — *deduction* preserves truth but adds nothing, *induction* generalizes but can be broken by the next case, and *abduction* leaps to the best explanation — and inside deduction, validity (good form) is a wholly separate thing from soundness (good form plus true premises).

BEST ANALOGY

Sherlock Holmes's "deductions" are really abductions — the best explanation of the clues, not a guaranteed conclusion — and a valid-but-unsound argument is a beautifully built pipe carrying sewage.

LIVE CONTROVERSY

Whether logic is discovered or invented (and whether there's one true logic or a toolkit of them), now sharpened by a real frontier where proof assistants like Lean verify cutting-edge mathematics at zero tolerance and AI has reached medal-level — but emphatically has *not* "solved mathematics."

THREADS TODAY › computation (Curry–Howard: proofs correspond to programs; Boolean algebra in silicon; proof assistants) · information (formalization makes a proof's content machine-checkable) · emergence (massively collaborative proof settling about 22 million implication relations) — with deduction and induction tying back to Day 1 and Day 2.

TOMORROW → DAY 04

Probability as Extended Logic

Today the unreliable engine was induction, and abduction left us needing a way to say which explanation is *best*. Tomorrow we tame both with numbers. Probability turns out to be not a separate subject from logic but its natural extension to partial belief — and the Monty Hall problem will show how badly our intuitions misfire, and how Bayes' theorem sets them right. Bring today's distinction between airtight and merely-plausible inference; you're about to learn the calculus of the merely plausible.

SOURCES

Sources & further reading

1. "Validity and Soundness." *Internet Encyclopedia of Philosophy* (accessed 2026). iep.utm.edu/val-snd — the form-based definition of validity and the validity-vs-soundness distinction.

2. "Deductive and Inductive Arguments." *Internet Encyclopedia of Philosophy.* iep.utm.edu/ded-ind — truth-preserving vs ampliative inference.

3. Douven, I. "Abduction." *Stanford Encyclopedia of Philosophy* (rev. 2021). plato.stanford.edu/entries/abduction — Peirce, inference to the best explanation, and the scholarly debate over what abduction is.

4. "Aristotle's Logic." *Stanford Encyclopedia of Philosophy.* plato.stanford.edu/entries/aristotle-logic — the syllogistic, *Prior Analytics*, and term logic.

5. Bobzien, S. "Ancient Logic." *Stanford Encyclopedia of Philosophy.* plato.stanford.edu/entries/logic-ancient — Chrysippus, the Stoic indemonstrables, and propositional logic; Łukasiewicz's reassessment.

6. Boole, G. (1854). *An Investigation of the Laws of Thought.* London: Walton & Maberly. See "George Boole, The Laws of Thought," *PhilPapers.* philpapers.org/rec/BOOTLO-4 — logic as algebra; "Logic and Mathematics."

7. "Origins of Boolean Algebra in the Logic of Classes." *Mathematical Association of America (Convergence).* old.maa.org — Boole, Venn, Peirce, and the path to digital logic via Shannon (1937).

8. "Frege's Logic." *Stanford Encyclopedia of Philosophy.* plato.stanford.edu/entries/frege-logic — the *Begriffsschrift* (1879), quantifiers, predicate logic, and Russell's paradox.

9. "Intuitionistic Logic." *Stanford Encyclopedia of Philosophy.* plato.stanford.edu/entries/logic-intuitionistic — Brouwer, Heyting, the rejection of excluded middle, the BHK interpretation.

10. Priest, G., Berto, F. & Weber, Z. "Dialetheism" and "Paraconsistent Logic." *Stanford Encyclopedia of Philosophy.* plato.stanford.edu/entries/dialetheism — explosion, paraconsistency vs dialetheism, the Logic of Paradox.

11. "Fuzzy logic." *Wikipedia* (accessed 2026). en.wikipedia.org/wiki/Fuzzy_logic — Zadeh (1965), truth in [0,1], many-valued / Łukasiewicz roots.

12. Garson, J. "Modal Logic." *Stanford Encyclopedia of Philosophy.* plato.stanford.edu/entries/logic-modal — necessity/possibility and applications to computer science and verification.

13. "Curry–Howard correspondence." *Wikipedia* (accessed 2026). en.wikipedia.org/wiki/Curry–Howard_correspondence — propositions as types, proofs as programs.

14. "Mathlib statistics." *Lean community* (accessed June 2026). leanprover-community.github.io/mathlib_stats.html — current theorem and definition counts.

15. "100 theorems in Lean." *Lean community* (accessed June 2026). leanprover-community.github.io/100.html — 84 of Wiedijk's 100 theorem benchmarks formalized in Lean.

16. Commelin, J. & Topaz, A. et al. "Liquid Tensor Experiment." *Lean community blog* (completion 14 July 2022); Scholze's original challenge (Dec 2020). leanprover-community.github.io — machine-checking a Fields Medalist's uncertain proof.

17. Tao, T. "Formalizing the proof of PFR in Lean4." *terrytao.wordpress.com* (Nov 2023). Gowers, Green, Manners & Tao, "On a conjecture of Marton," *Annals of Mathematics* (2025). terrytao.wordpress.com

18. Tao, T. et al. "The Equational Theories Project." Project announced Sept 2024; retrospective paper Dec 2025 (arXiv:2512.07087). teorth.github.io/equational_theories – 22,033,636 ordered pairs including self-implications; 22,028,942 nontrivial graph edges; 50+ contributors, Lean-verified.

19. Buzzard, K. "Fermat's Last Theorem project." *Lean community blog* (launch 30 April 2024); EPSRC grant EP/Y022904/1 (2024–2029), Imperial College London. leanprover-community.github.io – in progress; "at least a 5 year project."

20. Leroy, X. et al. "CompCert" – a formally verified C compiler. Yang, Chen, Eide & Regehr, "Finding and Understanding Bugs in C Compilers," *PLDI* (2011). compcert.org – six CPU-years and no wrong-code bugs found.

21. Klein, G. et al. (2009). "seL4: Formal Verification of an OS Kernel." *SOSP '09.* sel4.systems – first machine-checked proof of OS-kernel functional correctness (Isabelle/HOL).

22. "AI achieves silver-medal standard solving International Mathematical Olympiad problems." *Google DeepMind blog* (25 July 2024). deepmind.google – AlphaProof + AlphaGeometry 2; 28 points; works in Lean.

23. Hubert, T., Mehta, R., Sartran, L. et al. (2026). "Olympiad-level formal mathematical reasoning with reinforcement learning." *Nature* 651: 607–613. doi:10.1038/s41586-025-09833-y. nature.com/articles/s41586-025-09833-y – the AlphaProof method paper; published online 12 Nov 2025, version of record 13 Mar 2026; ~80 million Lean problems.

24. "Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the IMO." *Google DeepMind blog* (July 2025). deepmind.google – 35/42 officially certified; natural-language proofs within the contest time limit.

25. "66th IMO 2025." *International Mathematical Olympiad*. imo-official.org/editions/2025 and individual results – 630 contestants; gold cutoff 35; human score distribution.

26. "Our First Proof submissions." *OpenAI* (2026). openai.com/index/first-proof-submissions – OpenAI's later summary of its July 2025 IMO gold-medal-level result, 35/42 points.

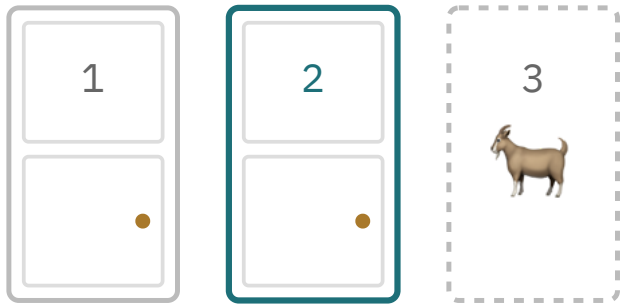
27. "Philosophy of logic" & "Logical realism." *Wikipedia / Stanford Encyclopedia of Philosophy* (accessed 2026). plato.stanford.edu/entries/logical-pluralism – realism, conventionalism, Quine/Putnam on revising logic, logical pluralism.

END OF DAY 03 · 177 DESCENTS REMAIN

BLOCK I · FOUNDATIONS OF KNOWLEDGE & REASONING · DAY 04 / 180

Probability as *Extended Logic*

A game show host opens a door. Your gut says it can't matter. Your gut is about to lose two out of three times.



● YOU PICKED 1 · HOST OPENED 3 · SHOULD YOU SWITCH TO 2?

The whole of Bayesian reasoning, hiding inside a 1970s game show.

You pick Door 1. Somewhere behind these three doors sits a car; behind the other two, goats. The host — who knows exactly where the car is — strolls over to Door 3, swings it open to reveal a goat, and asks, almost kindly: *would you like to switch to Door 2?* Two doors left, one car. Fifty-fifty, surely. Switching can't possibly matter.

It matters enormously. Stay, and you win the car one time in three. Switch, and you win **two times in three** — you double your odds by doing nothing but changing your mind. This is the Monty Hall problem, and when it ran in a magazine in 1990 it triggered one of the great public meltdowns in the history of mathematics. Today we'll see why the answer is not just correct but *inevitable* —

and how the same machine that solves it turns out to be the deepest available theory of what it means to reason under uncertainty at all.

On **Day 1** we met *credence* — belief as a dial from 0 to 1 — and the Dutch book argument showing that incoherent dials can be turned into a guaranteed loss. Today we learn the law that says how the dial must *move* when evidence arrives: Bayes' theorem. On **Day 2** we watched science struggle to draw the line between signal and noise, and saw the replication crisis as that struggle under live fire; today's frontier — a quiet revolution replacing the p-value with a *bet* — is aimed squarely at fixing it. Threads lit today: *information* (evidence as bits that update belief), *computation* (the mind and the lab as inference engines), and a flicker of *energy* when the "Bayesian brain" returns.

THE MELTDOWN

The smartest people in the country, all wrong at once

In September 1990, Marilyn vos Savant — listed in the *Guinness Book* for the highest recorded IQ, writing the "Ask Marilyn" column in *Parade* magazine — answered a reader's question about a game show. Switch doors, she wrote; you'll win two-thirds of the time. The answer is correct. The response was apocalyptic.

Monty Hall's actual stage makes the puzzle less like a parlor trick: the host was never a random door-opener, but a knowledgeable agent whose action carried information.

By her own count she received some **10,000 letters**, the overwhelming majority telling her she was wrong — and roughly **1,000 of them signed by people with PhDs**. Mathematicians wrote in to scold her. One professor offered the immortal line:

> *"You blew it, and you blew it big! … There is enough mathematical illiteracy in this country, and we don't need the world's highest IQ propagating more. Shame!"*
>
> — Scott Smith, Ph.D., University of Florida, in a letter to Parade (1990)

He was the one who'd blown it. So had, by the strict statistics of the thing, most of his colleagues. Vos Savant held her ground across three more columns, eventually asking schoolteachers across the country to run the experiment with paper cups

and a coin. They did. The data came back exactly as she'd said: switching wins twice as often. The professors, slowly and not always graciously, retreated.

THE MAN WHO NEEDED TO SEE IT TO BELIEVE IT

Even **Paul Erdős** — one of the most prolific mathematicians who ever lived, a man who proved theorems most of us can't even read — refused to accept the answer. When his friend Andrew Vázsonyi laid out the logic, Erdős was unconvinced. Only when Vázsonyi ran a *computer simulation*, playing the game hundreds of times and watching switching win about two-thirds of the rounds, did Erdős concede. And even then he was annoyed: the simulation showed him *that* it was true without showing him *why*. (Recounted in Paul Hoffman's biography *The Man Who Loved Only Numbers*, 1998.) If it tripped Erdős, you are in excellent company.

Here's the thing the meltdown reveals. The Monty Hall problem isn't a trick or a word game — its answer is provably, simulation-confirmably true. What it exposes is that human intuition about uncertainty is *systematically* miscalibrated, and that we badly need a formal tool to override it. That tool is the subject of today's descent. But first, let's actually break our intuition on the rocks — and then rebuild it.

The Monty Hall Machine

FIRST PICK	HOST ACTION	STAY	SWITCH
Car, probability 1/3	Opens either goat door	Win	Lose
Goat, probability 2/3	Forced to open the other goat door	Lose	Win

So staying keeps the original 1/3 chance; switching captures the 2/3 chance that the first choice was wrong.

WHY IT WORKS

The host is doing you a favor (and leaking information)

The cleanest way to feel the answer: **your first pick is right one time in three.** That number never changes. When you pointed at Door 1, there was a 1/3 chance the car was behind it and a 2/3 chance it was behind "one of the other two." The host then opens a goat door — but crucially, the host is *not* choosing at random. He knows where the car is, and he is *required* to reveal a goat. So all of that 2/3 probability, which used to be smeared across two doors, gets **concentrated onto the single door he didn't open.**

The host's reveal isn't noise. It's *information* — the first appearance of one of our five recurring threads in hard quantitative form. Day 1's stopped clock taught that being right by luck is not knowledge; here, the host's constrained, knowledgeable action is evidence that moves the credence dial. Switch, and you're betting on that fat 2/3. Stay, and you're clinging to your original lonely 1/3.

If your intuition still resists, blow the problem up. Imagine **a thousand doors.** You pick one — a 1-in-1,000 shot. The host, who knows, then opens 998 other doors, every single one a goat, leaving just your door and one other. Do you really still think it's a coin flip? Almost certainly the car is behind *that* one door the host so pointedly avoided. The three-door version is the same logic, merely too small to feel.

OLDER THAN THE GAME SHOW

The puzzle didn't start with Monty Hall. The statistician **Steve Selvin** posed it in a 1975 letter to *The American Statistician* — and his follow-up was the first place the phrase "the Monty Hall problem" ever appeared in print. Its skeleton is older still: it's identical to **Bertrand's box paradox** (Joseph Bertrand, 1889) and Martin Gardner's **Three Prisoners problem** (1959). Mathematicians call this a *veridical paradox* — an answer that looks impossible but is provably true. Convergent re-discovery again, exactly like the Gettier case on Day 1: when minds keep tripping over the same stone for a century, the stone is real.

THE MODEL

Bayes' theorem: the law of belief revision

What we did to those doors by hand has a name and a formula. It's the single most important equation in the theory of evidence, and it is almost insultingly simple to state:

$$P(H \mid E) = P(H) \times P(E \mid H) \,/\, P(E)$$

posterior (belief after evidence) = prior (belief before) × likelihood (how well H predicts E), normalized

In words: your *posterior* belief in a hypothesis H after seeing evidence E equals your *prior* belief, multiplied by the *likelihood* — how strongly H predicted that you'd see E — divided by how expected E was overall. Strong evidence is evidence your hypothesis predicts and rival hypotheses don't. That's the entire engine. Belief flows toward whatever best predicted what actually happened.

In Monty Hall, H = "the car is behind Door 2" and E = "the host opened Door 3." If the car really is behind Door 2, the host is *forced* to open Door 3 (he can't open your door or the car's), so the likelihood is 1. If the car is behind your Door 1, he could've opened either 2 or 3, so the likelihood of opening 3 is only 1/2. That asymmetry in the likelihoods is exactly what tips the posterior to 2/3 in favor of switching. The formula just does the bookkeeping our intuition botches.

The trap that fools doctors

Bayes' theorem doesn't only rescue game-show contestants. It catches a mistake that, in one famous study, most *physicians* got wrong. Play with it below — it's worth feeling this one in your bones, because it governs every medical test, spam filter, and airport screening you'll ever encounter.

The base-rate trap

GROUP	COUNT OUT OF 1,000	POSITIVE TESTS
Sick people	10	About 10 true positives
Healthy people	990	About 50 false positives
All positive tests	About 60	Only about 10 are truly sick

The posterior is therefore about 10 / 60, or 16.7%. If prevalence is 1 in 1,000, as in the Casscells study, the same kind of test yields a posterior of about 2%.

THE DEEP IDEA

Why "extended logic," not just a formula

Here's the claim that gives today its title. Ordinary deductive logic — the syllogisms of Day 3 — is the logic of *certainty*: if all men are mortal and Socrates is a man, then Socrates is mortal, full stop. But almost nothing in real life is certain. We need a logic for the vast middle ground between "definitely true" (probability 1) and "definitely false" (probability 0). The startling result is that there is essentially **only one** such logic, and it is the probability calculus.

This was made precise by the physicist **R. T. Cox** in 1946. Cox asked: suppose you want to attach a number to "how plausible is this, given what I know?" and you insist on just a few common-sense rules — plausibilities are real numbers; if you can compute a plausibility two valid ways you must get the same answer (*consistency*); and the plausibility of "not-A" should depend only on the plausibility of "A." From those bare desiderata, Cox proved, you are *forced* — not encouraged, forced — into the standard rules of probability. After a harmless rescaling,

negation has to behave like 1 - P(A), conjunction has to obey the product rule, and learning evidence E has to mean conditionalizing on E. Any consistent system of graded belief *is* probability theory in disguise.

The physicist **E. T. Jaynes** built his great posthumous book *Probability Theory: The Logic of Science* (2003) on exactly this foundation. His slogan: deductive logic is just the special case of probability theory where all the probabilities happen to be 0 or 1. Probability is logic *extended* to handle uncertainty — which is to say, extended to handle reality. Notice this is the *third* independent road to the same destination: the Dutch book argument (Day 1) got there from "don't be exploitable," and we'll see decision theory get there from "don't make dominated choices." Coherence, no-sure-loss, and consistent reasoning all point at one calculus.

THE HONEST FOOTNOTE

Cox's original proof was a touch too quick. In 1999 the computer scientist **Joseph Halpern** showed it needs an extra technical assumption to be airtight (it can fail on certain finite domains), and later authors patched it properly. So the right thing to say is not "probability is the *only conceivable* logic of uncertainty" — that overstates it — but "under reasonable conditions, consistent graded belief is forced into the probability axioms." The theorem stands; it just wears a slightly smaller crown than Jaynes's prose sometimes suggests. [ESTABLISHED]

THE DEBATE

Two tribes, one equation

If probability is this beautiful and unified, why has it been the site of a century-long civil war? Because the equation is agreed on; what's fought over is *what the numbers mean.* Both tribes use the very same calculus — the axioms **Andrey Kolmogorov** wrote down in 1933, which deliberately decline to say what probability *is* and only fix how it must behave. Onto that neutral skeleton, two interpretations are draped.

Frequentist

PROBABILITY = LONG-RUN FREQUENCY

- A probability is the **frequency of an event in infinitely many repetitions.** "The coin is fair" means it lands heads half the time over endless flips.
- Parameters are **fixed unknown constants**; the data are random. You reason about how often your *method* would mislead you.
- Tools: **p-values, confidence intervals, Type I/II error** (Fisher; Neyman & Pearson, 1920s–30s).
- Can't coherently say "70% chance there was life on Mars" — Mars either had life or it didn't; there's no repetition to count.

Bayesian

PROBABILITY = DEGREE OF BELIEF

- A probability is a **credence** — your rational degree of confidence given what you know (straight from Day 1's dial).
- Parameters get **probability distributions**; you update them with Bayes' theorem as data arrive.
- Tools: **priors, posteriors, Bayes factors.** Lineage: Laplace → Jeffreys → Ramsey → de Finetti → Savage.
- **Happily** says "70% chance of past life on Mars" — a one-off claim with no repetitions is exactly what credence is for.

Frequentism dominated the 20th century partly for a good reason and partly for an accident. The good reason: its founders craved *objectivity* and distrusted the Bayesian *prior* as a smuggled-in opinion. (Fisher dismissed "inverse probability" as something that "must be wholly rejected.") The accident: Bayesian methods need heavy computation, which didn't exist until cheap computers arrived. The central Bayesian sore point remains the prior — where does your "before" belief come from, and why should anyone trust yours? Objective Bayesians (Jeffreys, Jaynes) hunt for rule-based priors; subjective Bayesians shrug and say all reasoning starts somewhere.

"PROBABILITY DOES NOT EXIST"

The Italian Bruno de Finetti opened his treatise with those four words, in capitals. His point was deliberately provocative: there is no probability "out there" in the world like mass or charge — there is only the coherent betting behavior of a reasoning agent. He backed the slogan with a real theorem (his 1937 *representation theorem*): if you treat a sequence of observations as *exchangeable* — order doesn't matter to you — then you are mathematically obliged to act *as if* there's some fixed unknown frequency with a prior over it. Subjective belief and objective-looking parameters turn out to be two views of one structure. A truce, written in math.

And note the practical wisdom that falls out: **Cromwell's rule** (named by Dennis Lindley after Oliver Cromwell's 1650 plea, "think it possible that you may be mistaken"). Never set a prior to exactly 0 or 1, because Bayes' theorem can never budge it afterward — a belief held with absolute certainty is, by construction, unteachable. Leave a sliver of doubt for the moon being green cheese, Lindley wrote, or no returning astronaut's cheese samples will ever move you. Calibration, again — the through-line of this whole block.

THE FRONTIER · 2026

The quiet mutiny against the p-value

For a century, the frequentist p-value has been science's gatekeeper: get below 0.05 and you may call your result "significant." On Day 2 we saw the bill come due — the replication crisis, in which mountains of "significant" findings simply evaporated on re-testing. A big culprit is structural: the p-value is fragile. **Peek at your data midway and stop the moment you hit p<0.05, and you've quietly inflated your false-positive rate** — a sin so common it has a name, "optional stopping." A new framework now circulating through statistics rebuilds testing from the ground up to fix exactly this. Its central object isn't a probability. It's a *bet.*

Edge 01 [ESTABLISHED]

The e-value: test a hypothesis by betting against it

An *e-value* is the payoff of a bet against the null hypothesis. You wager $1 that the null is false, under a betting contract designed to be *fair if the null is true* — meaning that if the null really holds, you can't expect to grow your money (in symbols, the expected value of an e-value under the null is at most 1). So if you walk away having multiplied your stake twentyfold, something is off with the null: either it's false, or you got astronomically lucky. A large e-value is literally **money won against the null**, and your accumulated wealth *is* your evidence. The reciprocal 1/e behaves like a conservative p-value, but the betting picture is the point.

In the coin demo below, the null is concrete: **the coin is fair, P(heads) = 0.5.** The displayed e-value is the wealth from two likelihood-ratio tickets. One ticket bets on a heads-heavy coin, P(heads) = 0.60: a head multiplies that ticket by 0.60 / 0.50 = 1.2, while a tail multiplies it by 0.40 / 0.50 = 0.8. The mirror ticket bets on a tails-heavy coin, P(heads) = 0.40, with the multipliers reversed. The demo splits the starting $1 evenly between those two tickets, so either kind of sustained bias can make wealth grow. If the coin is actually fair, each ticket has expected multiplier 1 on every flip; the game is fair under the null. In this toy game, *winning* means your wealth gets large enough to reject "fair coin"; *losing* means the wealth stalls or shrinks, so you have not earned evidence against fairness.

This isn't loose metaphor; it's a rigorous program — "game-theoretic statistics," built over two decades by **Glenn Shafer** and **Vladimir Vovk** and now carried forward by **Aaditya Ramdas**, **Peter Grünwald**, **Ruodu Wang** and others. Shafer's manifesto, "Testing by Betting," was read before the Royal Statistical Society in 2020 and published in its *Journal* (Series A) in 2021. His complaint about the p-value is partly that it's *too confusing to communicate*; "I won $20 betting against this hypothesis" is something a human can actually grasp.

Edge 02 [ESTABLISHED] [CONTESTED]

Why a bet beats a p-value: you can peek all you want

Bets compound. If you make a fair bet against the null, then another, then another, your running wealth forms what mathematicians call a *martingale*, and a classical result (Ville's inequality) guarantees it almost never balloons to huge values *if the null is true.* This gives e-values an almost magical property the p-value lacks: *anytime validity.* You may watch the experiment unfold, stop whenever you like, collect more data if it looks promising — **peek as often as you want** — and your error guarantee still holds. Grünwald, de Heide & Koolen call this *"safe testing"* (published in the RSS *Journal*, Series B, 2024); the broader machinery, including confidence intervals that are valid at every moment, is *"safe anytime-valid inference"* (Ramdas, Grünwald, Vovk & Shafer, *Statistical Science*, 2023). E-values also combine trivially: **multiply** independent ones, or even **average** dependent ones, and you still have a valid e-value — which makes pooling studies clean where p-values turn into a multiple-comparisons minefield.

Try it below: the same data stream, judged by a fragile peeking p-value versus an honest e-value. The toy task is intentionally narrow: it is trying to reject one claim, *"this coin is fair,"* not estimate the exact bias or prove unfairness with certainty.

What does this look like in science? In a living clinical meta-analysis, the null might be *"BCG vaccination has no clinically relevant effect on COVID-19 infection in healthcare workers."* New randomized trials report at different times, and researchers want to update the synthesis whenever fresh data arrive without letting the false-positive risk creep upward every time they look. The ALL-IN meta-analysis framework was built for exactly that kind of setting: it lets evidence from successive trials be added while preserving type-I error and interval-coverage guarantees. In one BCG/COVID application, "winning" for the evidence process would have meant accumulating strong enough evidence for a clinically relevant benefit; the anytime-valid analysis instead found no clinically relevant reduction in infections, and left hospitalization too sparse for a firm conclusion.

That is the same structure as the coin toy, with medical endpoints and trial streams replacing heads and tails.

The e-value ledger

QUANTITY	MEANING	USE
E = 1	No net betting gain against the null	Starting point
Coin-demo ticket	A likelihood-ratio payoff: 1.2 for the favored outcome, 0.8 for the other	Fair in expectation if the coin is truly P(heads) = 0.5
E = 20	A twentyfold payoff from a bet fair under the null	Level-0.05 rejection threshold because 1 / 20 = 0.05
Running wealth	A test martingale or e-process	Can be monitored continuously while preserving Type I error control

The tradeoff is conservatism: an anytime-valid ledger can need stronger or more sustained evidence than a fixed-horizon test when all modeling assumptions are exactly right.

Edge 03 [ESTABLISHED] [REVIEW]

How far has the mutiny actually spread?

Here's where the hype filter earns its keep. The *mathematics* of e-values is settled and elegant — peer-reviewed in the field's very best journals (*Annals of Statistics*, both RSS *Journals*, *Statistical Science*), and gathered into a 390-page Foundations and Trends monograph by Ramdas & Wang after its 2024 preprint. That part is `[ESTABLISHED]` beyond dispute.

Real-world *adoption* is a narrower and more honest story. The clearest uptake is in **tech-company A/B testing**, where "peeking" is the entire business model: **Optimizely** rebuilt its platform around "always-valid inference" (Johari, Koomen, Pekelis & Walsh), and **Netflix** and **Adobe** publicly run anytime-valid confidence sequences so product teams can monitor experiments continuously without cheating the statistics. That's genuine production use — but it's a long way from the world's biostatistics, psychology, and physics communities, where the p-value remains entrenched.

And the new tool is no free lunch. In fixed-horizon comparisons, e-values can need **more extreme data** than p-values to reach the same rejection threshold; Shafer's reply is that this is the cost of making the evidential scale honest rather than a simple defect. The efficiency of your bet depends on choosing a good betting strategy — arguably the same modeling judgment a Bayesian makes in choosing a prior, reappearing in new clothes. Critics including Samuel Pawel and Leonhard Held warn that branding tests as "safe" or "always valid" can mislead, since the guarantees still rest on assumptions (a correctly specified model, no publication bias) that can fail like any other. The honest verdict: a `[PROMISING]`, rigorous, genuinely useful complement to the p-value — emphatically *not* its science-wide replacement, at least not yet.

What would move the needle? If a drug regulator like the FDA or EMA blessed e-value designs for confirmatory clinical trials, or a top general-science journal wrote them into its author guidelines, the "replacement" claim could graduate from hype to hint to reality. Watch those two signals.

OPEN QUESTIONS

What's genuinely unsettled

- **What *is* a probability, really?** A frequency in the world, a degree of belief in a mind, or a fair betting rate? Three centuries on, the interpretation war has truces (de Finetti) but no surrender.
- **Where do priors come from?** Is there a principled, objective way to set your "before" belief, or does all reasoning rest on a choice no math can justify?
- **Will betting-based statistics actually take over?** Or settle in as a specialist tool for sequential experiments while the p-value rules on — and is "choose your bet" any less subjective than "choose your prior"?
- **Is the brain *literally* running Bayes?** Day 1's predictive-processing thread says perception is Bayesian inference in neural tissue. Today gives that claim its normative backbone — but "the brain approximates Bayes" and "the brain *is* Bayesian" are very different bets, and we'll return to them on **Day 119**.
- **Does Cox's theorem truly force probability on any rational agent** — including an artificial one — or only on agents that already accept his consistency axioms? (A question with teeth for the AI block, **Days 138–145**.)

◆ THE DAY IN THREE SENTENCES

BIG IDEA

Probability isn't merely a tool for dice and coins — it's the unique extension of logic into the realm of uncertainty (Cox, Jaynes), and Bayes' theorem is its law of motion: belief flows toward whatever best predicted what you actually saw.

BEST ANALOGY

Monty Hall opening a goat door — a knowledgeable agent's choice pours 2/3 of the probability onto one remaining door — and the gambler's ledger, where evidence against a hypothesis is literally money won betting against it.

LIVE CONTROVERSY

The frequentist–Bayesian split over what probability *means*, now joined by a 2020s mutiny that would replace the fragile, peek-sensitive p-value with the e-value — established as math, adopted in tech, but not (yet) the science-wide revolution its boldest fans promise.

THREADS TODAY › information (the host's reveal and the e-value both as evidence that updates belief) · computation (mind and lab as inference engines) · energy (a light callback to the Bayesian brain) — with calibration carried straight from Day 1 and Day 2.

TOMORROW → DAY 05

Causation

Today we learned how to update belief on evidence — but evidence of *correlation*. Ice cream sales and drownings rise together; neither causes the other. Tomorrow we confront the hardest upgrade in all of reasoning: telling what merely *moves with* something apart from what actually *makes it happen*. Confounders, counterfactuals, and Judea Pearl's do-calculus — the machinery for asking not "what do I expect?" but "what if I intervene?" Bring today's Bayesian instinct; you'll need to learn its limits.

SOURCES

Sources & further reading

1. Selvin, S. (1975). "A Problem in Probability" (Letter to the Editor). *The American Statistician* 29(1): 67. — and the follow-up, "On the Monty Hall Problem," 29(3): 134, the first print use of the name.

2. vos Savant, M. "Ask Marilyn." *Parade* (Sept 9, 1990, and follow-ups 1990–91). marilynvossavant.com/game-show-problem — the column, reader letters, and the ~10,000-letter / ~1,000-PhD estimates (vos Savant's own).

3. Tierney, J. (July 21, 1991). "Behind Monty Hall's Doors: Puzzle, Debate and Answer?" *The New York Times.* nytimes.com — includes Monty Hall and Persi Diaconis on the host-protocol caveat.

4. Hoffman, P. (1998). *The Man Who Loved Only Numbers.* Hyperion. — the Erdős / Vázsonyi simulation anecdote.

5. Bertrand, J. (1889). *Calcul des probabilités.* Gauthier-Villars. — Bertrand's box paradox, the structural ancestor. See also Gardner, M. (1959), "Mathematical Games," *Scientific American* (Three Prisoners).

6. Casscells, W., Schoenberger, A. & Graboys, T. B. (1978). "Interpretation by Physicians of Clinical Laboratory Results." *New England Journal of Medicine* 299(18): 999–1001. doi:10.1056/NEJM197811022991808. – only 11 of 60 clinicians gave the ~2% answer.

7. Cox, R. T. (1946). "Probability, Frequency and Reasonable Expectation." *American Journal of Physics* 14(1): 1–13. – the desiderata forcing the probability rules.

8. Jaynes, E. T. (2003). *Probability Theory: The Logic of Science.* Cambridge University Press (ed. G. L. Bretthorst). – probability as extended logic.

9. Halpern, J. Y. (1999). "A Counterexample to Theorems of Cox and Fine." *Journal of Artificial Intelligence Research* 10: 67–85. – the rigor caveat on Cox's theorem.

10. Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Foundations of the Theory of Probability). Springer. – the interpretation-neutral axioms.

11. de Finetti, B. (1937 / 1974). "La prévision..."; *Theory of Probability* (Eng. trans.). – "PROBABILITY DOES NOT EXIST"; the representation theorem.

12. Lindley, D. V. (1991). *Making Decisions*, 2nd ed. Wiley. – Cromwell's rule (p. 104).

13. Shafer, G. (2021). "Testing by Betting: A Strategy for Statistical and Scientific Communication." *Journal of the Royal Statistical Society Series A* 184(2): 407–431. doi:10.1111/rssa.12647. rss.onlinelibrary.wiley.com – with published discussion (incl. Vovk's comment, JRSS-A 184(2): 445–446).

14. Vovk, V. & Wang, R. (2021). "E-values: Calibration, combination, and applications." *The Annals of Statistics* 49(3): 1736–1754. doi:10.1214/20-AOS2020. pdf

15. Grünwald, P., de Heide, R. & Koolen, W. (2024). "Safe Testing." *Journal of the Royal Statistical Society Series B* 86(5): 1091–1128. doi:10.1093/jrsssb/qkae011 (read paper, with discussion incl. Shafer, Pawel & Held). academic.oup.com

16. Ramdas, A., Grünwald, P., Vovk, V. & Shafer, G. (2023). "Game-Theoretic Statistics and Safe Anytime-Valid Inference." *Statistical Science* 38(4): 576–601. doi:10.1214/23-STS894. arXiv:2210.01948

17. Ramdas, A. & Wang, R. (2025; first posted 2024). "Hypothesis Testing with E-values." *Foundations and Trends in Statistics* 1(1–2): 1–390. arXiv:2410.23614 – the comprehensive monograph.

18. ter Schure, J., Ly, A., Belin, L. et al. (2022). "Bacillus Calmette-Guérin vaccine to reduce COVID-19 infections and hospitalisations in healthcare workers." Prospective ALL-IN

meta-analysis preprint. Amsterdam UMC — exact e-value logrank tests and anytime-valid CIs in a living clinical meta-analysis.

19. Johari, R., Koomen, P., Pekelis, L. & Walsh, D. (2022). "Always Valid Inference: Continuous Monitoring of A/B Tests." *Operations Research* 70(3): 1806–1821. doi:10.1287/opre.2021.2135 — Optimizely's deployment; cf. Netflix Research on anytime-valid inference and Adobe's Experience Platform confidence sequences.

20. Wasserstein, R. L. & Lazar, N. A. (2016). "The ASA Statement on p-Values." *The American Statistician* 70(2): 129–133. — and Amrhein, Greenland & McShane (2019), "Retire statistical significance," *Nature* 567: 305–307.

END OF DAY 04 · 176 DESCENTS REMAIN