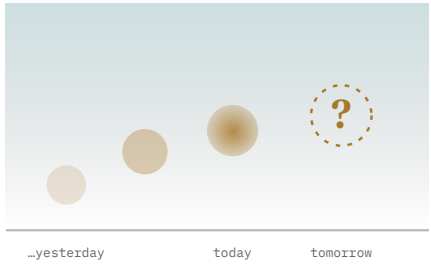
The Scientific Method & Demarcation

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

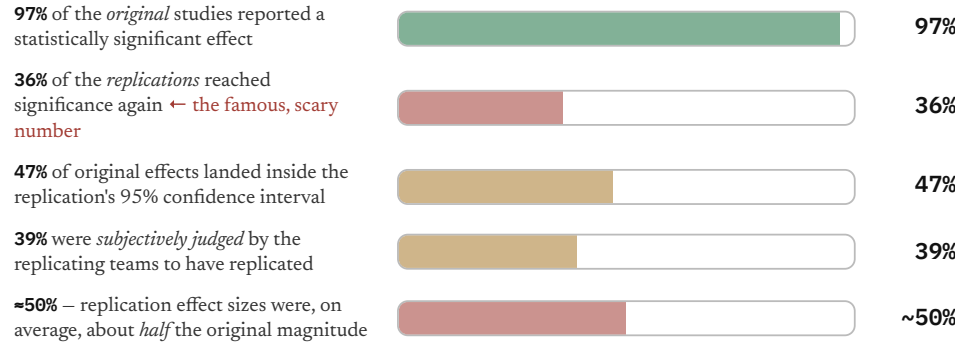
The replication crisis: demarcation under live fire

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

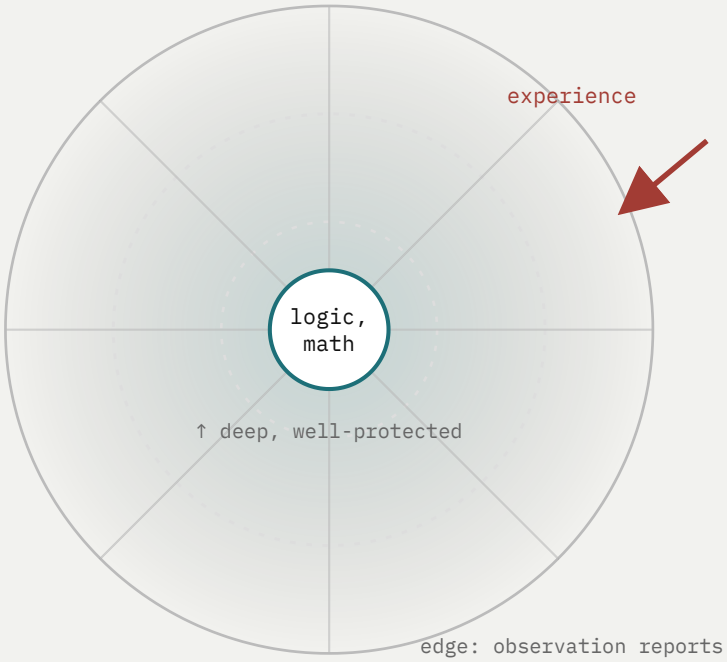
If single observations are theory-laden, the philosopher **W.V.O. Quine** showed in 1951 that single *tests* are theory-laden too — and turned it into a deeply influential paper in modern philosophy, "Two Dogmas of Empiricism." We met its child in the main lesson (the Duhem–Quine thesis: no hypothesis 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