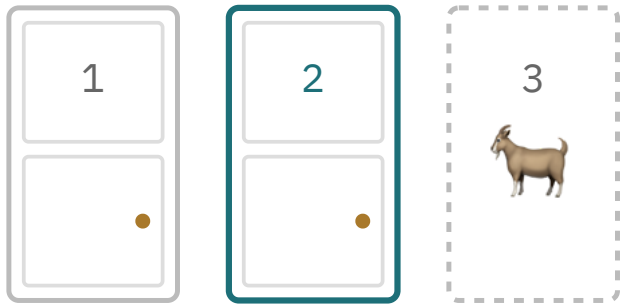
Probability as *Extended Logic*

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



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

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

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

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

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

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

THE MELTDOWN

The smartest people in the country, all wrong at once

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

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

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

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

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

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

THE MAN WHO NEEDED TO SEE IT TO BELIEVE IT

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

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

The Monty Hall Machine

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

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

WHY IT WORKS

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

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

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

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

OLDER THAN THE GAME SHOW

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

THE MODEL

Bayes' theorem: the law of belief revision

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

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

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

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

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

The trap that fools doctors

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

The base-rate trap

GROUP	COUNT OUT OF 1,000	POSITIVE TESTS
Sick people	1	About 1 true positive
Healthy people	999	About 50 false positives
All positive tests	About 51	Only about 1 is truly sick

The posterior is therefore about 0.99 / 50.94, or 1.9% — the Casscells result with the sensitivity made explicit.

THE DEEP IDEA

Why "extended logic," not just a formula

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

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

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

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

THE HONEST FOOTNOTE

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

THE DEBATE

Two tribes, one equation

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

Frequentist

PROBABILITY = LONG-RUN FREQUENCY

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

Bayesian

PROBABILITY = DEGREE OF BELIEF

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

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

"PROBABILITY DOES NOT EXIST"

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

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

THE FRONTIER · 2026

The quiet mutiny against the p-value

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

Edge 01 [ESTABLISHED]

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

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

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

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

Edge 02 [ESTABLISHED] [CONTESTED]

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

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

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

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

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

The e-value ledger

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

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

Edge 03 [ESTABLISHED] [REVIEW]

How far has the mutiny actually spread?

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

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

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

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

OPEN QUESTIONS

What's genuinely unsettled

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

◆ THE DAY IN THREE SENTENCES

BIG IDEA

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

BEST ANALOGY

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

LIVE CONTROVERSY

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

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

TOMORROW → DAY 05

Causation

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

SOURCES

Sources & further reading

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

END OF DAY 04 · 176 DESCENTS REMAIN