FOUNDATIONS TO THE 2026 RESEARCH FRONTIER,
WITH DEEP DIVE

The 180-Day Descent

By Claude Opus and GPT

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.

This deep-dive PDF includes optional appendices after the main lesson when a day has one. They are for readers with unusual interest and spare time; they are not prerequisites or building blocks for later chapters.

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 three 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, separating deduction, induction, and abduction before following valid inference into proof assistants and AI.

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

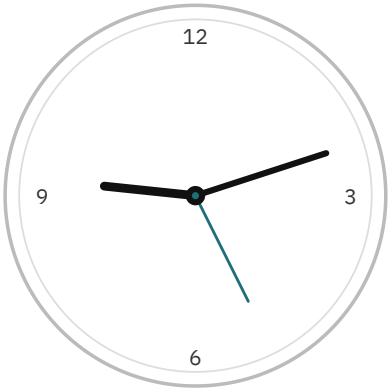
BLOCK I

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.

OPTIONAL APPENDIX

Appendix: The Rest of the Map

This section is optional supplemental reading. You can skip it without losing the main lesson.

We spent the main lesson on one belief, on one late morning. The field is far larger than one clock.

The main piece had a tight job: take a single belief — *it's 9:12* — and ask whether it counted as knowledge. To do that it leaned, quietly, on a stack of assumptions it never examined, and it strolled right past whole provinces of the subject without nodding. Does knowing require *certainty*? Can the skeptic who says you know *nothing* actually be answered? Does the word "know" even hold still from one sentence to the next? Why is knowledge worth *more* than a true belief that does the same job? And what about all the knowing that has nothing to do with facts — knowing how to swim, knowing a face, knowing a city? This appendix walks the rest of that map. Nothing here repeats the main lesson; it all hangs off its edges.

↩ CONTINUES DIRECTLY FROM

Day 1 — What Is Knowledge? There we built the three-legged stool (justified true belief), watched Gettier kick a leg out with three pages, toured the failed "fourth condition" patches, mapped Agrippa's trilemma, and ended at three frontiers: the cross-cultural test of "knowledge" intuitions, Bayesian credence, and the predictive brain. Keep two images from that day in your pocket — the *stopped clock* (right by luck, not connection) and the *archer* whose arrow is blown off course then back onto the bull (accurate, but not *apt*). Both come back transformed below.

◇ SEVEN ROOMS WE SKIPPED

1. **The trapdoors under Gettier** — the two hidden assumptions that make the trick possible, and the escape hatch (certainty) that drops you into skepticism.
2. **The skeptic at the door** — dreams, demons, brains in vats, and the 2020s simulation upgrade.

3. **"Knows" on a sliding scale** — the Bank Cases: same evidence, different stakes, opposite verdict.
4. **The luck we were really chasing** — anti-luck epistemology, which finally explains *why* the patch wars happened.
5. **Why knowing beats being right** — Meno's road, and the value of knowledge.
6. **The kinds of knowing we ignored** — how, and by acquaintance.
7. **Almost everything you know, someone told you** — testimony, disagreement, and epistemic injustice.

§1 THE MACHINERY

The two trapdoors under every Gettier case

Before we explore new rooms, look down. Gettier's three-page bomb only goes off because the floor has two trapdoors built into it — two assumptions so natural the main lesson never paused on them. Name them and the whole landscape reorganizes.

Trapdoor one: fallible justification. The classical picture lets you be *justified* in believing something that turns out *false*. Smith had excellent reason to believe "Jones will get the job" — the boss said so — and it was false. If justification had to *guarantee* truth, that step would be impossible and the case couldn't even start. **Trapdoor two: closure.** Justification (and knowledge) is assumed to travel across *entailment*: if you're justified in believing something, you're justified in believing what it obviously implies. Smith reasons from "Jones will get it (and has ten coins)" to the weaker "the winner has ten coins" — a valid inference — and carries his justification along for the ride. Knock out either plank and Gettier cases evaporate.

That hands us a tempting exit. Slam trapdoor one shut: insist that real knowledge needs *infallible* justification — reasons that make error literally impossible. No more Gettier cases, ever. This is the dream of *infallibilism*, and it is very old. Descartes went looking in 1641 for a single belief no demon could fake, and found exactly one that survives even the supposition that an all-powerful deceiver is fooling you about everything else: *I think, therefore I am*. You cannot be tricked into wrongly believing you exist, because the tricking requires a you to be tricked.

The trouble is what the demon takes with him on the way out. If knowledge demands that kind of certainty, then you do not know you have hands, that the sun will rise, that the

person across the table is your friend and not an android — because a clever enough deception could fake any of it. Buy certainty and the price is **skepticism**: the bar is set so high that almost nothing clears it. Peter Unger argued exactly this in *Ignorance* (1975) — that "knows," used strictly, applies to virtually nothing, much as "flat" strictly applies to no real surface. So infallibilism doesn't dissolve the problem; it trades a small puzzle (the odd lucky belief) for a total one (you know next to nothing). Which is our cue to open the next door, where that skeptic is already knocking.

GETTIER'S OTHER CASE, IN ONE BREATH

The main lesson used the coins. Gettier's *second* case shows trapdoor two even more nakedly. Smith, with great evidence, believes "Jones owns a Ford." From that he validly deduces "Jones owns a Ford, *or* Brown is in Barcelona" — a disjunction he's entitled to, since a true disjunct makes the whole thing true. But Jones doesn't own a Ford after all… and Brown, by pure fluke, *is* in Barcelona. The disjunction is true, justified, believed — and obviously not known. Closure carried the justification; luck supplied the truth. Same skeleton, fancier clothes.

§2 THE BIGGEST OMISSION

The skeptic at the door

Western epistemology has a recurring houseguest who refuses to leave: the figure who says you can't know *anything* about the world outside your own mind. The main lesson kept the door shut. Open it, because every modern theory of knowledge is partly built to deal with what's standing there.

The skeptic's tools are thought experiments, escalating in cruelty. First, the **dream**: right now, how do you know you're not asleep? Dreams feel utterly real from the inside; you've been fooled before. (The Daoist Zhuangzi, around 300 BCE, dreamt he was a butterfly and woke unsure whether he was a man who'd dreamt a butterfly or a butterfly now dreaming a man — the same wound the Buddhist Dharmottara reopened in the main lesson, proof again that minds keep tripping over this independently.) Descartes raised the stakes to an **evil demon** bent on deceiving you about everything. The twentieth century updated the hardware: you might be a *brain in a vat*, nerves wired to a computer feeding you exactly the experiences you're having now (Hilary Putnam, *Reason, Truth and History*, 1981). You cannot tell from the inside. That's the whole point.

Spelled out, the skeptic's argument is brutally clean — and it runs on the very closure principle from §1:

> *(1) You don't know you're not a handless brain in a vat being fed a hand-experience.*
> *(2) If you know you have hands, then (since having hands entails not being a handless vat-brain) you know you're not one.*
> *(3) So you don't know you have hands.*

Each line looks reasonable; together they seem to prove you know nothing about the external world. The interactive below lets you try every way out — and discover that each "way out" is a named philosophical position with a price tag.

The Skeptic's Syllogism, as four exits

MOVE	LINE REFUSED	REPRESENTATIVE VIEW	COST
Accept all three	None	Skepticism	You do not know you have hands, or much about the external world.
Reject P1	You do not know you are not a vat-brain	Moore's common-sense reply	Can feel like insisting rather than explaining.
Reject P2	Closure	Dretske / Nozick relevant alternatives	Closure is deeply intuitive and useful elsewhere.
Change the standard	A fixed meaning of "know"	Contextualism	The skeptic wins in the seminar; ordinary speakers win in ordinary life.

The doors are worth naming in full. **G. E. Moore** (1939) simply ran the argument backwards: *I am far more sure that here is one hand* (holding it up) *than I am of any fancy premise the skeptic offers* — so if the premises imply I don't know it, so much the worse for the premises. Cheeky, and strangely hard to beat. **Fred Dretske** (1970) and Robert Nozick (1981) took the surgical route: *deny closure.* On Dretske's *relevant alternatives* view, to know something you only need to rule out the *relevant* ways you could be wrong, not every bizarre one. At the zoo you know the animal is a zebra — you've ruled out "it's a horse," "it's a goat" — even though you haven't ruled out "it's a mule cleverly painted to look like a zebra," because in this context that's not a live possibility. Knowledge doesn't automatically transmit to every entailment. The cost is steep: closure is intuitive, and giving it up has consequences elsewhere. **Contextualism** (our next section) offers the diplomat's solution: maybe the skeptic and Moore are *both* right, because "know" means something stricter in the skeptic's seminar than in ordinary life.

The 2020s upgrade: are we in a simulation?

The vat got a software update. Nick Bostrom's **simulation argument** (*Philosophical Quarterly*, 2003) makes a careful probabilistic case that at least one of three things is true: civilizations almost never reach the technology to run ancestor-simulations; or they reach it but choose not to; or *we are almost certainly living in one*. David Chalmers, in **Reality+** (2022), takes the next step and bites a bullet most people won't: he argues we *can't know* we're not simulated and should assign the possibility real probability — but that this **isn't a catastrophe**, because *"virtual reality is genuine reality."* A simulated tree, on his *simulation realism*, is a real digital object, not an illusion; if you've always lived in a perfect simulation, your belief "that's a tree" is *true*, just realized in silicon. The skeptic assumed a fake world means false beliefs; Chalmers denies the link.

Two honest labels before we move on. The simulation *hypothesis* — that we are in fact simulated — is, as it stands, **untestable metaphysics, not science**: there's no agreed observation that would confirm or refute it, which puts it on the wrong side of the demarcation line we'll draw tomorrow. `[UNFALSIFIABLE]` The *philosophical* payoff is real all the same: it sharpens what we even mean by "real" and "know." And there's a famous reply that turns the screw the other way. Putnam argued that "I am a brain in a vat" is **self-refuting**: your words only mean what they do because of your causal history, so a lifelong brain-in-a-vat's word "vat" couldn't refer to real vats (it never causally touched one) — meaning that if you *were* a vat-brain, your sentence "I am a brain in a vat" would come out *false*. Whether that works is still argued, which is precisely why this thread runs straight into the AI block: when a system trained only on text outputs "Paris is in France," does it *know* that — or is it

the purest brain-in-a-vat of all, with words that never touched the world? Hold the question for **Days 138–145**.

§3 THE MOVING TARGET

"Knows" on a sliding scale

Here is a possibility the main lesson never entertained: maybe sixty years of hunting the perfect definition of "knows" failed because the word was never aiming at a fixed point. Consider a pair of cases from Keith DeRose (*Philosophy and Phenomenological Research*, 1992) that have launched a thousand papers — the **Bank Cases**.

It's Friday. You drive past your bank, which has a long Saturday line, and decide to come back tomorrow. Your wife asks if it'll be open Saturday. *Low stakes* version: nothing much rides on it; you say, "Yes, I know it's open Saturdays — I was here two Saturdays ago." That sounds true. You know it. *High stakes* version: there's a check that *must* be deposited by Monday or you bounce your mortgage and lose the house, and your wife points out, reasonably, that banks do change their hours. Now the very same sentence — "I know it's open Saturday" — curdles in your mouth. "Well… I'd better go in and check." Same person, same memory, same evidence, same day. Only the stakes (and whether someone raised the chance of error) have changed. Yet the knowledge seems to come and go. The dial below lets you slide between the two and watch it flip.

The Bank Cases, as a stakes table

CASE	EVIDENCE	STAKES	NATURAL VERDICT	WHAT IT TESTS
Low stakes	You were there two Saturdays ago.	A minor errand.	"I know it is open."	Ordinary standards are easy to meet.
High stakes	The same memory.	A mortgage deadline.	"I had better check."	Whether practical stakes affect knowledge.
Error raised	The same memory plus a live doubt.	Any serious consequence.	The claim to know weakens.	Whether context shifts the word or the knower's state.

Three camps, three diagnoses of the same data. **Contextualism** (DeRose; David Lewis, "Elusive Knowledge," 1996; Stewart Cohen, 1988) locates the shift in the *word*: "knows" is like "tall" or "here" — context-sensitive. Raising the stakes or mentioning error raises the standard a belief must meet for the sentence "S knows" to count as true. Both utterances are correct, in their own conversations. The skeptic is even right in the seminar — he's just jacked the standard sky-high. **Pragmatic encroachment** (Jason Stanley, *Knowledge and Practical Interests*, 2005; Fantl & McGrath; John Hawthorne, *Knowledge and Lotteries*, 2004) puts the shift in the *knower*: what *you* know genuinely depends on what's practically at stake *for you*, because knowledge is supposed to be the thing you can act on. High stakes really can deprive you of knowledge you'd have had when it didn't matter — a startling idea, since it lets practical pressure "encroach" on a supposedly purely factual state. **Invariantism** (the traditional holdout) digs in: "knows" means one fixed thing, the standards don't move, and one of your two verdicts is simply mistaken — you either knew all along or never did, and the stakes just changed how *willing* you were to *say* so. `[AGREED]` `[UNRESOLVED]` The data is robust; its explanation is one of the most active fault lines in the field.

§4 THE PATTERN BEHIND THE PATCHES

The luck we were really chasing

Return to the patch wars from the main lesson — no-false-lemmas, sensitivity, safety, virtue. They looked like a grab-bag of clever fixes that each met a nastier counterexample. Step back and they snap into focus: every one was chasing the *same ghost*. Duncan Pritchard gave it a precise name in *Epistemic Luck* (Oxford, 2005). The enemy of knowledge is a specific species he calls *veritic luck*: your belief is true in the actual world, but in *almost all the nearby ways things could have gone*, you'd have believed the same thing and been wrong. The truth and your believing it are only accidentally in step.

This is the deep content of the "safety" idea, and it's worth *seeing*. Picture the actual world as a dot, ringed by the nearby possible worlds — the small, realistic variations on how things might have been. A belief is *safe* (knowledge-grade) when it stays true across that neighborhood, and *unsafe* (merely lucky) when a slight nudge flips it to false. Toggle the three scenarios below and watch the neighborhood light up.

Safe vs. Lucky, as nearby-worlds cases

SCENARIO	ACTUAL WORLD	NEARBY WORLDS	VERDICT
Working clock	Your belief is true.	Small variations still leave you right.	Safe: knowledge-grade.
Stopped clock	Your belief is true at 9:12.	A minute earlier or later, the same belief is false.	Unsafe: veritic luck.
Fake-barn country	You see the one real barn.	Most nearby looks would have landed on facades.	Unsafe: environmental luck.

That single picture retroactively explains the whole mess. The stopped clock fails *hard* — a minute either side and you're wrong, so the neighborhood is a sea of red. Fake-barn country is subtler: the barn you're looking at is genuinely there (the core is green), but you're

surrounded by façades, so a glance a hundred meters either way would have fooled you — red neighborhood, no knowledge, even with a true justified belief and no false premise. The patches all failed because each tried to capture "green neighborhood" with a slightly different yardstick, and luck kept finding the gaps.

Two more patches the main lesson didn't name, now that we have the frame. **Defeasibility theory** (Lehrer & Paxson, 1969) said knowledge is *un-defeated* justified true belief: there must be no true fact out there that, if you learned it, would dissolve your justification. It handles many cases elegantly — until the "misleading defeater" twist, where there's a true-but-misleading fact that *shouldn't* rob you of knowledge but technically does, forcing ever-finer distinctions. And reaching back further, the **causal theory** (Goldman, 1967, before he turned reliabilist) demanded that the fact *cause* your belief — no causal chain, no knowledge. Beautiful for perception; fatal for mathematics, since the number 7 and the Pythagorean theorem don't cause anything (Paul Benacerraf pressed exactly this "access problem" in 1973). You can't shake hands with an abstract object.

And the deepest crack in reliabilism, which the main lesson only gestured at: the **generality problem** (Conee & Feldman, 1998). Reliabilism says a belief is justified if produced by a *reliable process* — but *which* process? Your belief that it's 9:12 was produced by "reading a clock," and also by "reading *that* clock," and "using vision in dim light," and "trusting instruments on Tuesdays" — each as real as the others, each with a different reliability score. Pick the type and you've picked the verdict. Specifying the "right" grain, in a principled way, has proven stubbornly hard.

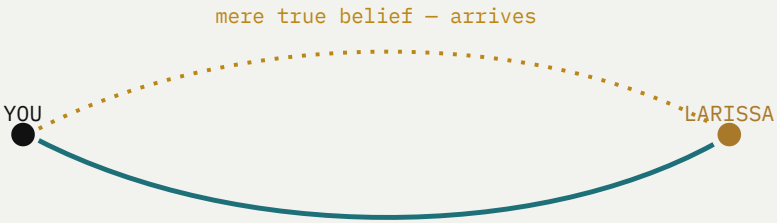
Where does Pritchard land? At *anti-luck virtue epistemology*: knowledge needs *both* conditions, because they catch different failures. You need **safety** (a green neighborhood — no veritic luck) *and* you need **aptness** (the belief is true *through your own ability* — the archer's skill from the main lesson). Neither alone suffices: the stopped clock can fail safety, fake-barn country can have local skill but bad luck. It's not a tidy three-word formula — and that, by now, may be the lesson. Knowledge might just *be* the kind of thing that takes two independent guarantees, one about you and one about your world.

§5 THE QUESTION UNDER THE QUESTION

Why is knowing worth more than just being right?

Step back from "what is knowledge?" to a question Plato asked first and nobody has fully answered: *why do we care?* If a true belief gets the job done, what does the extra machinery of knowledge buy you? Plato put it as a traveler's problem in the *Meno* (~380 BCE).

Suppose you want to walk to the town of Larissa. A person who *knows* the road will get you there. But so will a person who merely has a *true belief* about the road — who's never been, but happens to be right. For the purpose of arriving, the two are worth exactly the same. So why has the entire tradition prized knowledge above true belief? This is the *value problem*, and it's a load-bearing question: a theory of knowledge that can't say why knowledge is *better* has arguably missed the point of the concept.



If both roads reach Larissa, what is the second one worth?

The value problem turns into a precise weapon against one of the main lesson's theories. It's called the *swamping problem* (Linda Zagzebski, 2003). Reliabilism says knowledge is true belief from a reliable process. But ask *what the reliability adds in value*. Reliability is good only because it tends to produce truth. So once you *already have* the truth, what does it add that this particular truth also came from a reliable source? Zagzebski's homely analogy: a cup of good coffee is no *better* to drink for having come from a reliable coffee machine rather than an unreliable one that happened to produce an identical cup. The good-making feature (deliciousness / truth) is already present; the source's reliability gets *swamped*, adding nothing. If that's right, reliabilism can't explain why knowledge beats lucky true belief — the very thing a theory of knowledge most needs to deliver.

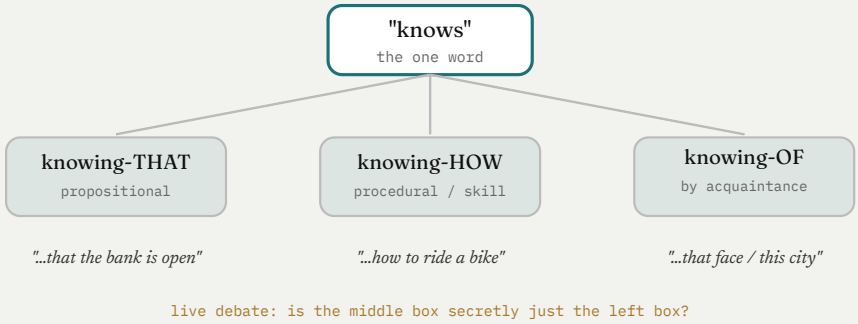
This is where **virtue epistemology** earns its keep, and where the archer finally pays off. Its answer: knowledge isn't valuable as a *better-stocked* true belief; it's valuable as an *achievement* — a success that's *yours*, that came about *through your own competence*. And achievements carry a kind of worth that lucky successes never do, the way a bullseye you actually aimed is worth something a lucky gust-blown hit isn't, even though the arrow lands in the same spot. A true belief reached through your own cognitive skill is a *cognitive achievement*; a lucky true belief is not. That's the extra value — not in the result, but in the

getting there. The road to Larissa you can actually find again is worth more than the one you stumbled onto, even on a day you both arrive.

§6 THE OTHER KNOWINGS

The kinds of knowing we ignored

Everything so far — the entire main lesson — was about *propositional knowledge*, knowledge-*that*: knowing *that* it's 9:12, *that* Jones got the job. But look how much of "know" in plain English isn't that at all. You know *how* to ride a bicycle. You know your mother's face. You know Lisbon. None of these is a stockpile of facts, and philosophers have argued for a century about how they relate.



One English verb, at least three different relations to the world.

Knowing-how. Gilbert Ryle, in *The Concept of Mind* (1949), insisted that knowing how to do something is not knowing a set of facts. A brilliant cyclist may be unable to state a single law of balance; a person who has memorized every fact about bicycles may topple on the first try. Worse, Ryle argued, reducing skill to facts triggers a regress: if every skilled act required first *knowing the proposition* describing the rule, you'd then need the skill of *applying* that rule, which would need another rule, forever. So skill must be its own kind of knowing. The twist: Jason Stanley and Timothy Williamson fired back in **"Knowing How"** (2001) with *intellectualism* — the claim that knowing-how just *is* a species of knowing-that after all (knowing, of some way to ride, *that* it is a way to ride), dressed in different grammar. Whether skill collapses into propositions is genuinely unsettled. [CONTESTED]

Knowing by acquaintance. Bertrand Russell (1911) drew a second cut: between knowledge *by acquaintance* — your direct, unmediated grip on a patch of red you're seeing, a pain you're feeling, a face you're looking at — and knowledge *by description*, the facts you know *about* things you've never directly met ("the first person to stand on the Moon," whom you know only as the one satisfying that description). You can know a stupendous amount *about* Bismarck and never have known *him*; you know the color red in a way the world's greatest blind physicist, who knows every fact about wavelengths, does not. That gap — facts about an experience versus the experience itself — is a quiet seed for the hardest problem in the entire course, the one waiting on **Day 123**: why there's *something it is like* to see red at all.

§7 THE SOCIAL TURN

Almost everything you know, someone told you

The main lesson, like most of traditional epistemology, imagined a lone mind facing the world — one person, one clock. But run an audit of what you actually know. That the Earth is about 4.5 billion years old. That Antarctica exists. Your own date of birth. The boiling point of water. You verified essentially none of it first-hand; you were *told*, by teachers, books, parents, instruments, strangers. *Testimony* is the overwhelming bulk of any human being's knowledge — and for centuries epistemology treated it as an afterthought.

The central question is whether trusting testimony is something you have to *earn* or something you're *entitled* to by default. **David Hume** (1748) took the demanding line: testimony is only as good as your own inductive track record of when testimony has proved reliable — it *reduces* to evidence you've personally gathered. **Thomas Reid** (1764) found this absurd: no child could bootstrap a track record before trusting anyone, and in fact we're built with a "principle of credulity," a default disposition to believe what we're told, exactly as we're built to trust our senses. On Reid's *anti-reductionist* view, testimony is a *basic* source of knowledge, not a derived one — and it has to be, or knowledge couldn't get off the ground in a social animal. The modern field mostly agrees that some default trust is unavoidable; the fights are over how much, and when it's defeated.

Two newer rooms branch off this one, and both matter enormously in 2026. The first is **disagreement**. When someone you regard as an *epistemic peer* — as smart, as informed, as careful as you — looks at the same evidence and concludes the opposite, what should you do? The *conciliationist* or "equal-weight" view (Adam Elga, *Noûs*, 2007; David Christensen, 2007) argues you should move substantially toward them: to stay put is to claim, with no independent reason, that *you're* the one who got it right and they made the mistake. The *steadfast* view answers that sometimes you can rationally hold your ground, because your

own reasoning is evidence too. It sounds abstract until you notice it's the whole epistemology of echo chambers, expert consensus, and what to do when half your sources contradict the other half. [DEBATE]

The second is sharper still: **epistemic injustice**, named by Miranda Fricker (*Epistemic Injustice: Power and the Ethics of Knowing*, 2007). Because so much knowing runs on testimony, *who gets believed* becomes an ethical question, not just an epistemic one. Fricker isolates two wrongs. *Testimonial injustice*: a speaker's word is given less credence than it deserves because of prejudice about who they are — the patient whose pain is dismissed, the witness disbelieved for their accent or gender. *Hermeneutical injustice*: subtler and deeper — a person can't even make sense of their own experience, to themselves or others, because the surrounding culture hasn't yet developed the *concept* for it (her example: the experience we now call sexual harassment, suffered by people who had no word for it and so couldn't name the wrong). Knowledge, it turns out, has a politics: the tools for understanding are unevenly distributed, and that unevenness can itself be an injustice.

THE FUNCTION-FIRST ESCAPE HATCH

There's a radical way to end the whole 180-page hunt for a definition, and it threads the social turn back to the start. Edward Craig, in *Knowledge and the State of Nature* (1990), proposed: stop asking *"what is knowledge?"* and ask *"what is the concept FOR — why would creatures like us ever invent it?"* His answer: a social, language-using species desperately needs a way to flag **good informants** — to mark out whose word you can act on. "Knowledge" is the tag we evolved to pin on reliable sources of true information. That instantly explains the things the analyses struggled with: why knowledge must be *true* (a tip that's false is worthless), why *luck* disqualifies (you can't rely on a fluke next time), and why we care at all (survival in a world where most of what you need to know, you must get from others). It rhymes with Williamson's "stop trying to define it," and it cashes out the main lesson's open question — did *evolution* install the instinct that luck-based knowing doesn't count? Craig's answer is essentially: yes, and here's why it would.

§8 THE FORMAL EDGE

Two more frontiers, beyond Bayes

The main lesson's formal frontier was Bayesian credence. Two further formal ideas deserve a place on the map, because both bite ordinary intuitions and both feed straight into computer science and AI.

The logic of knowing. Jaakko Hintikka, in *Knowledge and Belief* (1962), treated "knows" as a formal operator you can reason with, like "necessarily" — launching *epistemic logic*, now a workhorse in computer science (reasoning about what distributed agents and AI systems "know"). It immediately surfaces deep puzzles. The *KK principle*: if you know *p*, do you thereby know *that you know p*? Tempting, but Williamson (from the main lesson) argues it's false — you can know something without being in a position to know that you know it, because knowledge has blurry margins. And *logical omniscience*: the clean logic implies that if you know some axioms, you know *every* logical consequence of them — which would make every mathematician instantly aware of every theorem. Obviously false for real, bounded minds, and a central headache for modeling actual reasoners (and machines).

The preface paradox. A companion to the lottery paradox from the main lesson, and arguably nastier. You write a long, careful book. For *each* claim in it, you've checked your work and rationally believe it's true. Yet you also write, sincerely, in the preface: "no doubt errors remain, and they are mine alone" — because you know that across hundreds of claims, you've almost certainly slipped *somewhere*. So you rationally believe each individual claim, and *also* rationally believe that *at least one of them is false* (David Makinson, "The Paradox of the Preface," 1965). Those can't all be true together. The moral lands on the main lesson's open question with full force: ordinary all-or-nothing belief isn't *closed under conjunction* — believing each of many things doesn't license believing their grand conjunction — which is one more reason the field keeps drifting from yes/no belief toward graded credence. The dial, again, doing what the switch can't.

◆ THE APPENDIX IN THREE SENTENCES

BIG IDEA

The main lesson made knowledge look like one tidy puzzle — find the fourth condition — but it's really a constellation: whether certainty is required (and the skeptic that demand invites), whether "knows" even holds still as stakes change, what knowledge is *worth* over mere true belief, and the fact that almost all of it comes from *other people*.

BEST NEW ANALOGY

The neighborhood of nearby possible worlds: knowledge is a belief whose neighborhood stays green (safe), while luck is a belief one nudge from red — and the road to Larissa you can find *again* is worth more than the one you stumbled onto, even when both arrive.

LIVE CONTROVERSY

Why the Bank-Case verdict flips — context shifting the *word* "knows" (contextualism), stakes shifting what the *knower* knows (pragmatic encroachment), or neither (invariantism) — is among the field's hottest open fault lines, alongside whether closure can be denied and whether knowing-how is secretly knowing-that.

THREADS HERE › information (testimony & the social transmission of knowledge; preface/credence) · computation (epistemic logic; modal "neighborhoods" of worlds) · evolution (Craig: the concept of knowledge as a good-informant detector built for a social species) — picking up the same five we're tracking all 180 days.

OPEN QUESTIONS

What this appendix leaves unsettled

- **Certainty or not?** Is the infallibilist right that real knowledge needs error-proof reasons (inviting skepticism) — or is fallible knowledge the only kind worth wanting?
- **Can closure be denied without disaster?** Dretske and Nozick block the skeptic by giving it up; the cost elsewhere is still being counted.
- **Does "knows" move?** Context-sensitive, stakes-sensitive, or fixed — and if it moves, what exactly is moving, the word or the world?
- **Can the value of knowledge be explained at all**, or does every account leave knowledge looking no better than lucky true belief?
- **Is knowing-how just knowing-that** in disguise, or its own irreducible kind of grip on the world?
- **Is testimony basic or earned?** — and, downstream, when a peer disagrees, must you really meet them halfway?
- **And the function-first wager:** if the concept of knowledge exists to flag good informants, does that *dissolve* the analysis project — or just relocate it?

SOURCES

Sources & further reading

Classical works are cited by original date; all are standard, widely available editions. Verified secondary anchors and reference entries are linked.

1. Descartes, R. (1641). *Meditations on First Philosophy.* — methodic doubt, the evil demon, and the cogito as the one indubitable point.
2. Unger, P. (1975). *Ignorance: A Case for Scepticism.* Oxford University Press. — infallibilism pushed to its skeptical conclusion ("knows," like "flat," applies to almost nothing).
3. Moore, G. E. (1939). "Proof of an External World." *Proceedings of the British Academy* 25: 273–300. — "Here is one hand": running the skeptical argument in reverse.
4. Dretske, F. (1970). "Epistemic Operators." *Journal of Philosophy* 67(24): 1007–1023. — denying closure; the relevant-alternatives view; the zebra/painted-mule case.

5. Nozick, R. (1981). *Philosophical Explanations.* Harvard University Press. — sensitivity / truth-tracking and its own denial of closure.

6. Putnam, H. (1981). *Reason, Truth and History.* Cambridge University Press. — the brain-in-a-vat, and the semantic-externalist argument that "I am a BIV" is self-refuting.

7. Bostrom, N. (2003). "Are You Living in a Computer Simulation?" *Philosophical Quarterly* 53(211): 243–255. simulation-argument.com

8. Chalmers, D. J. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy.* W. W. Norton / Allen Lane. — "virtual reality is genuine reality"; simulation realism. consc.net/reality

9. DeRose, K. (1992). "Contextualism and Knowledge Attributions." *Philosophy and Phenomenological Research* 52(4): 913–929. — the Bank Cases. See also DeRose (1995), "Solving the Skeptical Puzzle," *Philosophical Review* 104(1): 1–52.

10. Lewis, D. (1996). "Elusive Knowledge." *Australasian Journal of Philosophy* 74(4): 549–567. — contextualism and the rule of attention.

11. Cohen, S. (1988). "How to Be a Fallibilist." *Philosophical Perspectives* 2: 91–123. — the airport cases.

12. Stanley, J. (2005). *Knowledge and Practical Interests.* Oxford University Press. — pragmatic encroachment / interest-relative invariantism. See also Hawthorne, J. (2004), *Knowledge and Lotteries* (OUP); Fantl, J. & McGrath, M. (2009), *Knowledge in an Uncertain World* (OUP).

13. Pritchard, D. (2005). *Epistemic Luck.* Oxford University Press. — the modal account of luck; veritic luck; the safety condition; later, anti-luck virtue epistemology. Overview: IEP, "Epistemic Luck."

14. Lehrer, K. & Paxson, T. (1969). "Knowledge: Undefeated Justified True Belief." *Journal of Philosophy* 66(8): 225–237. — the defeasibility analysis.

15. Goldman, A. (1967). "A Causal Theory of Knowing." *Journal of Philosophy* 64(12): 357–372. — and Benacerraf, P. (1973), "Mathematical Truth," *J. Phil.* 70(19): 661–679, on why it fails for abstract objects.

16. Conee, E. & Feldman, R. (1998). "The Generality Problem for Reliabilism." *Philosophical Studies* 89(1): 1–29.

17. Plato. *Meno* (~380 BCE). — the road to Larissa; the value problem (knowledge vs. true belief).

18. Zagzebski, L. (2003). "The Search for the Source of Epistemic Good." *Metaphilosophy* 34(1–2): 12–28. — the swamping problem. See also Kvanvig, J. (2003), *The Value of Knowledge and the Pursuit of Understanding* (Cambridge UP).

19. Ryle, G. (1949). *The Concept of Mind.* University of Chicago Press. — knowing-how vs. knowing-that; the regress of rules.

20. Stanley, J. & Williamson, T. (2001). "Knowing How." *Journal of Philosophy* 98(8): 411–444. — intellectualism: knowing-how as a species of knowing-that.

21. Russell, B. (1910–11). "Knowledge by Acquaintance and Knowledge by Description." *Proceedings of the Aristotelian Society* 11: 108–128.

22. Hume, D. (1748). *An Enquiry Concerning Human Understanding*, §X. – the reductionist view of testimony. Reid, T. (1764). *An Inquiry into the Human Mind on the Principles of Common Sense.* – testimony as a basic source (anti-reductionism).

23. Elga, A. (2007). "Reflection and Disagreement." *Noûs* 41(3): 478–502. doi:10.1111/j.1468-0068.2007.00656.x. And Christensen, D. (2007), "Epistemology of Disagreement: The Good News," *Philosophical Review* 116(2): 187–217.

24. Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing.* Oxford University Press. – testimonial and hermeneutical injustice.

25. Craig, E. (1990). *Knowledge and the State of Nature: An Essay in Conceptual Synthesis.* Oxford University Press. – the function-first / good-informant genealogy of the concept.

26. Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions.* Cornell University Press. – epistemic logic; the KK principle; logical omniscience.

27. Makinson, D. C. (1965). "The Paradox of the Preface." *Analysis* 25(6): 205–207.

28. Reference surveys: *Stanford Encyclopedia of Philosophy* – "Skepticism," "Epistemic Contextualism," "The Value of Knowledge," "Epistemological Problems of Testimony," "Epistemic Injustice."

OPTIONAL APPENDIX

Appendix: The Edge of the Map

This section is optional supplemental reading. You can skip it without losing the main lesson.

The coastline still being drawn. Recent, high-voltage, and — every line of it — not yet safe to stand on.

The first appendix charted the *settled* hinterland — skepticism, contextual "knows," the value of knowing, the social web — territory mapped decades or centuries ago. This one sails to the edge, where the cartographers are still arguing about where the shore is. Everything below is peer-reviewed work from **2020 onward** that could genuinely redraw what we mean by "knowledge" — and precisely *because* it's that new, the hype filter does the heavy lifting. Nothing here is bankable. Each frontier comes with its own counter-literature already forming, and every claim wears a tag: `[ESTABLISHED]` `[PROMISING]` `[CONTESTED]` Read it the way you'd read a dispatch from an expedition still underway — thrilling, partial, and subject to revision by the next ship back.

↩ THIRD IN A SEQUENCE

Day 1 — What Is Knowledge? built the stool and watched Gettier kick a leg out. **Appendix I — "The Rest of the Map"** walked the settled provinces: skepticism, contextualism, anti-luck epistemology, the value problem, testimony and epistemic injustice. This piece is the live edge of that same continent. Where the social-turn rooms of Appendix I (testimony, disagreement, who gets believed) described the *structure* of knowing-from-others, several frontiers here describe what happens when that structure comes under deliberate *attack* — by manipulators, by machines, by the feed.

◈ SIX PLACES THE SHORELINE IS MOVING

1. **The zetetic turn** — epistemology pivots from *belief-states* to the *act of inquiry*, and finds its old rules in conflict with the new ones.

2. **Knowledge before belief** — cognitive science flips the furniture: maybe representing *knowledge* is more basic than representing belief.
3. **Do machines know — or bullshit?** — the philosophy of large language models, and a deliberately rude diagnosis.
4. **The epistemic backstop collapses** — deepfakes quietly remove a support that's been holding up testimony all along.
5. **Hostile epistemology** — echo chambers, manufactured clarity, and trust as something that can be *weaponized*.
6. **Accuracy-first** — a formal refoundation that re-derives Day 1's Dutch book from *truth* instead of *money*.

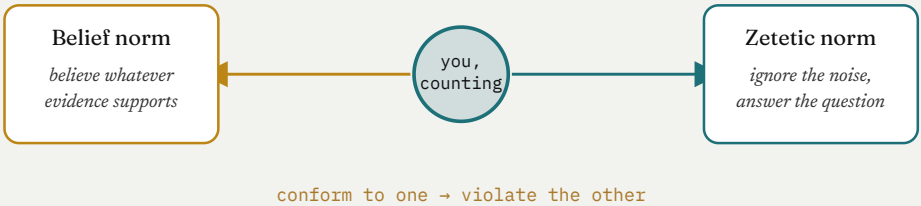
§1 THE ACTIVITY TURN

Epistemology forgot to ask about *inquiry*

[PROMISING][CONTESTED]

Here is a strange omission, and once you see it you can't unsee it. For a century, epistemology has been almost entirely a theory of *states* — of belief, justification, knowledge: snapshots of a mind that has already finished thinking. It has had remarkably little to say about the *activity* that produces those states — the messy business of *inquiry*: asking a question, deciding what evidence to gather, knowing when to stop. Jane Friedman put a name to the missing half and lit a fuse under the field. The norms of inquiry she calls *zetetic* norms (from the Greek *zētein*, to seek), and her landmark paper "The Epistemic and the Zetetic" (*Philosophical Review*, 2020) argues something genuinely destabilizing: the norms of *inquiry* and the norms of *belief* are not merely separate — they actively **conflict**.

The engine is a principle so obvious it sounds like a truism — the *Zetetic Instrumental Principle*: if you want to figure out the answer to a question, you ought to take the necessary means to figure it out. Now watch it collide with a bedrock epistemic norm — the evidentialist's command that you may believe whatever your evidence already supports. Suppose you're trying to count the windows on the building across the street. Good inquiry says: *focus, go count the windows, don't get distracted.* But at every passing instant your senses hand you sufficient evidence to form, and be permitted to believe, a thousand idle truths — the color of that car, the number of people on the corner, the shape of a cloud. The belief-norm *permits* all of them. The inquiry-norm tells you to *ignore* all of them and count windows. Conform to one and you flout the other. The diagram makes the squeeze concrete.



Friedman's tension: good inquiry and permissible belief pull opposite ways.

Why does this matter beyond the seminar? Because it suggests epistemology has been studying the wrong unit. If belief-norms and inquiry-norms genuinely clash, then a theory built only on belief is incomplete — maybe even *backwards*. The radical proposal, the "zetetic turn," is that **all epistemic norms are ultimately norms of inquiry** (suspension of judgment becomes a *question-directed* attitude; believing an answer is a way of *closing* a question). The field has not swallowed this whole — and that's the honest part. Arianna Falbo ("Should epistemology take the zetetic turn?", *Philosophical Studies*, 2023) and others argue the inquiry-norms are really *practical*, not distinctively epistemic, and that a purely zetetic epistemology can't explain why some beliefs are irrational even when believing them would *help* your inquiry. So: the *puzzle* is now taken with great seriousness across the field; the *grand thesis* that inquiry swallows everything is a genuine, unresolved fight. Either way, the question "what is knowledge?" is quietly being reframed as "what is it to inquire *well*?" — and that reframing reaches all the way forward to **Day 2**, where the scientific method is exactly a set of norms for collective inquiry.

§2 THE COGNITIVE-SCIENCE TURN

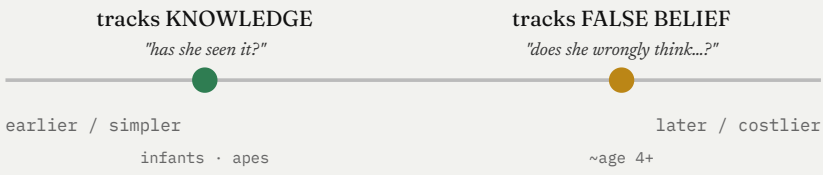
What if knowledge comes *before* belief?

[PROMISING][CONTESTED]

Day 1 treated knowledge as something *built up from* belief: take a belief, add truth, add justification, screen out luck. Almost every theory we met assumed belief is the raw material and knowledge the finished product. A large interdisciplinary team led by Jonathan Phillips and Joshua Knobe dropped a target article in *Behavioral and Brain Sciences* — "Knowledge

before belief" (2021) — arguing that, as a matter of how human (and animal) minds actually work, this may be exactly **upside down**.

The standard story in psychology is that our "theory of mind" — our capacity to model other minds — is centered on *belief*, and matures when a child finally passes the *false-belief test* (understanding that someone can hold a belief the child knows to be false) around age four. Phillips and colleagues marshal converging evidence that representing *knowledge* is the more basic feat. The threads: developmentally, infants and toddlers track who has *seen* or has *access to* information — who *knows* — well before they can handle false beliefs. Comparatively, non-human great apes show robust signs of tracking what others can perceive and know, while convincing evidence of belief-tracking remains elusive. And in adults, attributions of knowledge are made at least as fast as — sometimes faster than — attributions of belief, which is hard to explain if "X knows p" is computed by first building "X believes p" and then checking extra conditions. The picture they paint is a *factive theory of mind*: the mind's first and most basic tool for modeling others is "what do they know?", with the trickier, error-tolerant "what do they merely *believe* (perhaps falsely)?" coming later and costing more.



The proposed ordering — the reverse of the textbook "belief-first" story.

If it holds, the payoff for Day 1 is direct and large. It would be empirical ammunition for Timothy Williamson's *knowledge-first* program — the philosophical claim (which we met as pure armchair theory) that knowledge is the basic, unanalyzable state and belief should be explained in terms of *it*, not the reverse. Suddenly that's not just a logician's hunch; it's a candidate fact about the architecture of cognition, with an evolutionary rationale (a social animal urgently needs to track *who has reliable information* — echoing Edward Craig's "good-informant" genealogy from Appendix I). But honesty demands the asterisk, and here it's loud: a BBS target article arrives wrapped in dozens of peer commentaries, and many dissent hard. Critics argue the knowledge/belief line is blurrier than the authors allow, that "tracking who saw what" needn't be a full representation of *knowledge*, and that culture and

language shape the whole picture. So: the *behavioral findings* (kids and apes track informational access early) are reasonably solid; the *strong interpretation* (knowledge-representation is metaphysically and computationally prior, with belief assembled from it) is very much live. The furniture of the mind is being rearranged in real time, and the movers don't yet agree on the floor plan.

§3 THE MACHINE TURN, PART ONE

Does a language model know anything — or just bullshit?

[ESTABLISHED][CONTESTED]

Day 1 ended on a needle of a question: when a system like the one that drafted these pages outputs a true, well-supported sentence, does it *know* anything — or is it the ultimate Gettier case, right for reasons that have nothing to do with the truth? The 2020s turned that closing flourish into one of the hottest debates in the field, and the most-discussed entry has a title that sailed past peer review unblunted: "ChatGPT is bullshit" (Hicks, Humphries & Slater, *Ethics and Information Technology*, 2024).

Their move is precise, not merely rude. They borrow Harry Frankfurt's technical sense of *bullshit* (from his 1986 essay *On Bullshit*): bullshit is speech produced with *indifference* to truth. The liar at least tracks the truth — he has to, in order to steer you away from it. The bullshitter doesn't care either way; he says whatever serves his purpose, and whether it's true is simply beside the point. Now consider what a large language model fundamentally *is*: a system trained to emit the statistically likely next token, to produce fluent, plausible-sounding text. It has no representation of truth that it is trying to honor. So when it states a real fact and when it "hallucinates" a fake citation, it is doing the *very same thing* — generating likely-looking text — and succeeding equally at its actual task in both cases. On this view, "hallucination" is a flattering misnomer that implies a malfunction; the truer description is that the system is **indifferent to truth by design**, which is bullshit in Frankfurt's exact sense. They distinguish *soft* bullshit (no intent to deceive, just truth-indifference) from *hard* bullshit (additionally posing as a sincere truth-teller), and argue an LLM is at minimum a soft bullshitter.

Why this could redraw the map: it cuts directly against the loose talk of machines "knowing," "understanding," or "believing." If the argument is right, an LLM's true outputs aren't knowledge and aren't even really *assertions* in the full sense — they're a new category of truth-apt-looking text with no one home who cares whether it's true. That reframes how

we should trust, cite, and regulate these systems. And it is, predictably, *contested* — the rebuttals are already a small literature. Some argue the "bullshit" label smuggles in a stance on whether models have intentions at all (Sarah Fisher, "Large language models and their big bullshit potential," 2024; David Gunkel & Simon Coghlan, "Cut the crap," 2025); others that as models are trained with reinforcement to be truthful and to express calibrated uncertainty, "indifferent to truth" is too crude. What's *settled* is the unglamorous core: a base language model has no built-in commitment to truth, and fluency is not knowledge. What's *open* is whether "bullshit," "instrument," "testifier," or some entirely new epistemic category is the right home for what these systems produce. It is the brain-in-a-vat from Appendix I made of silicon and shipped to a billion users — words that may never have touched the world, now answering our questions.

§4 THE MACHINE TURN, PART TWO

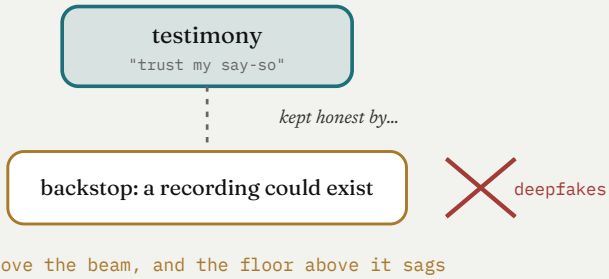
The support beam you never noticed: the epistemic backstop

[PROMISING][CONTESTED]

Appendix I established how much of what you know arrives by *testimony* — other people's say-so. Regina Rini's "Deepfakes and the Epistemic Backstop" (*Philosophers' Imprint*, 2020) identifies a hidden structural support that has quietly kept that whole edifice honest — and shows how a new technology is sawing through it.

The insight is subtle. Why is testimony as reliable as it is? Part of the answer, Rini argues, is a silent regulator: the ever-present *possibility* of a recording. When someone can be contradicted by a photograph, an audio clip, or a video, they have a standing incentive to testify truthfully — because a recording might surface and catch them out. Recordings function as an *epistemic backstop*: not because we constantly check them, but because their mere availability disciplines testimony, the way an unused referee still shapes a game. For roughly a century — since photography and audio became hard to fake — we have built our norms of public truth-telling on this backstop without ever naming it. Deepfakes — convincingly fabricated video and audio, generated by the same machine-learning wave as §3 — dissolve it from both sides. They flood the channel with convincing fakes *and*, just as corrosively, they hand every caught wrongdoer a new escape: *that recording of me is probably a deepfake.* The "liar's dividend." Once any recording can be waved away, the backstop stops disciplining testimony — and testimony itself, our single largest source of knowledge, loses a support we didn't know it leaned on. Don Fallis sharpens the same worry in information-theoretic terms ("The Epistemic Threat of Deepfakes," *Philosophy & Technology*, 2021):

deepfakes *reduce the information* a video carries about what actually happened, degrading it as a signal.



Knock out a support no one was looking at, and the structure it held still falls.

This is impactful precisely because it reframes deepfakes as an *epistemological* problem, not just a fraud or privacy one: the threat isn't only the specific lies, it's the erosion of a background condition for trusting recordings at all. But — true to this appendix — the magnitude is contested, and the pushback is sharp and worth taking seriously. Joshua Habgood-Coote ("Deepfakes and the epistemic apocalypse," *Synthese*, 2023) argues the doom framing is overblown: we have never relied on recordings as infallible, we already cross-check testimony against many sources, and societies have absorbed media-manipulation panics before. Atencia-Linares and Artiga ("Deepfakes, shallow epistemic graves," *Synthese*, 2022) defend the residual epistemic robustness of photography and video. So the *mechanism* Rini names — recordings as a silent regulator of testimony — is a genuine and illuminating contribution; the *prediction* of an "epistemic apocalypse" or wholesale collapse of public knowledge is a live dispute, not a settled forecast. Bring this one forward to **Day 2**, where science's answer to "how do you trust a report you can't personally verify?" is a whole institutional machinery of replication and recording — and to the AI block, where it meets §3 head-on.

§5 THE ADVERSARIAL TURN

Hostile epistemology: when the environment is built to fool you

[ESTABLISHED][CONTESTED]

Traditional epistemology pictured a lone, neutral mind facing a neutral world. C. Thi Nguyen's program — he calls it *hostile epistemology* — starts from a darker and more modern premise: your epistemic environment is not neutral. It is increasingly *engineered*, often by parties with an interest in what you come to believe, to exploit the predictable shortcuts your mind must use. Three of his post-2020 moves have reshaped how the whole field talks about online life.

The first is a distinction that sounds academic and turns out to be the key to everything: the difference between an *epistemic bubble* and an *echo chamber* ("Echo Chambers and Epistemic Bubbles," *Episteme*, 2020). They are *not* the same thing, and conflating them is why so many well-meaning fixes fail. In a bubble, outside voices are merely *absent* — you simply haven't been exposed to them (think a filter that only ever shows you agreeable sources). In a chamber, outside voices are *present but actively discredited* — you've been trained to *distrust* them in advance ("the mainstream media lies," "experts are corrupt"). The consequence is stark and counterintuitive, and the interactive below lets you feel it: the obvious intervention — *expose people to the other side* — pops a bubble but can *strengthen* a chamber, because inside a chamber, encountering the enemy's argument is exactly what the chamber predicted, and so confirms it.

Bubble vs. Chamber, as exposure outcomes

STRUCTURE	OUTSIDE VOICES	EXPOSURE OUTCOME	LESSON
Epistemic bubble	Absent, not refuted.	New sources can connect and puncture the bubble.	Exposure can work when the problem is missing information.
Echo chamber	Present but pre-discredited.	Exposure can reinforce distrust because the chamber predicted hostile outsiders.	The obvious repair can backfire when distrust is built into the structure.

The second move names a vulnerability inside your own head. In "The Seductions of Clarity" (*Royal Institute of Philosophy Supplement*, 2021), Nguyen argues that the *feeling* of

clarity — that satisfying click when everything seems to fall into place — functions as a *thought-terminating heuristic*. We use the sense that a matter has become clear as a signal that we've inquired enough and can stop. Usually fine. But it means clarity can be *weaponized*: a manipulator who can manufacture an exaggerated sense of clarity — a tidy ideology that explains everything, a conspiracy theory where every fact slots satisfyingly into place — can get you to *halt your inquiry early*, before you notice the holes. Notice how this snaps together with §1: clarity is dangerous precisely because it *terminates the zetetic process*. The slickest, most "it all makes sense now" account is, for that very reason, the one to interrogate hardest. The third move completes the toolkit: in "Trust as an Unquestioning Attitude" (*Oxford Studies in Epistemology*, 2022), Nguyen analyzes trust itself as the stance of *not questioning* — of taking something as a settled background you build on without re-checking. Indispensable (you can't re-derive everything from scratch), and exactly therefore exploitable: capture what someone trusts unquestioningly, and you've captured where they'll never think to look.

The honest tag here is twofold. The *conceptual* contributions — bubble vs. chamber, clarity as inquiry-terminating, trust as unquestioning — have been rapidly and widely adopted because they're genuinely clarifying and action-guiding. But two cautions earn their chips. First, philosophers are already pushing on the construct itself (Carey & Ventham, "There is no fresh air: a problem with the concept of echo chambers," *Episteme*, 2025). Second — and this is a hype-filter point the course insists on — the *empirical* social-science picture of how prevalent real-world echo chambers actually are is genuinely mixed; several large studies find most people's media diets are more varied than the "sealed echo chamber" image suggests. So treat the *conceptual machinery* as a sharp and durable tool, and the *empirical scale* of the phenomenon as an open measurement question. The framework is the contribution; the size of the fire it describes is still being measured.

§6 THE FORMAL REFOUNDATION

Re-deriving Day 1's Dutch book — from truth, not money

[ESTABLISHED][CONTESTED]

On Day 1 we justified the laws of probability with a *bribe*. The Dutch book theorem showed that if your credences break the probability rules, a clever bookie can sell you a set of bets you each accept as fair but which together guarantee you lose money. Powerful — but faintly unsatisfying as *epistemology*. Who cares about money? Shouldn't a *belief* be irrational for some reason to do with *truth*, not with your wallet? A research program that matured

through the 2010s and is in full bloom now — *accuracy-first epistemology*, also called epistemic utility theory — answers exactly that, and it's one of the most elegant results in the modern subject.

The idea (seeded by James Joyce's 1998 "Nonpragmatic Vindication of Probabilism" and built out by Richard Pettigrew's *Accuracy and the Laws of Credence*, 2016, with a wave of 2020–2023 papers refining and contesting it) is to measure how good a set of credences is by a single epistemic yardstick: *accuracy*, its closeness to the truth. Full confidence in a truth is perfectly accurate; full confidence in a falsehood, maximally *in*accurate. Now the theorem. For any *incoherent* credence — one that violates the probability laws — there is guaranteed to exist a *coherent* credence that is **more accurate in every possible world at once**. The incoherent one is, in the technical term, *accuracy-dominated*: strictly beaten on truth-closeness no matter how things turn out. So you don't need the bookie at all. Incoherent confidence is irrational for a purely *epistemic* reason — it leaves guaranteed accuracy on the table; there's a better-aimed set of credences available that's closer to the truth whatever the world does. The interactive lets you see the domination geometrically.

Accuracy domination, as credence geometry

CREDENCES	SUM	GEOMETRY	VERDICT
P(S)=0.50, P(not-S)=0.50	1.00	On the coherence line.	Undominated: no other credence is closer in every world.
P(S)=0.80, P(not-S)=0.80	1.60	Above the coherence line.	Dominated by a coherent projection closer to both truth-corners.
P(S)=0.20, P(not-S)=0.20	0.40	Below the coherence line.	Dominated by a coherent projection closer to both truth-corners.

What makes this a frontier rather than a footnote: it's an attempt to rebuild the *foundations* of rationality on a single epistemic value — getting close to the truth — and to derive not just probabilism but the update rule (conditionalization) and more besides from accuracy-dominance arguments. If it fully succeeds, the whole Bayesian edifice we started sketching

on Day 1 rests on truth, not on betting behavior or psychology. The chip, though, is earned. The *core* dominance theorem for probabilism is established mathematics. The *ambition* — that all epistemic norms fall out of accuracy alone — is contested: the cleanest results lean on technical assumptions (additivity, finitely many propositions) that critics argue smuggle in more than pure "closeness to truth" warrants (Chad Marxen, "Epistemic utility theory's difficult future," *Synthese*, 2021), and rival measures of accuracy can deliver different verdicts. So: a beautiful, genuinely illuminating reframing with a rock-solid center and a contested perimeter — which is, fittingly, the exact shape of this entire appendix.

◆ THE FRONTIER IN THREE SENTENCES

BIG IDEA

Since 2020 the question "what is knowledge?" has been pushed from five directions at once — reconceiving epistemology as the study of *inquiry* not belief (zetetic), reordering the mind so knowledge comes *before* belief, and confronting machines, deepfakes, and engineered information environments that strain or attack the very notion of a knower — while a formal program quietly rebuilds rationality's foundations on *truth* itself.

BEST NEW ANALOGY

The epistemic *backstop*: testimony has been kept honest all along by a support beam no one was looking at — the mere possibility of a recording — and deepfakes saw through it; pair it with the echo chamber, where the obvious fix (show them the other side) is precisely what makes the trap stronger.

LIVE CONTROVERSY

Every item here is genuinely unsettled — whether inquiry-norms swallow belief-norms, whether knowledge-representation is really more basic than belief, whether "bullshit" is the right word for what LLMs do, whether deepfakes bring collapse or just friction, and whether accuracy alone can ground all of epistemic rationality — which is exactly why each wears a hype-filter tag.

THREADS HERE › information (testimony's hidden backstop; LLMs as truth-indifferent text engines; accuracy as the epistemic good) · computation (the mind's factive theory-of-mind; epistemic utility as decision theory for belief) · evolution (why a social species evolves to track *knowledge* first). The five threads, now at the waterline.

OPEN QUESTIONS

What the edge of the map leaves blank

- **Is inquiry the real unit?** Do the norms of seeking truly conflict with the norms of believing — and if so, which is fundamental?
- **Knowledge or belief first?** Is "factive theory of mind" the basic cognitive tool, with belief a later, costlier add-on — or is the knowledge/belief line itself too crisp?
- **What *do* machines produce?** Knowledge, assertion, testimony, instrument-readings, or a genuinely new category of truth-apt text with no one home who cares?
- **Friction or collapse?** Do deepfakes merely add cost to verifying recordings, or dissolve a load-bearing condition of public knowledge?
- **How engineered is your mind's environment** — and how big, really, are the echo chambers we can now so clearly *describe*?
- **Can truth alone ground rationality?** Does accuracy-first reach all the way, or only as far as its technical assumptions carry it?
- **And a quieter contender for a future day:** several of these point past knowledge toward *understanding* as the thing we actually prize — a pivot we'll feel again whenever a model can predict without explaining.

SOURCES · ALL 2020+ UNLESS A FOUNDATIONAL ANCHOR

Sources & further reading

1. Friedman, J. (2020). "The Epistemic and the Zetetic." *The Philosophical Review* 129(4): 501–536. doi:10.1215/00318108-8540918. link See also Falbo, A. (2023), "Should epistemology take the zetetic turn?" *Philosophical Studies* 180(10–11): 2977–3002; Flores, C. & Woodard, E. (2023), "Epistemic norms on evidence-gathering," *Philosophical Studies* 180(9): 2547–2571.

2. Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., Santos, L. & Knobe, J. (2021). "Knowledge before belief." *Behavioral and Brain Sciences* 44: e140. doi:10.1017/S0140525X20000618 (target article + ~30 peer commentaries, several dissenting). link

3. Hicks, M. T., Humphries, J. & Slater, J. (2024). "ChatGPT is bullshit." *Ethics and Information Technology* 26: 38. doi:10.1007/s10676-024-09775-5. link Anchor: Frankfurt, H. (2005), *On Bullshit* (Princeton UP). Replies: Fisher, S. A. (2024), "Large language models and their big bullshit potential," *Ethics and*

Information Technology 26; Gunkel, D. & Coghlan, S. (2025), "Cut the crap: a critical response to 'ChatGPT is bullshit'," *Ethics and Information Technology* 27.

4. Rini, R. (2020). "Deepfakes and the Epistemic Backstop." *Philosophers' Imprint* 20(24): 1–16. link And Fallis, D. (2021). "The Epistemic Threat of Deepfakes." *Philosophy & Technology* 34(4): 623–643. doi:10.1007/s13347-020-00419-2.

5. Habgood-Coote, J. (2023). "Deepfakes and the epistemic apocalypse." *Synthese* 201(3). And Atencia-Linares, P. & Artiga, M. (2022). "Deepfakes, shallow epistemic graves: On the epistemic robustness of photography and videos in the era of deepfakes." *Synthese* 200(6). – the principal skeptical replies to the "collapse" framing.

6. Nguyen, C. T. (2020). "Echo Chambers and Epistemic Bubbles." *Episteme* 17(2): 141–161. doi:10.1017/epi.2018.32. link

7. Nguyen, C. T. (2021). "The Seductions of Clarity." *Royal Institute of Philosophy Supplement* 89: 227–255. And Nguyen, C. T. (2022). "Trust as an Unquestioning Attitude." *Oxford Studies in Epistemology* 7: 214–244. See also Nguyen (2023), "Hostile Epistemology," *Social Philosophy Today* 39: 9–32; and the critique Carey, B. & Ventham, E. (2025), "There is no fresh air: A problem with the concept of echo chambers," *Episteme* First View. doi:10.1017/epi.2024.43.

8. Pettigrew, R. (2016). *Accuracy and the Laws of Credence.* Oxford University Press. Foundational anchor: Joyce, J. M. (1998), "A Nonpragmatic Vindication of Probabilism," *Philosophy of Science* 65(4): 575–603. Recent development & critique: Pettigrew, R. (2022), "Accuracy-First Epistemology Without Additivity," *Philosophy of Science* 89(1): 128–151; Marxen, C. (2021), "Epistemic utility theory's difficult future," *Synthese* 199(3–4): 7401–7421. Survey: *SEP*, "Epistemic Utility Arguments for Epistemic Norms."

Hype-filter note: classical anchors (Frankfurt 2005, Joyce 1998) are cited only as the roots of the post-2020 work that is this appendix's actual subject. No claim above should be treated as settled; that is the point of the chips.

TOMORROW → DAY 02

The Scientific Method & Demarcation

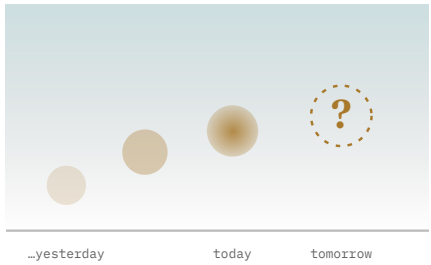
Today we asked when a *single* belief counts as knowledge. Tomorrow we scale the question up to an entire institution: how does science decide which claims even get to enter the arena? Popper's demand that a real theory be *falsifiable*, Kuhn's paradigm shifts, Lakatos's research programmes — and the 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

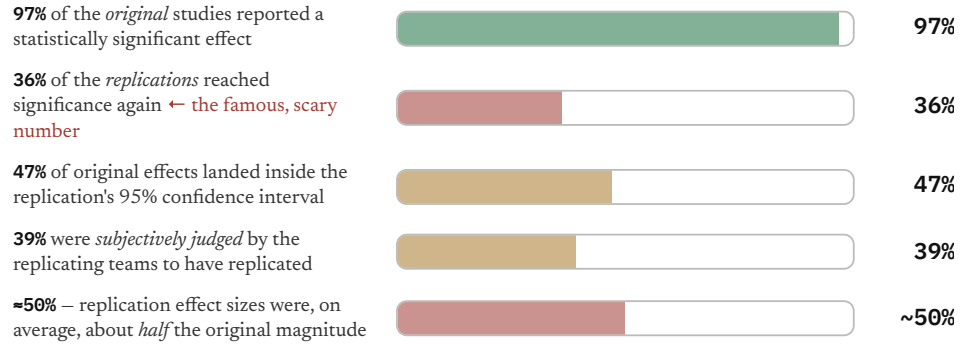
The replication crisis: demarcation under live fire

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.

OPTIONAL APPENDIX

Appendix: Foundations Without Bedrock

This section is optional supplemental reading. You can skip it without losing the main lesson.

We kept saying falsify, test, observe. Now we go down a level — and find there's nothing solid underneath.

The main descent gave you the tour: Hume's hole, Popper's escape, Kuhn's mess, Lakatos's repair, and the replication crisis testing the whole quarrel under live fire. This appendix takes the same building and walks you into the basement — past the floorboards, to look at the foundations. And the discovery waiting down there, made over and over by very different people, is strangely consistent: **there are no foundations.** No theory-neutral observation to settle disputes. No non-circular justification for expecting tomorrow. No purely logical algorithm that stamps a claim "science." Just piles driven into a swamp, deep enough to hold for now.

This continues directly from the main Day 2 lesson, which ended on the replication crisis and the question *"is science failing, or working as designed?"* Here we deepen four things we waved at in passing: (1) what Hume's problem becomes once you take it seriously — and the *worse* riddle hiding behind it; (2) the cracks in Popper's own machinery he honestly admitted; (3) the deep reason a "neutral test" may not exist; and (4) the actual *mathematics* that makes most published findings inflated. Keep the calibration instinct from **Day 1** close — by the end you'll see exactly why it's the only safe attitude.

PART 1 · THE HOLE GETS DEEPER

Hume answers his own riddle — then Goodman makes it worse

We left Hume having argued that nothing non-circularly justifies our faith in the sunrise. But Hume didn't actually stop there, and the part the textbooks skip is the most human bit. Having shown that *reason* can't ground induction, he asked the obvious follow-up: so why do we do it anyway, every second of every day, without falling apart? His answer is almost

tender. We infer by *custom* — by habit. Burned once, the child fears the flame; it isn't deduction, it's the worn groove of repeated experience:

> ***Having found, in many instances, that any two kinds of objects... have always been conjoined together; if flame or snow be presented anew to the senses, the mind is carried by custom to expect heat or cold... This belief is the necessary result of placing the mind in such circumstances.***
>
> — Hume, Enquiry, §V (1748)

This is the move worth naming, because it recurs through the whole course. Hume splits one question into two. There's the **justificatory** problem (can induction be *deductively* or non-circularly proven? — no, and that wound never heals) and the **descriptive** problem (why do minds infer anyway? — because we're built to, by custom). He surrenders the first and answers the second. We are not reasoning machines that happen to have instincts; we are instinct-machines that have learned to dress our habits in the language of reason. (You'll feel this exact split again on **Day 11**, heuristics and biases, and **Day 119**, the predictive brain.)

Four ways out of the hole

For two and a half centuries, philosophers have tried to climb out of Hume's pit. None has fully succeeded — but the attempts are gorgeous, and each is a different temperament made into an argument.

Strawson

DISSOLVE THE QUESTION

To ask "is induction rational?" is confused. Reasoning *well* just *means*, in part, proportioning belief to evidence inductively. Demanding an outside stamp of approval is like asking whether the law is legal. There's no question left to answer.

Reichenbach

MAKE THE PRAGMATIC BET

We can't prove induction works — but we can show it's the *best bet available*. If *any* method can track nature's regularities, induction will eventually find them. It can't do worse than the alternatives, so use it. Justified as a means, not as a truth.

Popper

DENY THE PREMISE

His radical claim: there *is* no induction. Science never generalizes from instances; it conjectures boldly and tries to refute. With no inductive step in the method, Hume's problem simply has nothing to bite on. (Critics: but then science can never tell us a theory is *reliable* for prediction — which we plainly need.)

Bayes

QUANTIFY THE UPDATING

Treat learning as revising *degrees of belief* by Bayes's theorem — the credence dial from Day 1. This beautifully *formalizes* learning from evidence, but it doesn't slay Hume: the priors and the updating rule themselves still need grounding. (Picked up properly on **Day 4**.)

And just when you think the worst is behind you, a Harvard logician named **Nelson Goodman** stands up in 1955 and detonates a *second* bomb — one that goes off even if you grant that induction works perfectly. It's called the *new riddle of induction*, and its weapon is a single nonsense word.

The gem that turns blue: meet "grue"

Define a new color predicate, *grue*. An object is grue if it has been examined before some future date — say, January 1, 2050 — and found **green**; or else it has *not* been examined by then and is **blue**. Strange, artificial, useless. But watch what it does.

Every emerald ever examined has been green. So every emerald ever examined is also, by definition, *grue* (examined before 2050, and green). Which means your mountain of evidence supports **both** of these with exactly equal force:

- **H1:** "All emeralds are green." → predicts the next emerald you dig up in 2051 is green.
- **H2:** "All emeralds are grue." → predicts the next emerald you dig up in 2051 is *blue*.

The evidence cannot choose between them, because *every observation confirms both equally*. Induction, even granting it works, doesn't tell you which regularity you're allowed to project into the future. Play with it below — drag your observation horizon and watch the two theories sit in perfect agreement right up until the moment they violently disagree.

Green vs. Grue, as a projection table

PERIOD	OBSERVED EVIDENCE	"ALL GREEN" PREDICTS	"ALL GRUE" PREDICTS	LESSON
Before 2050	Every examined emerald is green.	Green emeralds.	Green emeralds.	The evidence confirms both descriptions equally.
After 2050	New observations finally enter the divergent region.	Green emeralds.	Blue emeralds.	Reality can break the tie only after the cutoff is crossed.
Goodman's point	Past regularity alone does not choose a projectible predicate.	Project green.	Project grue.	Induction needs background habits about which predicates are natural or entrenched.

The obvious objection — "but *grue* is gerrymandered nonsense, *green* is natural!" — is exactly the trap. Goodman's needle: from *inside* the grue-language, it's *green* that looks weird. Define "bleen" (blue-before-t-or-green-after) and you can define plain old "green" as "grue-before-t-or-bleen-after" — green becomes the funny-looking compound, and grue the simple primitive. There's no view from nowhere that crowns green the natural one. Goodman's own escape was to say we project the predicates that are *entrenched* — the ones our language has used successfully many times before. Which is honest, and also slightly deflating: it grounds the lawfulness of nature not in nature but in the contingent habits of human vocabulary. Hume said our *inferences* rest on custom; Goodman says even the *concepts* we infer with do too. The hole, it turns out, has a basement of its own. [DEBATE]

PART 2 · THE CRACKS POPPER ADMITTED

Falsification, looked at closely

In the main lesson Popper was the hero with the clean rule. He was also, to his great credit, his own most honest critic — and three subtleties he conceded matter enormously for everything downstream.

First: demarcation is not about meaning

Popper is constantly confused with the *logical positivists* of the Vienna Circle (Schlick, Carnap, and their English megaphone A.J. Ayer, whose *Language, Truth and Logic* landed in 1936). The positivists had their own famous criterion — the *verifiability theory of meaning*: a statement is *meaningful* only if it can be empirically verified (or is true by definition). Everything else — metaphysics, theology, ethics — is not false but literally *nonsense*, "pseudo-statements." It was a wood-chipper for whole branches of philosophy.

Popper thought this was both arrogant and self-defeating (the verifiability criterion isn't itself verifiable, so by its own rule it's nonsense). His point was sharper and more modest. Falsifiability sorts the **scientific** from the **non-scientific** — but it says *nothing* about meaning. Unfalsifiable claims can be perfectly meaningful, often profound, sometimes the seeds of future science. "Every material body is attracted by every other" was untestable metaphysics long before it was Newton. Demarcation draws a line on a map; it does not burn down the other country. Forgetting this turns Popper into a philistine he explicitly refused to be.

Second: the boldest theory is the *least* probable — and that's the point

Here's a delicious inversion of common sense. We tend to admire a "safe" theory that fits the data snugly. Popper admired the opposite. The *more* a theory forbids — the more ways the world could prove it wrong — the higher its *empirical content*, and the *lower* its probability of being true by chance. "Einstein's light bends by exactly 1.75″" is a tightrope; "the economy is shaped by many factors" is a sofa. A theory can be highly probable precisely *because* it says almost nothing. So Popper flipped the prize: science should seek **bold, improbable, high-content** conjectures and expose them to brutal tests. Probability is what cowards optimize. Testability is what science optimizes. (Hold this thought — it sets a genuine tension with the Bayesian, probability-maximizing picture we meet on **Day 4**.)

Third: there is no bedrock — only piles in a swamp

This is the crack that gives this whole appendix its title, and it's the one quick summaries of Popper often skip. A falsification needs a fact to do the falsifying — a "basic statement," an observation report like *"the needle points to 1.75."* But where do those come from? Not from pure, theory-free looking. Every observation is shot through with assumptions (that the instrument works, that light behaves, that "needle" and "point" carve the world correctly). So basic statements aren't *given* by nature; they're *accepted* — by agreement, by decision, provisionally. Popper said so himself, in the most beautiful passage he ever wrote:

> *The empirical basis of objective science has thus nothing 'absolute' about it. Science does not rest upon solid bedrock. The bold structure of its theories rises, as it were, above a swamp... The piles are driven down... but not down to any natural or 'given' base; and if we stop driving the piles deeper, it is not because we have reached firm ground. We simply stop when we are satisfied that the piles are firm enough to carry the structure, at least for the time being.*
>
> — Popper, The Logic of Scientific Discovery (1959)

Sit with what this costs him. If the facts that do the falsifying are themselves accepted by convention, then falsification is never the clean, absolute guillotine the slogan promises. A scientist *could* always reject the basic statement instead of the theory ("the instrument was faulty"). Popper's defense was a *methodological* one: agree, as a rule of the game, not to wriggle out with ad hoc rescues — not to keep re-driving the piles wherever it's convenient. Which is reasonable. But notice it's a *rule we choose*, not a fact we discover — uncomfortably close to the communal judgment Popper disliked in Kuhn's picture of normal science. The swamp swallows a little more certainty than the textbook version admits.

CORROBORATION IS NOT A DOWN PAYMENT ON TRUTH

One more Popperian fine print, because people get it wrong constantly. When a theory survives a savage test, Popper says it is *corroborated* — but corroboration is emphatically **not** a probability, and a much-tested theory does *not* become "probably true." It's just a report card on how severe a beating the theory has taken and survived, valid only "for the time being." Hilary Putnam pressed the obvious objection: if science never licenses calling any theory probable or reliable, how can we possibly justify *using* our best theories to build bridges and send probes to Mars? We clearly do rely on them. Popper's austere answer — rely provisionally on what has survived severe tests, without treating it as probable truth — many find too cold to be the whole story.

PART 3 · THE MISSING NEUTRAL GROUND

You can't even see the same sunrise

Popper's swamp suggested observation isn't bedrock. A philosopher-physicist named **Norwood Russell Hanson** pushed the knife further in *Patterns of Discovery* (1958) with a phrase that became a slogan: observation is *theory-laden*. There is, he said, "more to seeing than meets the eyeball." What you perceive is already shaped by what you believe.

His thought experiment is unforgettable. Put Tycho Brahe, who believes the Earth stands still, and Johannes Kepler, who believes it spins, on a hill at dawn. The same photons strike the same retinas; a camera would record identical images. And yet — do they see the same thing? Tycho sees the *Sun moving up* from a fixed horizon. Kepler sees a *fixed Sun* revealed as the horizon rolls *down* away from it. The raw sensation may be shared, but the seeing — the meaningful, conceptual act of seeing-as — is structured by theory all the way down.

Same photons, same retinas — two different sunrises. If observation is theory-laden, there is no neutral umpire to settle a clash of theories.

This is the quiet land-mine under the whole idea of a decisive experiment. The falsificationist picture needs a neutral observation language — facts both sides accept — to serve as referee between rival theories. Hanson (and then Kuhn, with his duck-rabbit and his student who sees "confused broken lines" where the physicist sees "a record of familiar subnuclear events") suggests the referee may be compromised before the match begins, quietly wearing one team's colors. *(Fairness check: Hanson himself admitted "something" in the two dawn experiences "is the same for both," so the strong claim — that they literally see different things — is genuinely contested. A minimal version is safe; the maximal version is a fight.* `[CONTESTED]`*)*

Quine pulls the thread, and the whole sweater moves

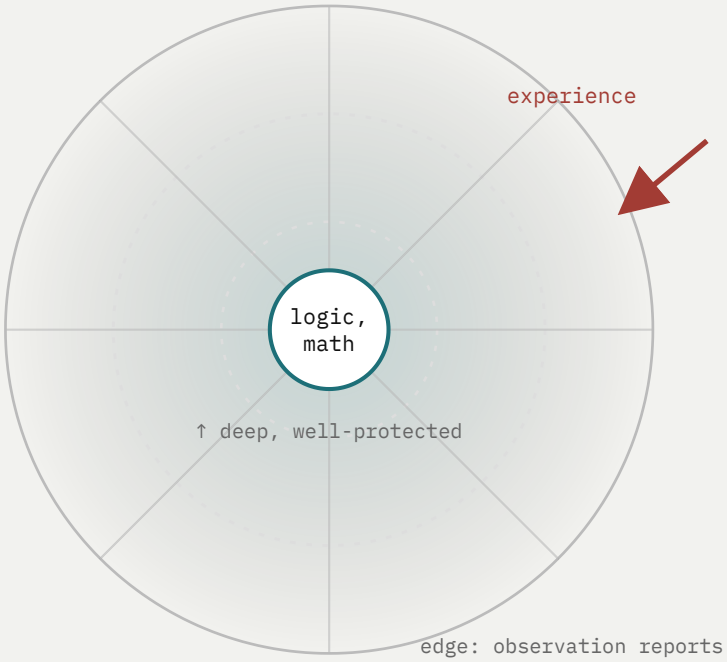
If single observations are theory-laden, the philosopher **W.V.O. Quine** showed in 1951 that single *tests* are theory-laden too — and turned it into a deeply influential paper in modern philosophy, "Two Dogmas of Empiricism." We met its child in the main lesson (the Duhem–Quine thesis: no hypothesis is tested alone). Here is the parent idea in full, and it's wilder. Quine pictured all of human knowledge — from "there's a cup here" to the laws of logic — as a single vast *web of belief*:

> *The totality of our so-called knowledge or beliefs... is a man-made fabric which impinges on experience only along the edges... total science is like a field of force whose boundary conditions are experience.*
>
> — Quine, "Two Dogmas of Empiricism" (1951)

Experience only ever touches the *edges* of the web. When a clash comes — a prediction fails — the shock propagates inward, but *you choose where to absorb it*. You can always protect any belief you like, however deep, by making adjustments elsewhere. Quine's two scandalous conclusions: experience meets our beliefs "not individually but only as a corporate body," and therefore —

> *Any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system... Conversely, by the same token, no statement is immune to revision.*
>
> — Quine (1951)

No statement is immune — not even logic or mathematics. (Quine noted that revising the law of the excluded middle had been floated to simplify quantum mechanics.) There is no privileged core of certainties; there's only a web held taut by experience at the rim and by our preference not to rip up more than we must. Which is the deepest version yet of "no bedrock": not even the laws of thought are nailed down.



Quine's web: shocks land at the rim and ripple in, but you decide what gives. The center can always be saved — at a price paid elsewhere.

Laudan slams the brakes: possible ≠ reasonable

If you've been feeling the floor tilt toward "so anything goes, it's all just choice" — good, because that's the abyss, and **Larry Laudan** (yes, the same demolition man from the main lesson) is the one who pulls everyone back from it. In "Demystifying Underdetermination" (1990), he argues the dramatic conclusions people draw from Quine are smuggled in by a single bad equation: treating *logically possible* as if it meant *rationally reasonable*.

Yes, Laudan concedes, pure deductive logic never *forces* a unique theory choice — you *can* always save a belief "come what may." But science was never running on deductive logic alone. It runs on logic *plus* a thick fabric of ampliative standards — simplicity, fruitfulness, consistency with established results, predictive track record. Quoting Duhem approvingly: "Pure logic is not the only rule for our judgments." That you *could* blame the telescope instead of the theory doesn't make it *reasonable* to; that you *could* hold the Earth flat by adding enough epicycles of excuse doesn't make it a live option for a sane inquirer. The web has no logical bedrock — but it has rational *tension*, and that tension is enough to do real

work. Underdetermination is true and mostly toothless. It's the difference between "I can't prove with certainty you're not a brain in a vat" and "therefore all bets are off." The first is correct; the second doesn't follow. [REVIEW]

PART 4 · A FAIRER TRIAL FOR FREUD

Grünbaum: psychoanalysis isn't un-science — it's failed science

In the main lesson we flagged that Popper may have caricatured Freud. The philosopher who turned that hunch into a forensic case was **Adolf Grünbaum**, in *The Foundations of Psychoanalysis* (1984) — and his verdict is far more interesting, and more damning, than Popper's.

Popper said psychoanalysis was *unfalsifiable* — it explained everything, forbade nothing, so it never even entered the arena of science. Grünbaum said: nonsense, and not in Freud's favor. Freud's theory *does* make testable claims. If repressed homosexuality is a *necessary cause* of paranoia, then a society that grows more tolerant of homosexuality should see paranoia decline — a real, checkable prediction. More centrally, Grünbaum excavated what he called Freud's *Tally Argument* (from Freud's 1917 lectures): Freud defended his method by claiming that *only* psychoanalytic interpretations that "tally with what is real" in the patient can produce a durable cure — so lasting therapeutic success would *vindicate* the interpretations.

That's a genuine scientific bet. On Grünbaum's reading, it *loses*. Durable remission happens through other therapies and through spontaneous remission with no analysis at all — so therapeutic success can't certify Freudian interpretations as uniquely correct. He also argued that the "evidence from the couch" is contaminated by the analyst's own suggestion: patients can oblige their analysts by producing the very memories and associations the theory predicts. So the data can't bear the causal weight Freud put on it. Grünbaum's conclusion reframes the whole demarcation question: psychoanalysis is not *non*-science safely quarantined outside the arena — it's **science that stepped into the ring and got knocked out**. Bad science, not non-science. (A genuinely different and arguably more respectful verdict: it takes Freud seriously enough to test him. [CONTESTED]) This distinction — *unfalsifiable* vs. *falsified* — is one you'll want in your pocket for every "is X a science?" fight to come.

PART 5 · THE ENGINE ROOM OF THE CRISIS

Why most findings are inflated: the actual math

The main lesson showed you the wreckage — 36% of psychology replications reaching statistical significance again, effects halving, power posing collapsing. It didn't show you the *machine* that can produce wreckage on that scale. The machine is not necessarily fraud. It's arithmetic, and once you see it you can't unsee it. Three gears mesh: **base rates**, **flexibility**, and **filtering**.

Gear one: the base-rate trap (Ioannidis's bombshell)

In 2005 the physician-statistician **John Ioannidis** published one of the most downloaded and most argued-over papers in the history of *PLoS Medicine*, with a title engineered to detonate: *"Why Most Published Research Findings Are False."* His argument isn't rhetoric; it's a formula. The thing we actually care about is the *positive predictive value* (PPV): given that a study reported a "significant" effect, what's the probability the effect is *real*? It depends on three numbers — the significance threshold α (conventionally 0.05), the study's statistical power (its chance of catching a real effect), and, crucially, the *pre-study odds* R: among all the hypotheses a field tests, what fraction are actually true?

That last number is the killer, and it's the one researchers forget. Here's the intuition, in dots. Suppose a field tests 1,000 hypotheses, of which only 100 are really true (because good ideas are rare and most guesses are wrong). Run them all at 80% power and the standard 5% threshold. You'll correctly flag about 80 of the 100 true effects. But among the 900 *false* hypotheses, the 5% false-positive rate hands you about 45 "significant" results that are pure noise. So of ~125 findings you'd publish as discoveries, ~45 — more than a third — are false. And that's the *rosy* case. Drop the power, or lower the fraction of true hypotheses, and the false discoveries swamp the real ones. The dial below lets you run Ioannidis's machine yourself.

The Discovery-Purity Engine, as base-rate scenarios

SCENARIO	TRUE HYPOTHESES	POWER	BIAS	PUBLISHED POSITIVES	PPV
Rosy baseline	100 of 1,000	80%	0%	80 true positives + 45 false positives	64% real
Low base rate	20 of 1,000	80%	0%	16 true positives + 49 false positives	25% real
Bias added	100 of 1,000	80%	20%	84 true positives + 216 false positives	28% real

Ioannidis's corollaries fall straight out of the machine, and they read like a map of where the replication crisis hit hardest: the smaller the studies, the smaller the true effects, the more analytical flexibility, the more financial interest, and the *hotter* the field (more teams racing the same question), the lower the chance any given published finding is true. It's not cynicism. It's the geometry of testing rare truths with imperfect instruments. [REVIEW]

It didn't go unchallenged, and the challenge is worth knowing. Statisticians **Steven Goodman and Sander Greenland** (2007) agreed with the broad moral but disputed the engineering: the model treats every significant *p* as if it were exactly 0.05 (throwing away evidence), bakes in its own bias parameters rather than measuring them, and the eye-catching "more teams → more falsehood" result is partly a modeling artifact. Ioannidis replied that the core stands and that even his own tables show findings can reach 85% credibility under good conditions. The honest takeaway: the *exact* false-positive rate of science is genuinely uncertain and field-dependent — but the *direction* of the argument, that low base rates plus low power can manufacture false positives, is hard to ignore. [CONTESTED]

Gear two: flexibility — how to find anything (the Beatles experiment)

The base-rate trap assumes honest 5% testing. Real research is leakier, and in 2011 three psychologists — **Simmons, Nelson, and Simonsohn** — demonstrated how leaky with one of

the great pieces of scientific theater. Their paper, "False-Positive Psychology," coined the phrase *researcher degrees of freedom*: all the small, innocent-looking choices a scientist makes along the way — when to stop collecting data, which outliers to drop, which control variables to include, which conditions to compare. Each choice is defensible. Together, they're a machine for manufacturing significance.

To prove it wasn't hypothetical, they ran a real experiment on real undergraduates and reported a real, statistically significant result: that listening to the Beatles' "When I'm Sixty-Four" **literally made people younger**. Not feel younger — *be* younger. After controlling for the participant's father's age, subjects who heard the song were calculated to be a year and a half younger in actual chronological age (adjusted mean 20.1 years) than those who heard a control track (21.5 years), $p = .04$. The effect is, of course, metaphysically impossible. That was the entire point. They got there using the ordinary flexibility the paper put on trial: choosing covariates, outcomes, comparisons, and stopping rules after seeing how the data are going. If you can prove a Beatles song reverses aging, you can prove anything. Their proposed cure — disclose every choice, ideally *before* you collect data — is the seed of the preregistration movement from the main lesson.

THE MOST UNSETTLING PART: YOU DON'T HAVE TO CHEAT

Andrew Gelman and Eric Loken gave this its sharpest form in 2013, the *garden of forking paths*. You might imagine p-hacking requires running 20 analyses and reporting the one that "worked." But suppose an honest researcher runs *only one* analysis and had the hypothesis in mind in advance — yet the *specific* test they chose was shaped by what the data happened to look like. Had the data come out differently, they'd have justifiably analyzed it differently. All those untaken paths still poison the *p*-value, because it silently assumes there was only ever one road. "The problem," they wrote, is that the many potential comparisons are "contingent on data" — so a perfectly sincere scientist, never consciously fishing, can still drift into a false positive. This is why good intentions don't save you, and why the reforms had to be *structural*.

Gear three: filtering — the literature is a survivor's gallery

The third gear was spotted earliest of all. Back in **1959**, Theodore Sterling noticed something damning about what gets *printed*. Surveying four major psychology journals, he found that of the articles using significance tests, **286 of 294 — a staggering 97.28%** — had rejected the null hypothesis and reported a positive result. And not one of the studies he surveyed was a replication. Journals print winners. Nulls die in the file drawer — a problem

Robert Rosenthal formalized in 1979 as the *file-drawer problem* (and quantified with a "fail-safe N": how many buried null results would it take to overturn a published effect?).

Stack the gears and the crisis is overdetermined. Most tested hypotheses are false (base rates) → flexibility inflates the false ones into "significance" (forking paths) → and only the significant ones ever see print (the file drawer), often re-skinned afterward as if predicted all along (a sin Norbert Kerr named *HARKing* in 1998 — Hypothesizing After the Results are Known, which quietly "translates Type I errors into theory"). The published literature isn't a map of what's true. It's a gallery of the lucky survivors of a brutal, invisible selection — a darkly perfect echo of the *evolution* thread, and of Day 1's Gettier worry: results that are "right," but for reasons that have nothing to do with the truth.

`The verdict from the statisticians` `[ESTABLISHED]`

What a *p*-value is not

In 2016, for the first time in its 177-year history, the **American Statistical Association** issued a formal public warning about a specific statistical practice — the *p*-value (Wasserstein & Lazar, *The American Statistician*). The fact that the field's central U.S. professional association broke its silence tells you how serious the problem had become. Its six principles are worth tattooing somewhere visible, because many misuses in the crisis violate one:

- A *p*-value measures how incompatible data are with a model — and nothing more.
- It does **not** give the probability that the hypothesis is true, nor the probability your result is "due to chance."
- Conclusions should never hinge on whether *p* crosses a "bright line" like 0.05.
- Proper inference demands full reporting and transparency (no hidden forking paths).
- A *p*-value says nothing about the *size* or importance of an effect.
- By itself, it is a poor measure of evidence for a hypothesis.

The most common confusion — that $p = 0.05$ means "95% chance the finding is real" — is flatly false, and the base-rate engine above is why: the probability a discovery is true depends overwhelmingly on how rare true hypotheses are, which the *p*-value never sees. A 2019 follow-up went further still, with some statisticians urging the field to retire the phrase "statistically significant" altogether. The reform isn't finished. `[REVIEW]`

PART 6 · THE DUEL THAT DEFINED A FIELD

London, July 1965: a famous fight in the philosophy of science

All four protagonists from the main lesson — Popper, Kuhn, Lakatos, Feyerabend — were not abstractions politely taking turns in a textbook. They were living rivals, and in July 1965 they (and others) collided in person at an international colloquium at Bedford College in London. The proceedings, delayed for years by the combatants' refusal to stop revising, finally appeared in 1970 as *Criticism and the Growth of Knowledge* — one of the most electric volumes in the field. It opens with Kuhn, is pelted with replies, and closes with Kuhn firing back.

The fault line was sharp. Popper accused Kuhn's "normal science" — heads-down puzzle-solving inside an unquestioned paradigm — of being not science at all but a kind of intellectual conformism, even "mob psychology": the very uncritical dogmatism falsification was meant to abolish. Kuhn shot back that Popper had mistaken the rare, thrilling revolutionary moments for the daily substance of science, which is overwhelmingly conservative and paradigm-bound — and that's a *feature*, the thing that lets a field accumulate deep results instead of forever relitigating its foundations.

TWENTY-ONE PARADIGMS IN ONE BOOK

The sharpest blow came from an unexpected quarter. The linguist **Margaret Masterman**, broadly sympathetic to Kuhn, sat down and counted the ways he used his own central word — and found Kuhn deploying "paradigm" in at least **21 distinct senses**, which she sorted into metaphysical, sociological, and concrete "artefact" types. Her assessment was a perfect double-edged sword: Kuhn's book was "at once scientifically perspicuous and philosophically obscure." It was a devastating critique and a vindication at once — the concept was muddled *and* it had clearly hit something real. Kuhn later conceded the point and spent much of his career trying to say more precisely what he had meant.

Two of Kuhn's deeper ideas deserve rescuing from the caricature, because both are routinely overstated:

- **Kuhn loss.** Scientific progress is not purely cumulative. When a paradigm falls, the successor can *lose* explanatory successes the old one had — phlogiston chemistry explained a few things early oxygen chemistry initially couldn't. Progress is real but

ragged; we trade one set of solved puzzles for a larger, different set, and sometimes drop a few on the way. *(Contested how much this threatens realism — most documented losses are anecdotal rather than quantitative.)*

- **The world-change thesis.** Kuhn's most notorious line is that after a revolution "the scientist afterward works in a different world." But read him exactly and he's careful — he writes "we may *want* to say" the world changes, hedging it as a way of speaking, not a flat claim that reality reshuffles itself. He spent his later years walking back the most radical reading, retreating to a narrow *taxonomic incommensurability* (only the interlocking technical vocabulary shifts, not whole realities) and insisting, against his relativist fans, that "the world is not invented or constructed." The Kuhn of legend is wilder than the Kuhn of the page.

And **Feyerabend**, the supposed wrecker, had a constructive heart underneath the provocation. His real proposal was *pluralism*: a healthy science should *maximize* the number of competing theories, not enforce consensus. Two slogans carry it. The *principle of proliferation*: actively invent and defend theories that contradict the reigning one. And *counterinduction*: deliberately develop ideas inconsistent with even well-confirmed facts — because, exactly as Hanson warned, observations are theory-laden, so the *only* way to expose the hidden assumptions baked into your current view is to look at the world through a rival lens. In later prefaces and replies, he stressed that "anything goes" was not a creed he preached but "the terrified exclamation of a rationalist who takes a closer look at history." His monster turns out to be an argument *for* intellectual diversity as the engine of discovery — which lands surprisingly close to where this whole appendix has been heading.

THE THROUGHLINE

No bottom, and it works anyway

Stand back and the whole appendix is one note held for a long time. Hume: no logical justification for expecting tomorrow. Goodman: not even our concepts are safe from the same rot. Popper, honestly: the facts that falsify rest on convention, on piles in a swamp. Hanson: even what you *see* is bent by theory. Quine: the entire web, logic included, floats — nothing is immune to revision. And the replication crisis is that abstraction made horribly concrete: when you actually audit some literatures, a third or more of high-profile findings fail strict replication tests, exactly as the math of base rates and forking paths predicts can happen.

You'd be forgiven for expecting the moral to be despair. It's the opposite, and Laudan handed us the key: *logically possible* is not *reasonable*. Science has no foundations and needs

none. It works the way a city works — no single unmovable stone at the bottom, just countless mutually supporting structures, constantly inspected, occasionally condemned and rebuilt, the whole thing standing not because it rests on rock but because it keeps correcting itself faster than it crumbles. The replication crisis isn't the swamp swallowing science. It's science driving fresh piles, in public, having noticed the old ones were getting soft. That's not the failure of the method. *That is the method.*

Which is why the only sane posture for the next 178 days is the one we built on Day 1: hold every belief by the dial, not the switch. Proportion your confidence to the evidence, keep a little aside for being wrong, and treat the splashiest claim with the most suspicion. There's no bedrock under any of it. Learn to build on piles.

◆ THE APPENDIX IN THREE SENTENCES

BIG IDEA

Dig beneath the scientific method and you find no foundations — no non-circular justification for induction (Hume), no safe concepts (Goodman's grue), no theory-neutral observation (Hanson), no belief immune to revision (Quine), only Popper's "piles driven into a swamp" — and the replication crisis is the empirical warning sign, with a mathematical engine (base rates × flexibility × filtering) behind it.

BEST ANALOGY

The building on piles in a bottomless swamp — driven down "only until firm enough, for the time being" — paired with the Beatles song that "proved" listeners grew younger, the demonstration that ordinary flexibility can manufacture any result.

LIVE CONTROVERSY

Whether foundationlessness collapses into "anything goes" (Quine's web) or is tamed by reasoned standards (Laudan: logically possible ≠ rationally reasonable) — and, empirically, the true false-positive rate of science (Ioannidis vs. Goodman & Greenland), still unsettled and field-dependent.

THREADS HERE › information (the *p*-value, base rates, and what evidence can and can't carry) · evolution (the literature as a survivor's gallery of lucky positives) · computation & emergence (science as a self-correcting system with no central foundation, holding itself up by mutual tension) — extending the main Day 2 threads one level down.

SOURCES

Sources & further reading

1. Hume, D. (1748). *An Enquiry Concerning Human Understanding*, §IV–V. — the "sceptical solution"; custom/habit as the basis of inference. See SEP, "The Problem of Induction."

2. Goodman, N. (1955). *Fact, Fiction, and Forecast.* Harvard University Press. — the new riddle of induction ("grue"); projectibility and entrenchment. See SEP, "Nelson Goodman."

3. Strawson, P. F. (1952). *Introduction to Logical Theory*, ch. 9 — the "dissolution" of the problem of induction. Reichenbach, H. (1938). *Experience and Prediction* — the pragmatic vindication.

4. Ayer, A. J. (1936). *Language, Truth and Logic.* — the English-language popularization of logical positivism and verificationism. See SEP, "Logical Empiricism" and SEP, "Alfred Jules Ayer."

5. Popper, K. (1959). *The Logic of Scientific Discovery* (orig. 1934). — degrees of falsifiability; the "piles into a swamp" passage (§30); corroboration ≠ probability; demarcation ≠ meaning. See SEP, "Karl Popper."

6. Putnam, H. (1974). "The 'Corroboration' of Theories," in *The Philosophy of Karl Popper.* — the objection that Popper leaves science unable to justify reliance on theories.

7. Hanson, N. R. (1958). *Patterns of Discovery.* Cambridge University Press. — theory-ladenness of observation; Tycho vs. Kepler at dawn.

8. Quine, W. V. O. (1951). "Two Dogmas of Empiricism." *The Philosophical Review* 60(1): 20–43. — the web of belief; "no statement is immune to revision"; confirmation holism. full text

9. Laudan, L. (1990). "Demystifying Underdetermination," in *Minnesota Studies in the Philosophy of Science* 14: 267–297. — logically possible ≠ rationally reasonable; the limits of underdetermination. See SEP, "Underdetermination."

10. Grünbaum, A. (1984). *The Foundations of Psychoanalysis: A Philosophical Critique.* University of California Press. — the Tally Argument; psychoanalysis as falsifiable-but-failed (bad science, not non-science).

11. Ioannidis, J. P. A. (2005). "Why most published research findings are false." *PLoS Medicine* 2(8): e124. — the PPV model; pre-study odds, power, bias. plos.org

12. Goodman, S. & Greenland, S. (2007). "Why most published research findings are false: problems in the analysis." *PLoS Medicine* 4(4): e168 — the main statistical critique; with Ioannidis's reply (e215).

13. Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011). "False-Positive Psychology." *Psychological Science* 22(11): 1359–1366. — researcher degrees of freedom; the "When I'm Sixty-Four" demonstration ($p = .04$).

14. Gelman, A. & Loken, E. (2014). "The Statistical Crisis in Science" ("The garden of forking paths," 2013 working paper). *American Scientist* 102(6): 460. — false positives without conscious p-hacking. PDF

15. Kerr, N. L. (1998). "HARKing: Hypothesizing After the Results are Known." *Personality and Social Psychology Review* 2(3): 196–217.

16. Sterling, T. D. (1959). "Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—Or Vice Versa." *JASA* 54(285): 30–34. – 286 of 294 (97.28%) significance-test articles rejected the null; none were replications.

17. Rosenthal, R. (1979). "The file drawer problem and tolerance for null results." *Psychological Bulletin* 86(3): 638–641. – publication bias; the "fail-safe N."

18. Wasserstein, R. L. & Lazar, N. A. (2016). "The ASA Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70(2): 129–133. – the six principles; the 2019 follow-up urged retiring "statistical significance." tandfonline

19. Lakatos, I. & Musgrave, A. (eds.) (1970). *Criticism and the Growth of Knowledge.* Cambridge University Press. – proceedings of the 1965 Bedford College colloquium; includes Kuhn, Popper, Lakatos, Feyerabend, and Masterman's "The Nature of a Paradigm" (the 21 senses).

20. Kuhn, T. S. (1962/1970). *The Structure of Scientific Revolutions*, ch. X & Postscript. – Kuhn loss; the world-change thesis ("we may want to say..."); later taxonomic incommensurability. See SEP, "Incommensurability."

21. Feyerabend, P. (1975). *Against Method.* – pluralism, proliferation, counterinduction; "anything goes" as the "terrified exclamation of a rationalist." See SEP, "Paul Feyerabend."

TOMORROW → DAY 03

Logic & Valid Inference

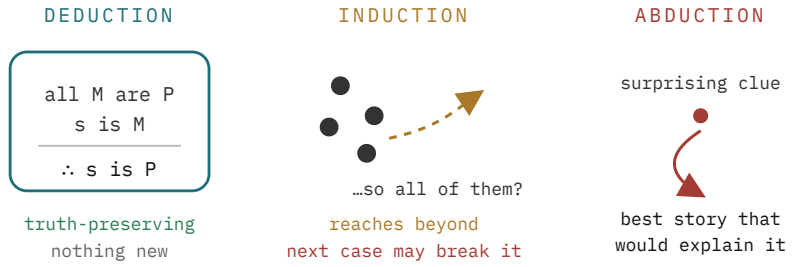
Today we leaned hard on words like "valid," "follows from," and "contradiction" — but what *are* the rules that make an argument actually hold together? Tomorrow we descend into logic itself: deduction (truth-preserving but never new), induction (Hume's wounded bird), and abduction (the detective's leap 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

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.

Revisability

EMPIRICAL?

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.

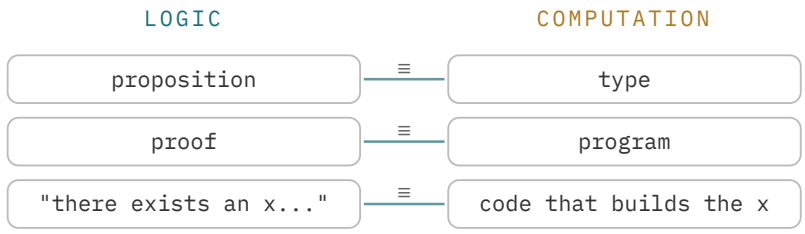
Paraconsistent logic drops *explosion*. In classical logic a single contradiction is apocalyptic: from "P and not-P" you can derive *literally anything* (the principle *ex contradictione quodlibet*) — one inconsistency and the whole system goes up in flames. Paraconsistent logics refuse this, letting 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