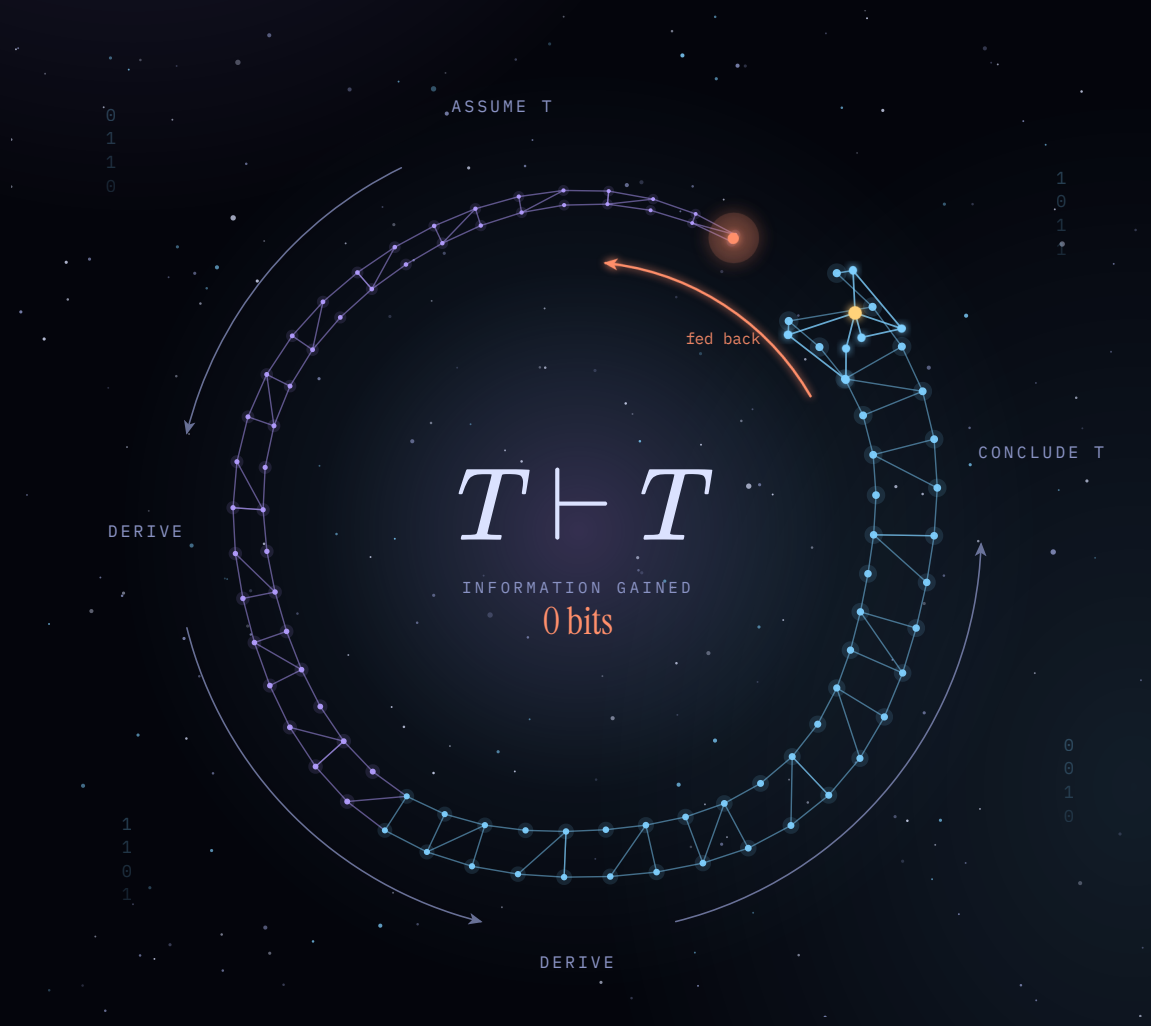
Christopher Yamba

The Ouroboros Problem

Why circular reasoning quietly erases the predictive value of AI agents that do science, how the same loop turns a mathematical proof into a null result, and how to build agents that cannot cheat.



Christopher Yamba October 2026 20-minute read christopheryamba.com

PROLOGUE

An oracle that was too good

Picture an AI agent asked to forecast the outcomes of 200 laboratory experiments. It plans, searches the literature, queries databases, writes analysis code and returns a prediction for each one. It scores 96%. The team is delighted, until someone reads the tool logs.

For 180 of the 200 experiments, a search call had returned a paper, a preprint or a database record that already reported the result. On those, the agent had found the answer and repeated it. On the remaining 20, the ones nobody had published yet, it did about as well as a coin.

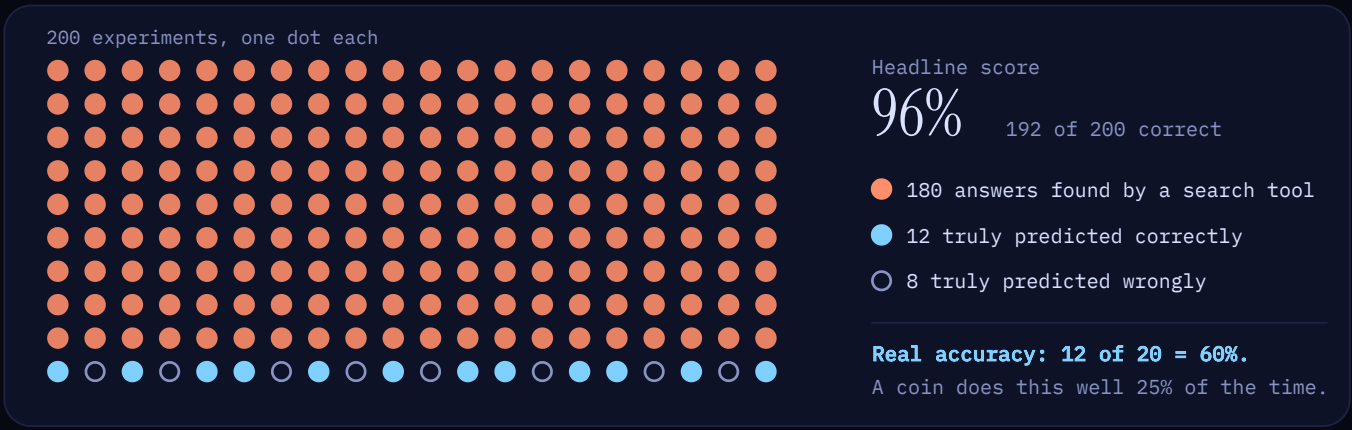


FIGURE 1 A hypothetical but very ordinary failure. Ember dots are answers the agent looked up. Only the bottom row was a real prediction, and 12 correct out of 20 is what chance produces a quarter of the time.

Nobody wrote code that said “cheat”. The loop closed by itself, through ordinary tools doing ordinary things. This essay is about that loop: the snake that eats its own tail. We will look at it in logic, where it turns a proof into a null result; in statistics, where it manufactures discoveries out of noise; and in agentic AI, where it hides in at least seven places. Then we will look at how to break it.

PART I · THE SHAPE

One shape, three disguises

A circular argument, the logicians’ *petitio principii* or “begging the question”, assumes what it sets out to show. The conclusion sits among the premises, usually in disguise. The disguise changes from field to field. The shape does not.

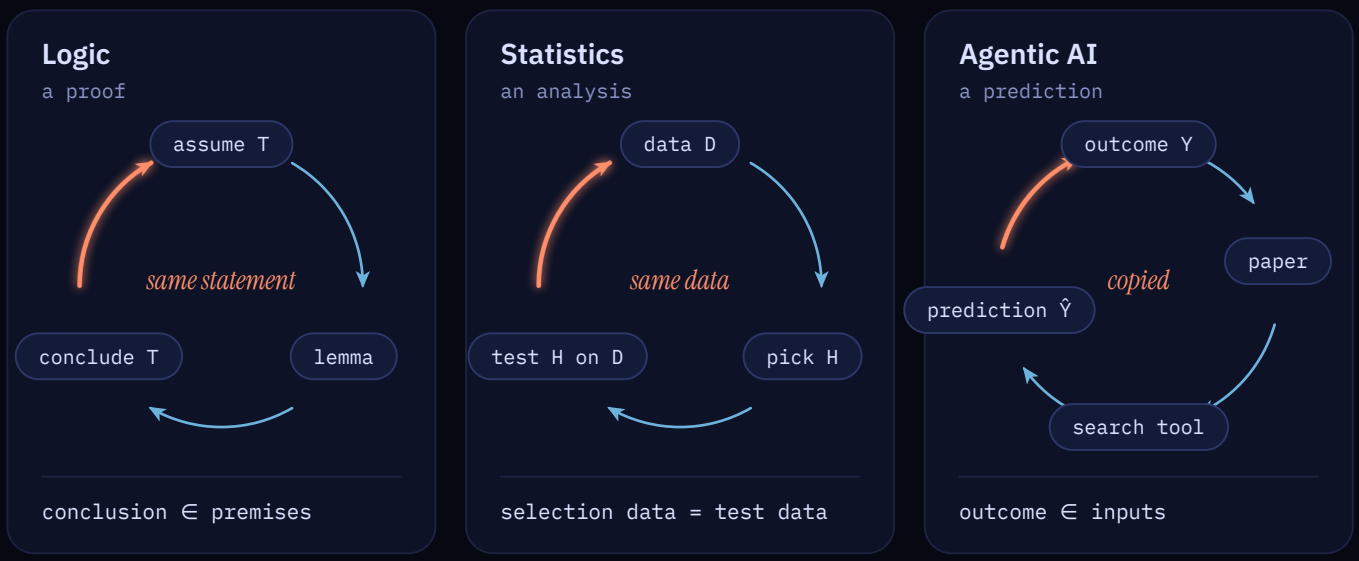


FIGURE 2 The same loop in three settings. Blue arrows are legitimate steps of reasoning. The ember arrow closes the circle: the end of the argument is fed back in as its beginning.

The cleanest way to see the shape is as a graph. Draw every claim as a node, and draw an arrow from A to B whenever B depends on A. Sound reasoning forms a **directed acyclic graph**: it starts from things accepted independently, such as axioms or measurements taken before a prediction, and flows toward the conclusion without looping back. Circular reasoning is a cycle.

DEFINITION

Let $\mathrm{Anc}(v)$ be the set of nodes that have a directed path to v. An argument for a conclusion C is circular when the conclusion is one of its own ancestors:

$$C \in \mathrm{Anc}(C).$$

Everything that follows is this one picture, seen from different angles.

PART II · THE CORE EQUATION

Evidence must be able to fail

Why is a cycle so damaging? Bayes' rule in odds form answers it in one line. Let H be a hypothesis and E an observation, such as "the check passed":

$$\underbrace{\frac{P(H \mid E)}{P(\neg H \mid E)}}_{\text{posterior odds}} = \underbrace{\frac{P(E \mid H)}{P(E \mid \neg H)}}_{\text{Bayes factor } K} \times \underbrace{\frac{P(H)}{P(\neg H)}}_{\text{prior odds}}$$

The Bayes factor K is the entire value of the evidence. Its logarithm $W = \log_2 K$ is what the statistician I. J. Good called the *weight of evidence*, here measured in bits. The quantity that matters most is in the denominator: $P(E \mid \neg H)$, the chance that the check passes even when the hypothesis is false.

A circular check passes whatever the truth is, because the answer was built into it. Then

$$P(E \mid H) = P(E \mid \neg H) = 1 \quad \Longrightarrow \quad K = 1 \quad \Longrightarrow \quad W = \log_2 1 = 0 \text{ bits.}$$

The posterior equals the prior. Nothing was learned, however long or sophisticated the argument looked.

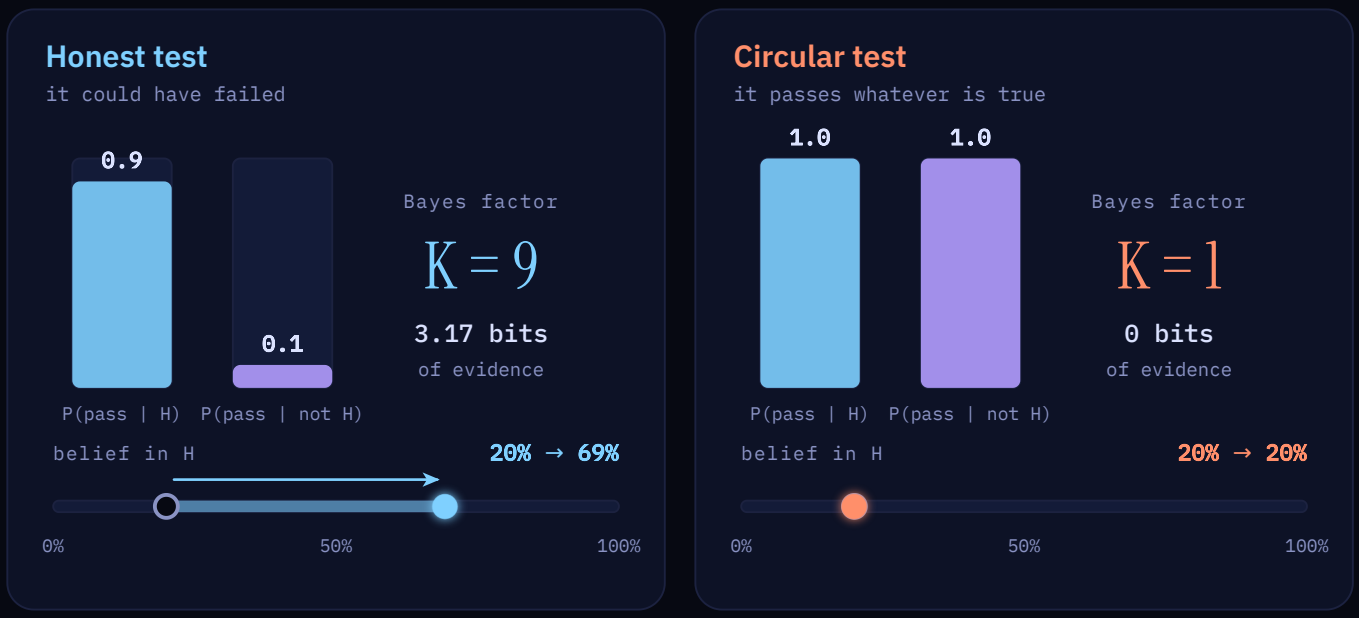


FIGURE 3 Same prior belief (20%), two checks. The honest check could have failed: $K = 9$, 3.17 bits, and belief rises to 69%. The circular check cannot fail: $K = 1$, zero bits, and belief stays exactly where it was.

KEY IDEA

A test is worth exactly as much as its ability to fail. Circular reasoning pushes the false-pass rate to one and the evidence to zero.

Karl Popper made the same point in words: a claim earns credibility only by surviving a test that could have refuted it. The Bayes factor puts a number on that idea. Keep it in mind, because every failure mode in the rest of this essay is a way of quietly pushing $P(E \mid \neg H)$ toward 1.

PART III · MATHEMATICS

A circular proof proves nothing

Mathematics looks like the one place where circularity should be easy to catch. Proofs are explicit and every step is written down. Yet circular proofs are among the most persistent errors in the history of the subject, because the assumption that closes the loop rarely looks like the theorem. It looks like something obvious.

What a proof actually rules out

Think of every possible world (in logic, every *model*) in which the axioms A hold. A statement T is true in some of them and false in others, until a proof shows otherwise. A valid proof shows that the second group is empty:

$$\mathrm{Mod}(A) \subseteq \mathrm{Mod}(T).$$

A circular proof quietly adds T to the premises. What it actually establishes is

$$\mathrm{Mod}\left(A \cup \{T\}\right) \subseteq \mathrm{Mod}(T),$$

which holds for *every* statement T, true or false, provable or not. It rules out no worlds. Its information content is zero, and that is the precise sense in which a circular proof is a null result.

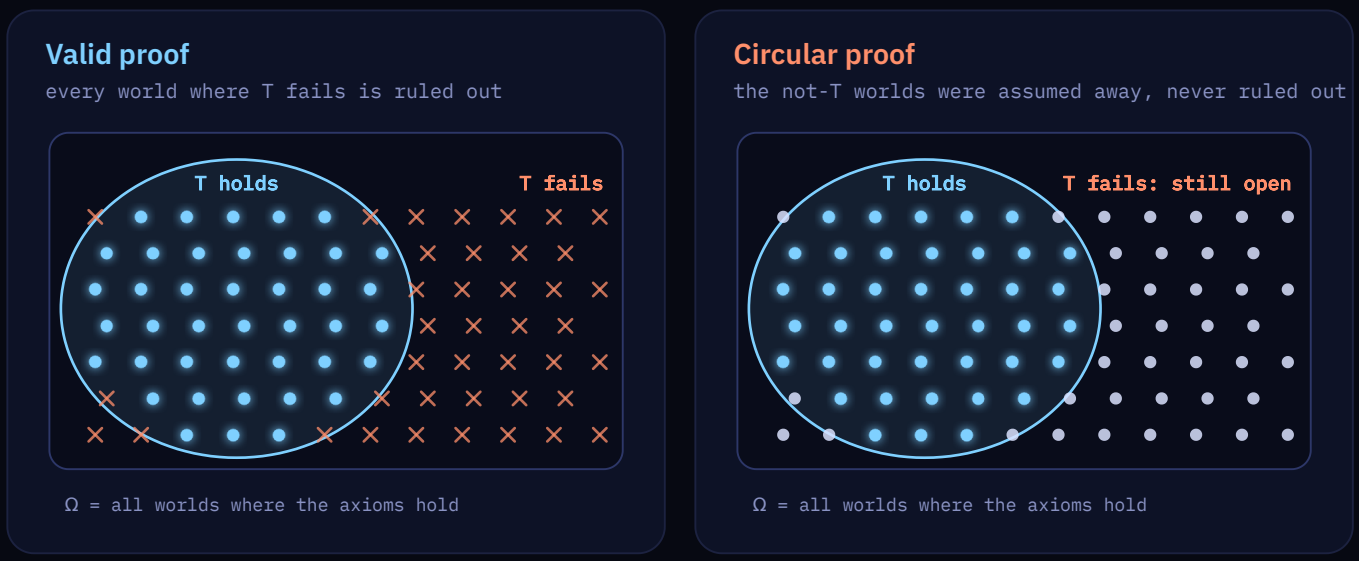


FIGURE 4 Each dot is a world consistent with the axioms. A valid proof eliminates every world in which T fails. A circular proof only ever looks at worlds where T already holds, so the others are never examined, let alone excluded.

Example: a famous limit

Every calculus student meets $\lim_{x\to 0} \frac{\sin x}{x} = 1$. A tempting one-line proof uses L'Hôpital's rule:

$$\lim_{x\to 0} \frac{\sin x}{x} \overset{?}{=} \lim_{x\to 0} \frac{\cos x}{1} = 1.$$

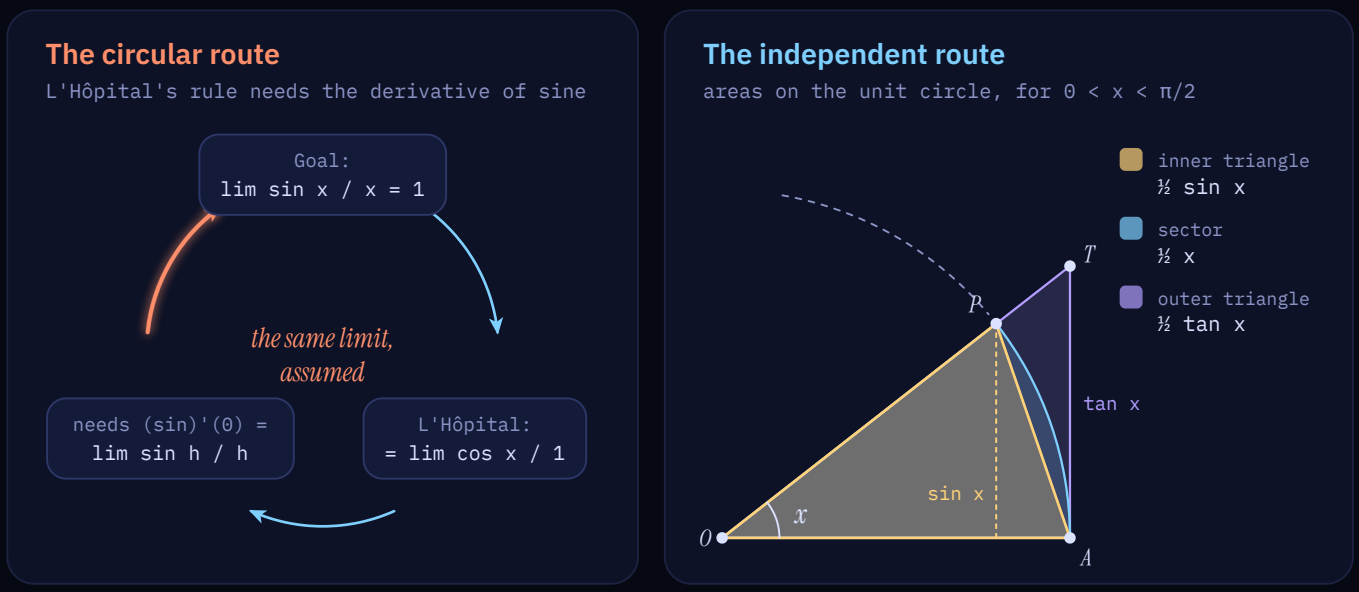
The rule needs the derivative of the sine. With sine defined geometrically, as in most courses, its derivative at zero is computed as

$$\sin'(0) = \lim_{h\to 0} \frac{\sin h - \sin 0}{h} = \lim_{h\to 0} \frac{\sin h}{h},$$

which is the limit we set out to find. The proof assumes its own conclusion. The repair is to anchor the limit in something independent. Three nested areas on the unit circle give, for $0 < x < \pi/2$,

$$\tfrac{1}{2}\sin x \le \tfrac{1}{2}x \le \tfrac{1}{2}\tan x \quad\Longrightarrow\quad \cos x \le \frac{\sin x}{x} \le 1,$$

and squeezing as $x \to 0$ gives the limit 1 without touching a derivative.



½sin x ≤ ½x ≤ ½tan x ⇒ cos x ≤ sin x / x ≤ 1 ⇒ squeeze as x → 0: the limit is 1, no derivative needed

FIGURE 5 Left: the circular route, in which the derivative that L'Hôpital needs is the very limit in question. Right: an independent route through three nested areas on the unit circle.

A subtlety worth keeping. If the sine is *defined* by its power series, its derivative comes for free and L'Hôpital's rule is perfectly legitimate. Circularity is a property of the dependency graph, not of the formula on the page. The same sentence can be honest in one system and circular in another.

Example: two thousand years of circular proofs

Euclid's fifth postulate, the parallel postulate, always looked less self-evident than his other four. For roughly two thousand years mathematicians tried to derive it from them: Ptolemy, Proclus, John Wallis, Girolamo Saccheri, Adrien-Marie Legendre and many more. Some attempts contained outright errors. Many, on close inspection, had assumed a statement that is *equivalent* to the postulate: that the angles of every triangle sum to 180°, that similar triangles of different sizes exist, that a circle passes through any three points not on a line.

The loop was finally broken in the nineteenth century, by a counter-model. Lobachevsky (1829) and János Bolyai (1832) developed a consistent geometry in which the postulate fails, and Eugenio Beltrami (1868) showed how to model it inside ordinary Euclidean geometry. That settles the matter through the soundness of logic:

$$\mathrm{Mod}\left(A \cup \{\neg T\}\right) \neq \varnothing \quad \Longrightarrow \quad A \nvdash T.$$

If there is a world in which the axioms hold and T fails, no valid proof of T from those axioms can exist. Every "proof" ever written was therefore either mistaken or smuggled T in through the back door. Centuries of effort carried zero bits about the postulate itself.

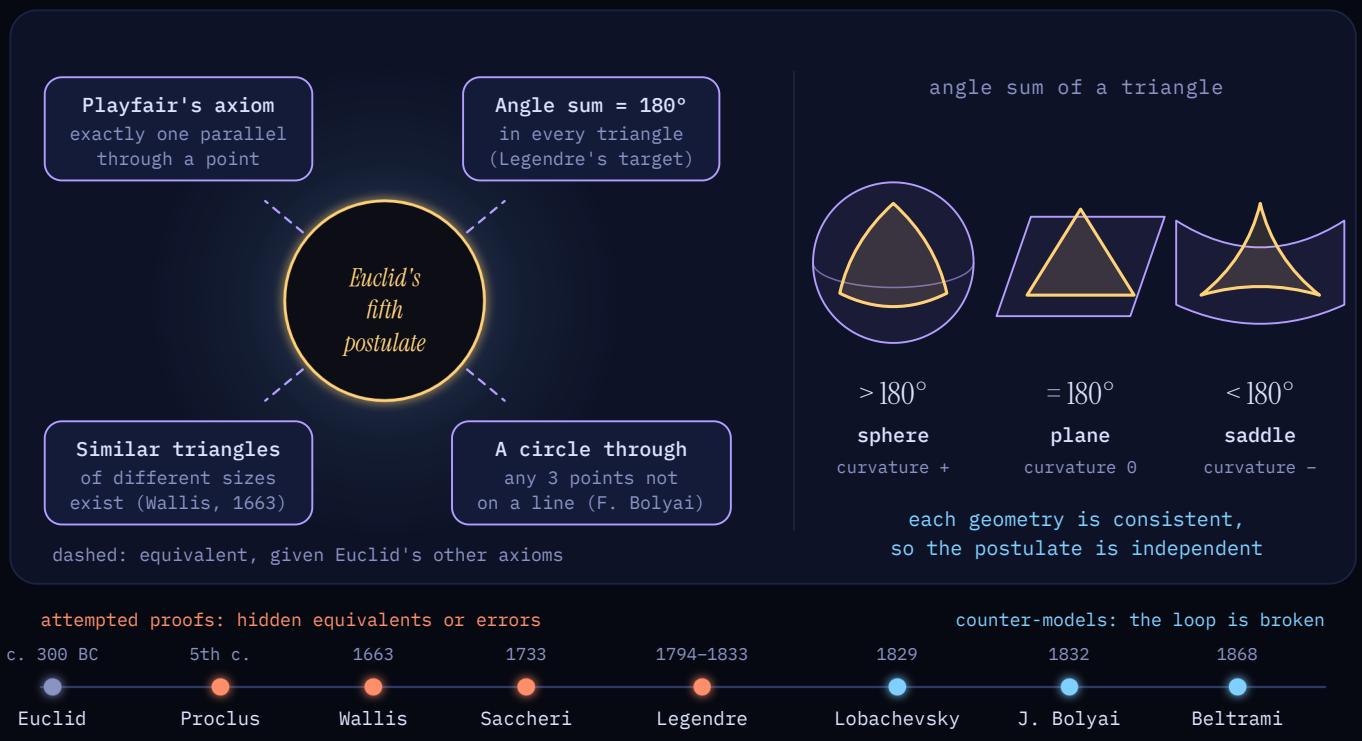


FIGURE 6 Left: statements that, given Euclid's other axioms, are each equivalent to the fifth postulate. Assuming any one of them assumes the postulate. Right: on a sphere, a plane and a saddle the angle sum of a triangle differs, and each geometry is internally consistent.

Example: reasoning backwards

A third pattern is subtler. Start from the claim, derive something true, and declare victory.

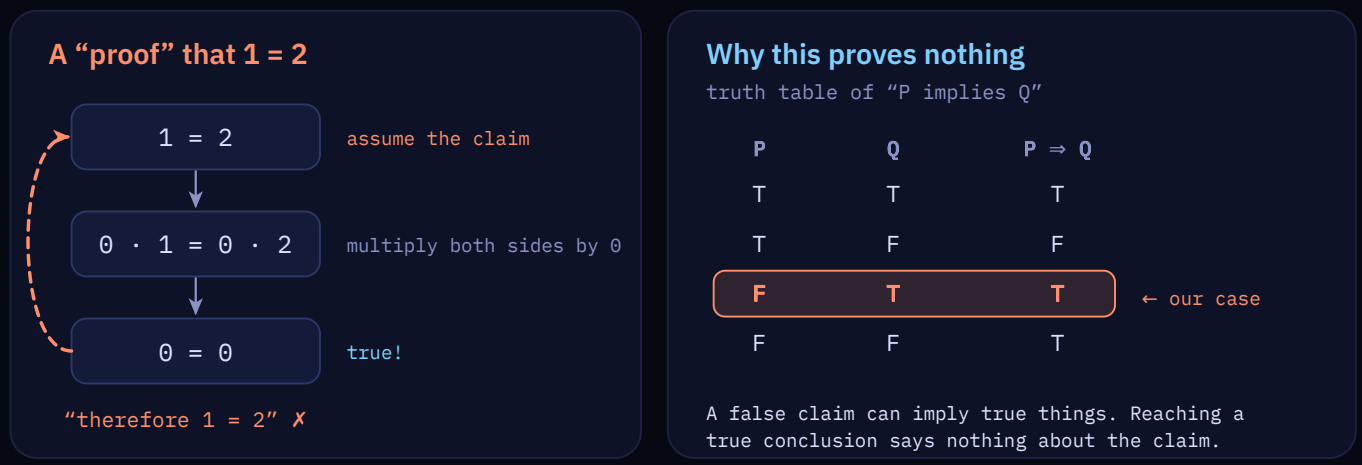


FIGURE 7 A false premise can imply a true conclusion. Reaching something true from a claim says nothing about the claim, because that row of the truth table is true either way.

In Bayesian terms, "the consequence checked out" has $P(E \mid \neg H)$ close to 1 whenever the consequence was going to be true anyway. This is exactly what an agent does when it "validates" a hypothesis by showing that it agrees with the literature it just retrieved.

KEY IDEA

Consistency is not confirmation. A hypothesis that agrees with what we already knew has passed a test it could not have failed.

PART IV · PREDICTION

In prediction, circularity is called leakage

Empirical science adds time to the picture. A prediction is a function of the information available when it is made, and the outcome arrives later. Write X for everything legitimately available at prediction time and Y for the outcome. An honest predictor is $\hat{Y} = f(X)$, and the honesty condition fits in one line:

THE HONESTY CONDITION

$$I(\hat{Y}; Y \mid X) = 0$$

Once you know X, the prediction carries no further information about the outcome. If this conditional mutual information is positive, something about Y reached the prediction through a side door.

Machine learning calls this **leakage** (Kaufman et al., 2012), and it is a widespread cause of irreproducible results in ML-based science (Kapoor & Narayanan, 2023). It is circular reasoning in a lab coat.

Information can only be lost

The data processing inequality makes the constraint precise. When the prediction depends on the world only through what was measured, $Y \to X \to \hat{Y}$ is a Markov chain and

$$I(Y; \hat{Y}) \leq I(Y; X).$$

No model, however clever, can know more about the outcome than its inputs contain.

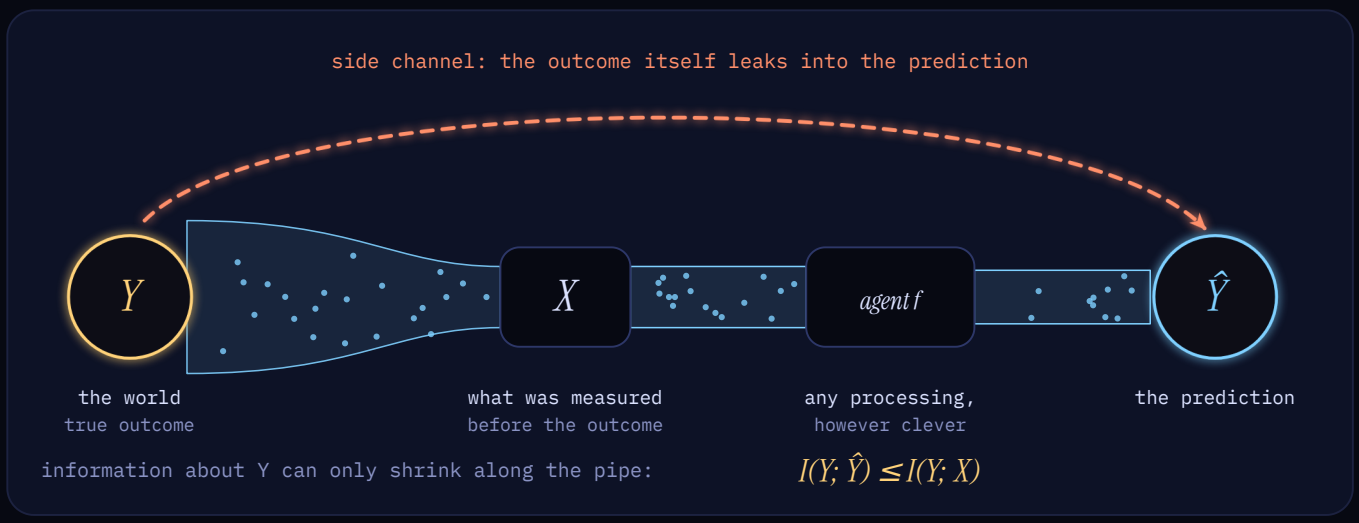
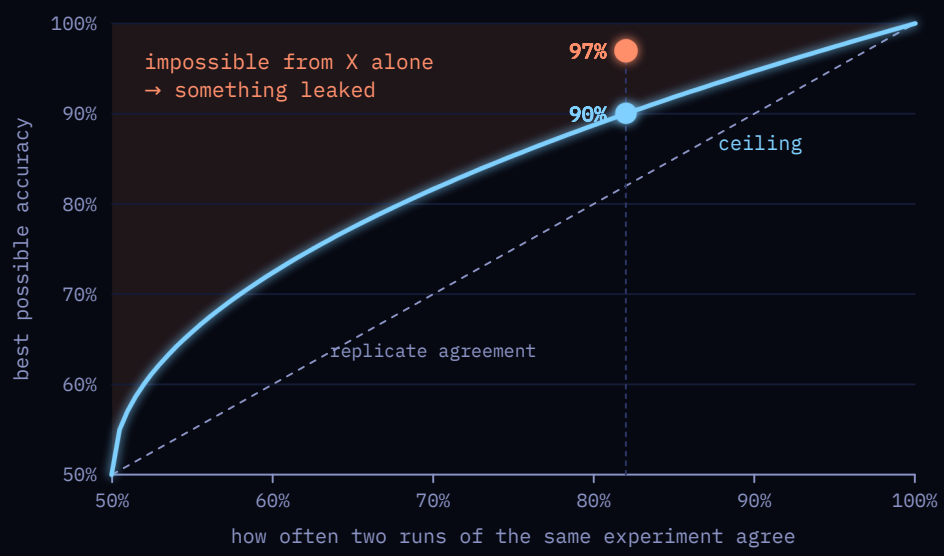


FIGURE 8 Information about the outcome can only narrow on its way through measurement and modelling. Performance that exceeds what the pipe can carry must have come through another channel.

This gives a practical alarm. Experiments are noisy, and two runs of the same assay never agree perfectly. Suppose each run reports a true yes-or-no outcome correctly with probability q, with independent errors. Two runs then agree with probability

$$a = q^2 + (1-q)^2 \quad \Longrightarrow \quad q = \frac{1+\sqrt{2a-1}}{2}.$$

Even a model that knows the true answer perfectly can match a single run only with probability q, because that run's noise is not in its inputs. So q is a hard ceiling, and you can estimate it from replicate agreement alone.



Example

Two runs of an assay agree 82% of the time.

Even a model that knows the true answer can match a single run at most 90% of the time.

A model that scores 97% against that run is predicting the run's noise. The noise was never in its inputs.

FIGURE 9 The replicate ceiling, assuming symmetric and independent measurement errors. If two runs agree 82% of the time, no honest model can exceed 90% against one run. A reported 97% means the model is reading the measurement, not predicting the phenomenon.

Double dipping: selecting and testing with the same data

The statistical form of the loop is to use one dataset twice: first to choose what to look at, then to test what you chose. Kriegeskorte and colleagues (2009) called it *double dipping* in neuroscience, and Vul and colleagues (2009) traced many implausibly high brain–behaviour correlations to it.

A simulation makes it concrete. Take 30 samples, an outcome made of pure noise and 10,000 candidate features, also pure noise. Pick the feature that correlates best with the outcome and report that correlation.

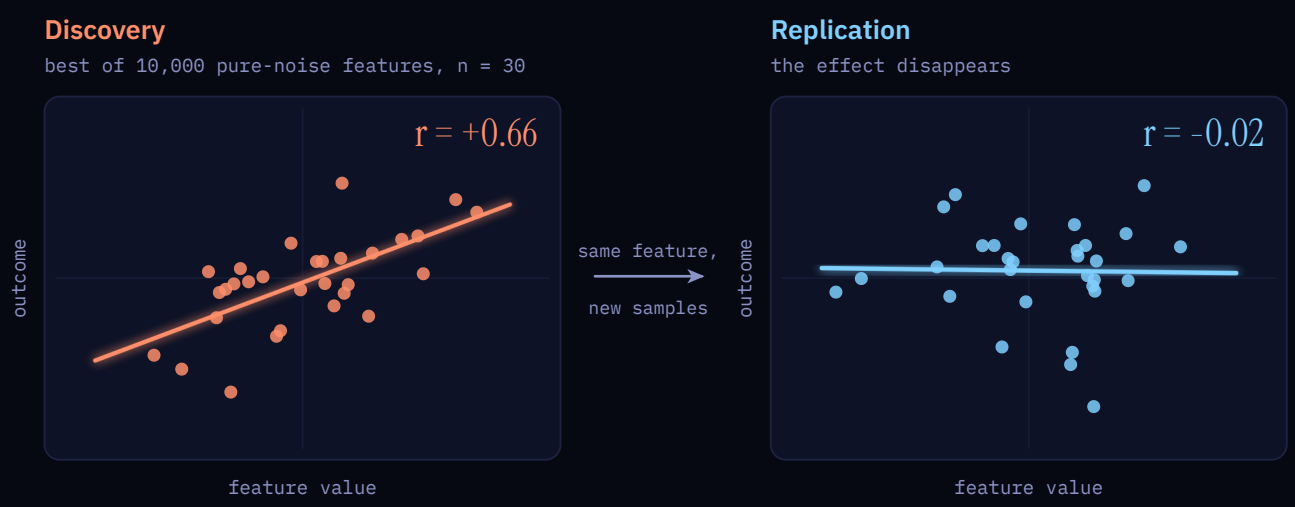


FIGURE 10 Discovery: a correlation of 0.66 from noise alone. Replication on new samples: −0.02. Nothing was faked; the selection step and the test step simply shared the same data.

The size of the illusion grows with the size of the search. With n samples and k hypotheses, the largest correlation found in pure noise grows roughly like

$$\max_{j \le k} |r_j| \approx \sqrt{\frac{2 \ln k}{n}},$$

which runs a little high when n is small. The simulated curve is below.

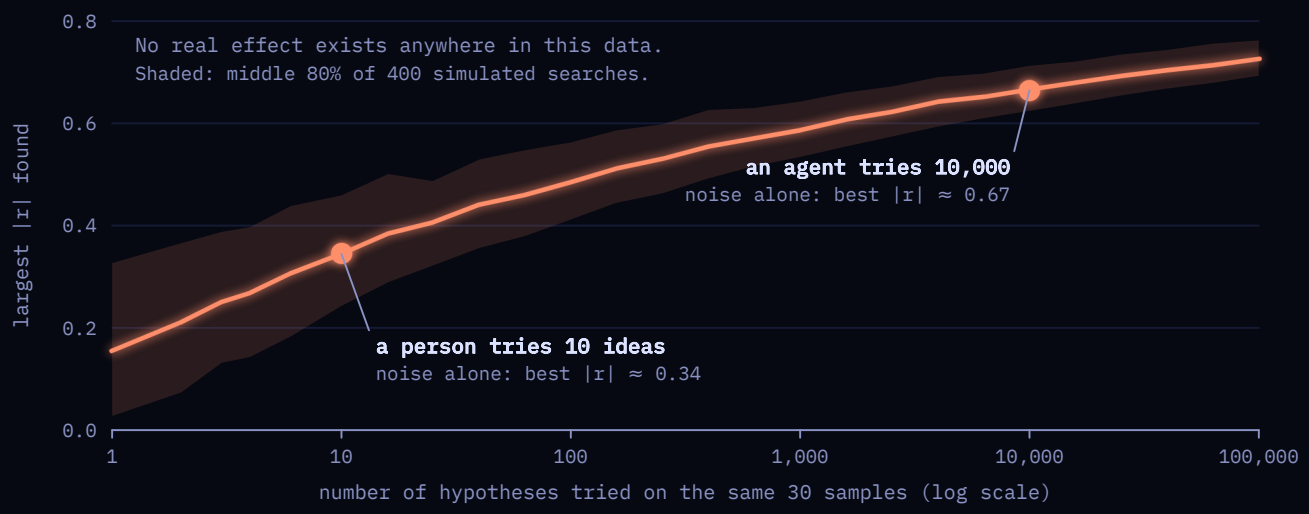


FIGURE 11 How much "signal" pure noise produces as a function of how hard you search. Each point is 400 simulated searches over 30 samples.

> **KEY IDEA**
>
> Agents search at a scale no person can. An agent that tests ten thousand hypotheses an hour is a double-dipping machine, unless the data that guides its search never touches the data that judges its result.

Counting the same evidence twice

A quieter version of the loop lets the same evidence into the calculation twice. If data D is used once to shape the prior and again as the likelihood, the posterior becomes

$$p(\theta \mid D, D) \;\propto\; p(D \mid \theta)^2\, p(\theta).$$

For Gaussian data this halves the variance: the standard error shrinks from $\sigma/\sqrt{n}$ to $\sigma/\sqrt{2n}$. Every interval looks $\sqrt{2}$ times tighter than it is, and a nominal 95% interval actually covers the truth with probability

$$P\left(|Z| < \tfrac{1.96}{\sqrt{2}}\right) \;\approx\; 0.83.$$

Agents do this naturally. A single experiment is described in a paper, a review, two press releases and a database entry; retrieval returns all five, and the agent counts five confirmations. Or an agent writes a summary of a source into its memory, later retrieves both the source and its own summary, and treats them as independent witnesses.

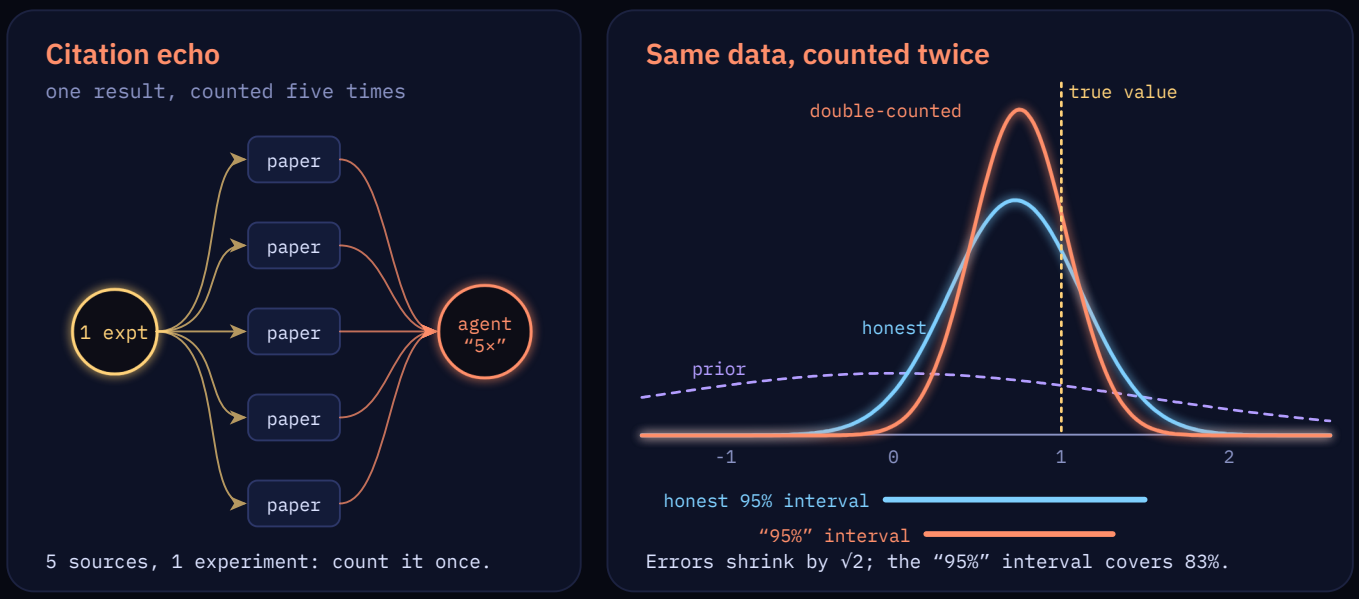


FIGURE 12 Left: citation echo, one result counted five times. Right: counting the same data twice produces a posterior that is narrower and more confident than the evidence supports. Simulated coverage of the "95%" interval: 83.5%.

Wearing out the test set

The most agentic form of the loop is iteration. An agent proposes a change, checks the score on the test set, keeps what helps and repeats. Every look leaks a little information about the test labels into the system. With m looks that each return one of s possible answers, the system can learn up to

$$m \log_2 s \ \text{ bits}$$

about the test set. A keep-or-discard decision is one bit per look, so a few hundred iterations can encode a few hundred bits, which is about as many labels as a 200-example yes-or-no test set has. Dwork and colleagues (2015) showed how to reuse a holdout safely by strictly limiting what each look reveals.

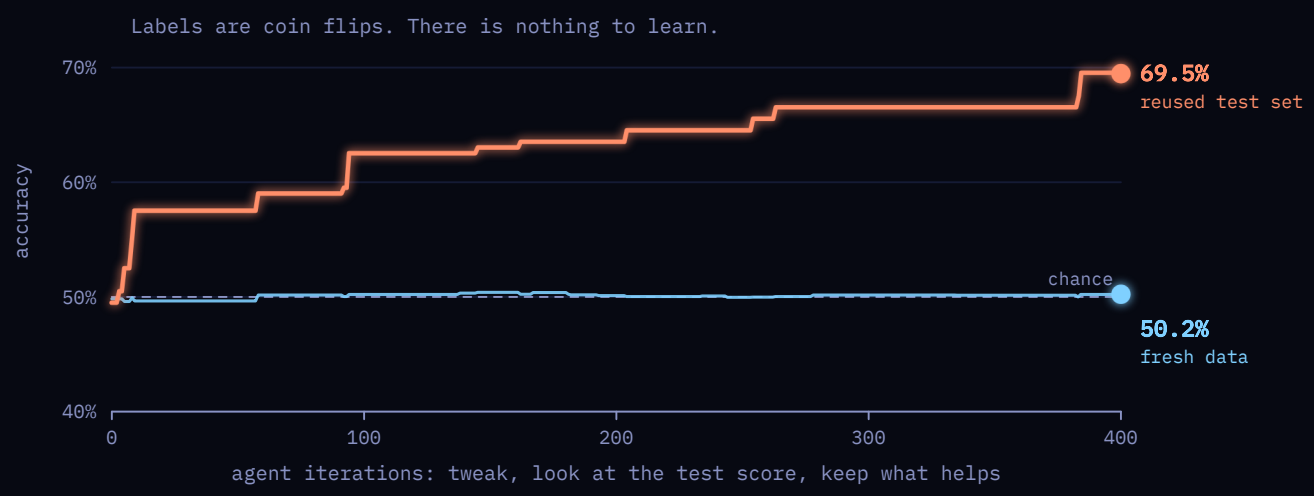


FIGURE 13 The labels are coin flips, so there is nothing to learn. A hill-climbing agent that keeps every change that does not lower the test score reaches 69.5% on the test set it keeps reusing. On fresh data it stays at 50.2%.

Where the loop hides in an agentic system

An agentic system multiplies the number of paths information can take. A language model brings everything it read in training. Tools reach the live literature. Memory and databases accumulate context. Evaluators and judges feed scores back into the next round. Each of these paths is useful, and each is also a way for the outcome to reach the prediction.

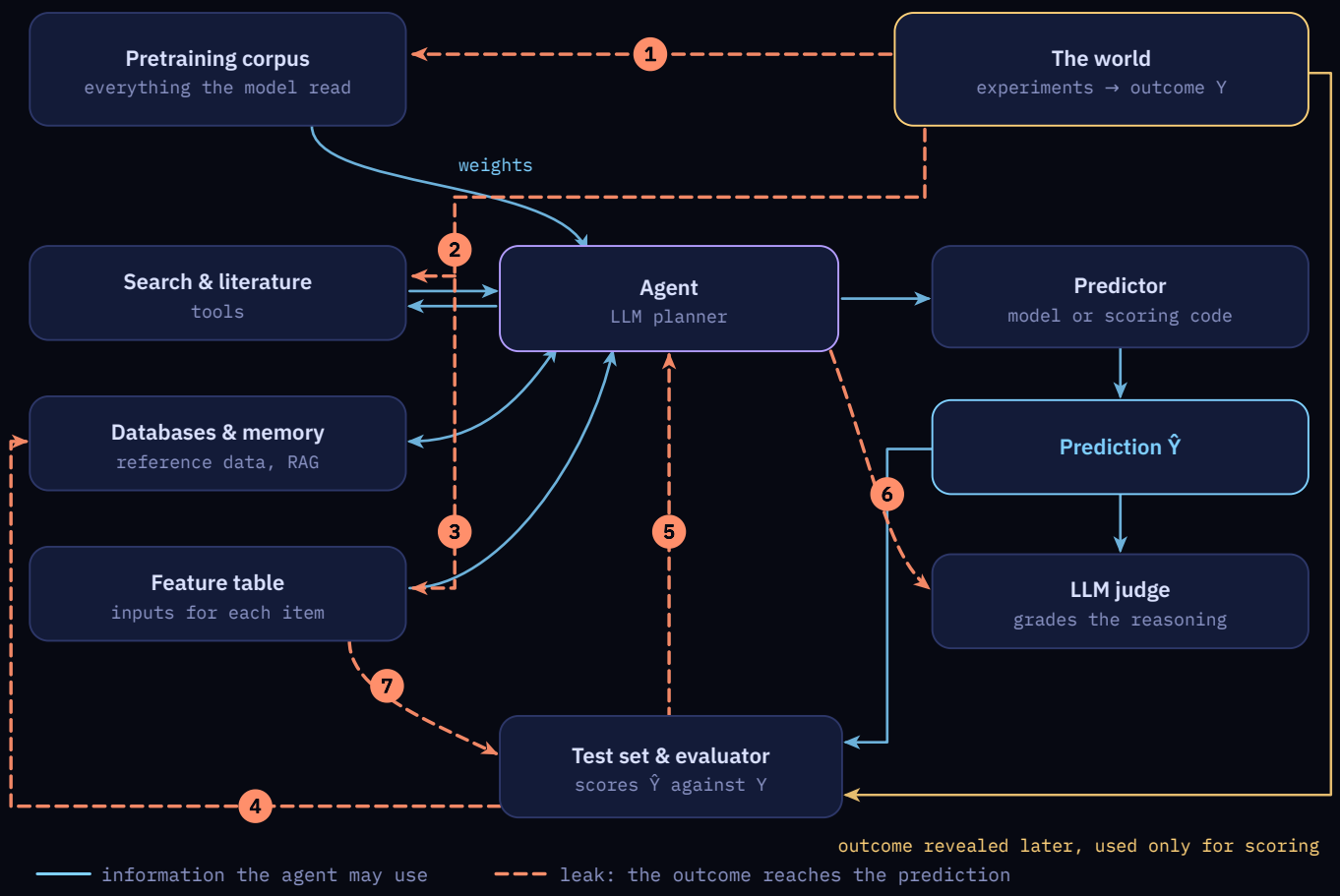


FIGURE 14 Anatomy of an agentic prediction system. Blue: information the agent may legitimately use. Ember: seven channels through which the outcome can reach the prediction. Gold: the outcome itself, which should reach only the evaluator, and only after the prediction is fixed.

1 Memorised answers

HOW THE LOOP CLOSES

The base model was trained on text published after the experiments it is now asked to predict. The result is in the weights.

BREAK IT

Evaluate on outcomes produced after the model's training cutoff, and probe for memorisation, for example by asking it to complete text it should never have seen (Golchin & Surdeanu, 2024).

2 Retrieved answers

HOW THE LOOP CLOSES

Search, retrieval or a database lookup returns the paper, preprint, patent, registry entry or later review that reports the outcome.

BREAK IT

Give every tool an as-of date. Log each retrieved document with its date, and fail the run if any of them postdates the prediction.

3 Proxy features

HOW THE LOOP CLOSES

Inputs that exist only because of the outcome: a “selected for follow-up” flag, an annotation added after the result, the number of papers written about an item.

BREAK IT

Ask one question of every input: could this value have been recorded before the outcome existed?

4 Near-duplicate splits

HOW THE LOOP CLOSES

A random split puts near-identical items on both sides, so the model recognises neighbours instead of learning a rule (Wallach & Heifets, 2018).

BREAK IT

Split by group, cluster or time, and report how similar each test item is to its nearest training item.

5 Test-set steering

HOW THE LOOP CLOSES

The agent iterates against the evaluation score until the test set has quietly become training data.

BREAK IT

Lock the test set and tune on a separate development set. Count every look, and use a reusable-holdout scheme if you must look often.

6 Self-grading

HOW THE LOOP CLOSES

The same model, or a close relative, both proposes and judges. Shared blind spots look like agreement, and LLM judges tend to favour their own text (Panickssery et al., 2024).

BREAK IT

Use judges with different information and different models, and tie every grade to a check that could fail.

7 Definitional loops

HOW THE LOOP CLOSES

The outcome label is defined using the same signal the model predicts from, for example “responders” defined by a marker that is also an input.

BREAK IT

Define outcomes before modelling, independently of the inputs, and write the definition down.

✓ The common thread

WHY IT KEEPS HAPPENING

None of these channels needs bad intent. They are the default behaviour of capable tools in a world where answers get published.

SO

The defence has to be designed into the system. It cannot be left to good intentions.

One useful reporting habit follows. The same accuracy means very different things depending on how far the test items are from what the system has already seen, so say where on this line your evaluation sits.

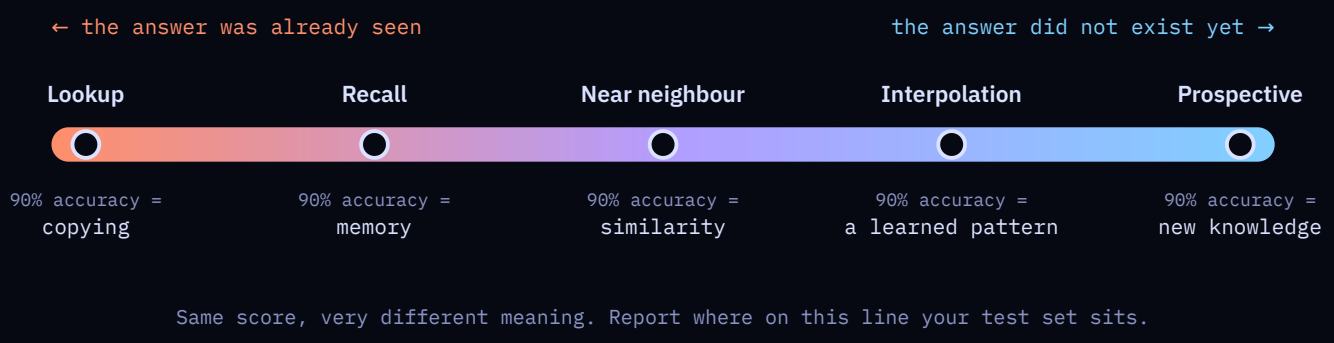


FIGURE 15 The leakage spectrum. “92% on outcomes published before the model’s training cutoff” and “92% on experiments run after the predictions were committed” are different claims, and only the second is a forecast.

PART VI · DEFENCES

Breaking the circle

The defences follow directly from the picture we started with. The goal is a reasoning graph with no path from the outcome to the prediction, and checks that could have failed.

01 Draw the information graph, and check it in code

Treat provenance as data. Every input, retrieved document and tool response gets a timestamp and a source. The test then becomes mechanical:

THE RULE

$$Y \notin \mathrm{Anc}(\hat{Y}), \qquad \text{and every node in } \mathrm{Anc}(\hat{Y}) \text{ is dated before } \hat{Y} \text{ is fixed.}$$

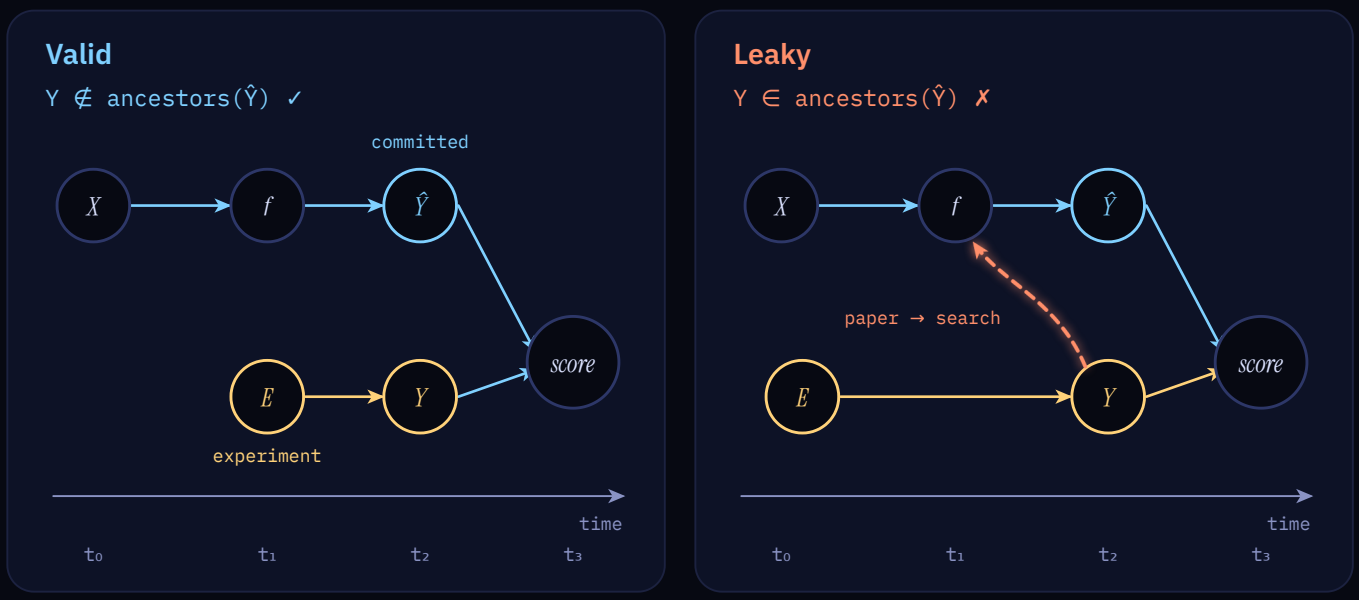


FIGURE 16 Left: the prediction is committed before the experiment runs, so the outcome cannot be among its ancestors. Right: the outcome reaches the model through a paper and a search tool, and the arrow of time is broken.

02 Put a time firewall around the agent

Fix an as-of date for every prediction. The model's training cutoff, every tool and every database snapshot must respect it. When they cannot, the evaluation is a test of recall and should be reported as one.

03 Commit, then reveal

The strongest guarantee is to predict outcomes that do not exist yet. Publish a cryptographic commitment before the experiment runs,

$$c = \mathrm{SHA\text{-}256}(\hat{y} \,\|\, s),$$

where s is a random salt. The hash reveals nothing about $\hat{y}$, and once the outcome is known anyone can check that the revealed prediction matches. Structural biology has run this experiment at scale since 1994: in CASP, research groups predict protein structures that have been solved but not yet released (Moult et al., 1995). That design is a large part of why results in the field are trusted. Preregistration applies the same logic to whole studies (Nosek et al., 2018).

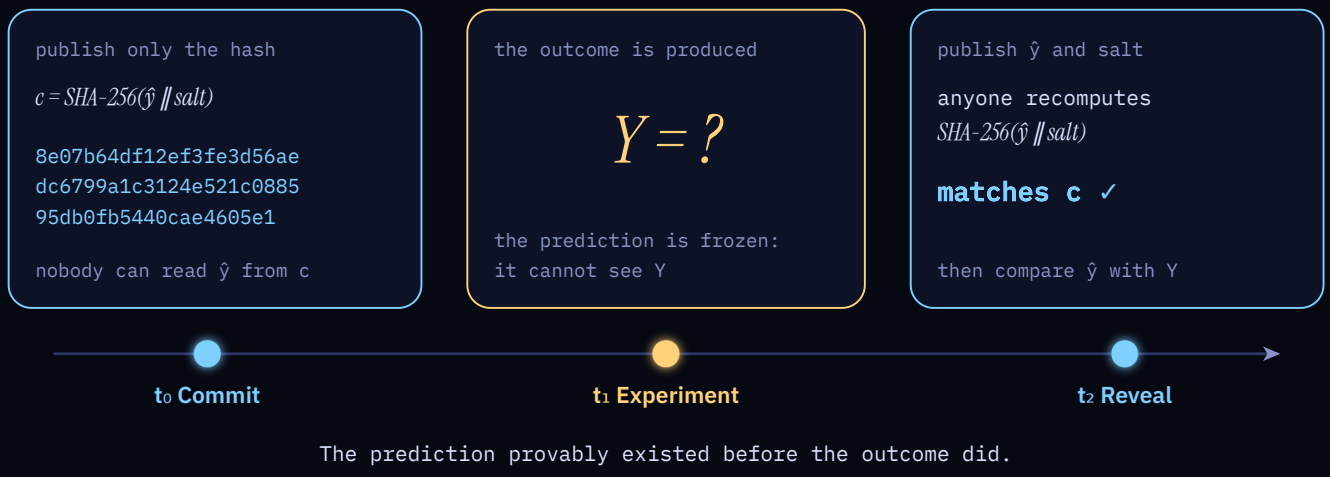


FIGURE 17 Commit and reveal. The hash shown is the real SHA-256 of an example prediction string, so the scheme can be checked by hand.

04 Run negative controls through the whole pipeline

Scramble the outcome labels and run everything again: every step, including feature selection, tuning and the agent's own choices. An honest pipeline must fall to chance. If it still scores well on noise, it is reading the answer from somewhere.

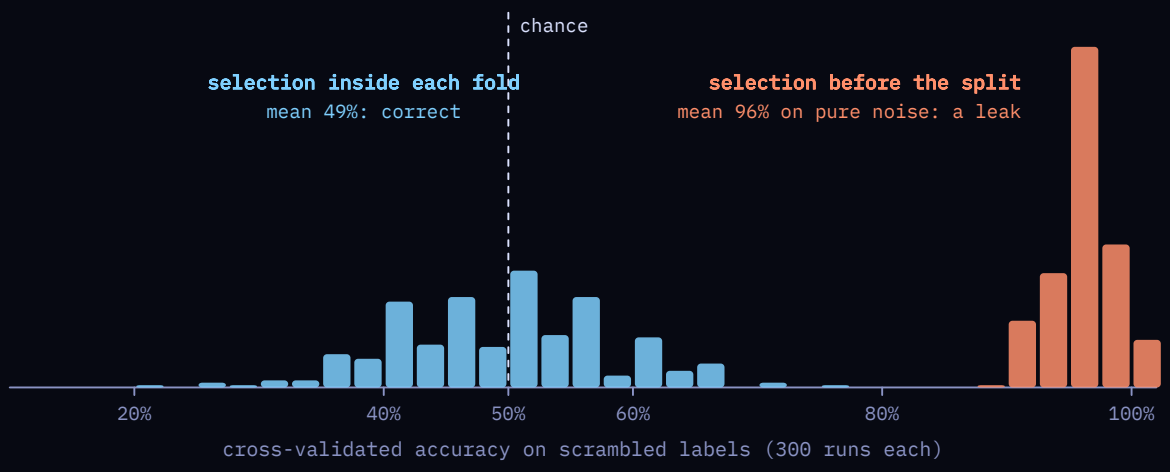


FIGURE 18 Sixty samples, 5,000 noise features, scrambled labels, 300 runs. Choosing the 20 "best" features before cross-validation gives 96% accuracy on pure noise, the trap Ambroise and McLachlan (2002) described in gene-expression studies. Choosing them inside each fold gives 49%, as it should.

```python
import numpy as np

def negative_control(pipeline, X, y, runs=200, seed=0):
    """Run the WHOLE pipeline on scrambled labels. An honest one falls to chance."""
    rng = np.random.default_rng(seed)
    scores = [pipeline(X, rng.permutation(y)) for _ in range(runs)]
    return np.mean(scores), np.percentile(scores, [5, 95])

chance = max(np.mean(y), 1 - np.mean(y))          # majority-class rate
real = pipeline(X, y)
null_mean, (lo, hi) = negative_control(pipeline, X, y)
assert null_mean < chance + 0.05, "Leak: the pipeline scores well on noise"
print(f"real {real:.2f} vs scrambled {null_mean:.2f} [{lo:.2f}, {hi:.2f}]")
```

For agents that can read the literature, scrambling labels is not enough, because the literature is not scrambled. Two complements help. **Mask identities**: replace names with neutral codes so that retrieval cannot find the item. And **split by time**: compare accuracy on outcomes published before and after the model's training cutoff. A model that predicts the past much better than the future is remembering.

05 Separate the generator from the judge

A check is informative only if it can disagree. Use evaluators that hold different information from the generator: a different model, a held-out data source, a physical experiment, or a human expert reviewing a random sample.

06 Budget your looks

Keep a development set for iteration and touch the final test set once. Log every look. If you must look often, use a mechanism that limits how much each look reveals.

07 Respect the ceiling

Estimate how reproducible the measurement is before celebrating. A model that beats the replicate ceiling is a leak until proven otherwise.

PART VII · CHECKLIST

The Ouroboros audit

Ten questions to ask of any system that claims to predict scientific outcomes, including your own.

- [ ] 01 Could the outcome have been known, anywhere, before the prediction was fixed?
- [ ] 02 Was the outcome, or anything written about it, published before the model's training cutoff?
- [ ] 03 Does every tool have an as-of date, and is every retrieved document logged with its date?
- [ ] 04 Could each input have been recorded before the outcome existed?
- [ ] 05 Are test items separated from training items by group or by time rather than at random?
- [ ] 06 How many times has anyone, or any agent, looked at the test score?
- [ ] 07 Is the judge independent of the generator, in model and in information?
- [ ] 08 Are repeated reports of a single result counted once?
- [ ] 09 Does the whole pipeline fall to chance on scrambled labels and masked identities?
- [ ] 10 Is the reported accuracy below the replicate ceiling of the measurement?

CODA

A prediction is a bet the world can still lose

Science moves forward on claims that could have turned out false and did not. That possibility of failure is what turns a prediction into evidence. Remove it, and the most impressive number in the world carries zero bits.

This matters most where experiments are slow and expensive, as in drug discovery, where a single validation can take months. There, an unnoticed loop is worse than having no model at all. It produces confidence with no information behind it, and that confidence decides where real time and money go.

AI agents will become very good at science. The task ahead is to make their successes mean something. Break the loop, and every correct prediction becomes what it claims to be: evidence that could have gone the other way.

References

1. Ambroise, C. & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS* 99(10), 6562–6566.
2. Cover, T. M. & Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley. (Data processing inequality.)

3. Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O. & Roth, A. (2015). The reusable holdout: preserving validity in adaptive data analysis. *Science* 349(6248), 636–638.
4. Golchin, S. & Surdeanu, M. (2024). Time travel in LLMs: tracing data contamination in large language models. *ICLR 2024*. arXiv:2308.08493.
5. Good, I. J. (1950). *Probability and the Weighing of Evidence*. Griffin.
6. Greenberg, M. J. (2008). *Euclidean and Non-Euclidean Geometries: Development and History*, 4th ed. W. H. Freeman.
7. Kapoor, S. & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* 4(9), 100804.
8. Kaufman, S., Rosset, S., Perlich, C. & Stitelman, O. (2012). Leakage in data mining: formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data* 6(4), 15.
9. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience* 12(5), 535–540.
10. Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins* 23(3), ii–v.
11. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. (2018). The preregistration revolution. *PNAS* 115(11), 2600–2606.
12. Panickssery, A., Bowman, S. R. & Feng, S. (2024). LLM evaluators recognize and favor their own generations. *NeurIPS 2024*. arXiv:2404.13076.
13. Popper, K. (1959). *The Logic of Scientific Discovery*. Hutchinson.
14. Vul, E., Harris, C., Winkielman, P. & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science* 4(3), 274–290.
15. Wallach, I. & Heifets, A. (2018). Most ligand-based classification benchmarks reward memorization rather than generalization. *Journal of Chemical Information and Modeling* 58(5), 916–932.

Transmissions from the space between code and cosmos.

Christopher Yamba builds agentic AI systems for science and works as an AI engineer, entrepreneur and consultant. More essays at christopheryamba.com.

Christopher Yamba wrote this essay in collaboration with AI, which helped with drafting, simulations and illustrations. Every simulation is a short NumPy script with a fixed random seed, so each number can be reproduced.

© 2026 Christopher Yamba.