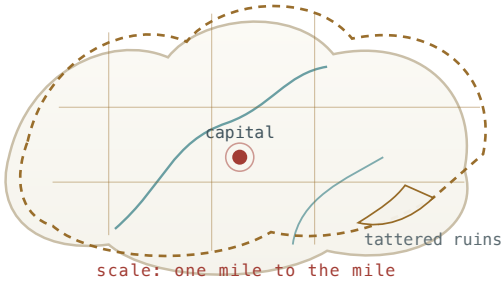
BLOCK I · FOUNDATIONS OF KNOWLEDGE & REASONING

Models, Maps & Idealization

A perfect map is a useless map. So what, exactly, is a good one for?



Borges imagined an empire that built a map the exact size of its territory. Their descendants let it rot.

In a single paragraph published in 1946, Jorge Luis Borges describes an empire whose cartographers grew so skilled, and so obsessed with exactitude, that they drew a map of the empire on the scale of **one mile to the mile**. It coincided with the territory point for point, lay over the country like a second skin, and became the most accurate map ever made. It was also worthless. The next generation abandoned it, and travelers found only "tattered ruins" in the desert.

The joke has teeth. A map is useful precisely because it is not the territory: it shrinks, flattens, selects, and throws nearly everything away. What survives is the thing you came for: structure you can use. Every scientific model performs the same trick. And once you see that, a harder question opens: when our best theories talk about things we can never see, are they describing reality, or only drawing very good maps?

CORE MODEL

SIMULATION FRONTIER ACTIVE

WHERE WE ARE

Yesterday (**Day 9**) we drew loops, bathtubs, and stability valleys. Every one of those diagrams was a model: a tool we trusted because it was simpler than the world. Today we turn the lens onto the tools themselves. We cash out **Day 7** on compression, reuse **Day 8** on descriptions that sit between order and randomness, and meet **Day 1** again in a new disguise: a model that predicts correctly for the wrong reason.

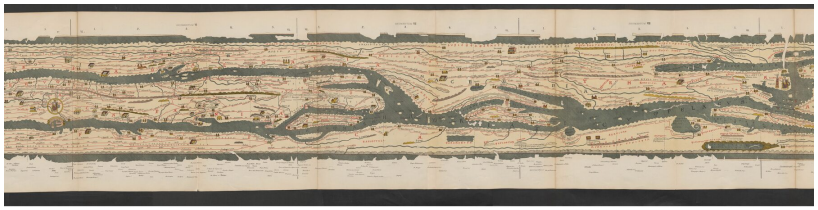
THE MODEL

The map is not the territory

The phrase belongs to Alfred Korzybski, who in 1933 was trying to cure a bad human habit: mistaking our words and symbols for reality itself. The full sentence matters:

> *"A map is not the territory it represents, but, if correct, it has a similar structure to the territory, which accounts for its usefulness."* Korzybski, Science and Sanity, 1933.

That clause about **similar structure** is the entire theory of modeling in miniature. A map is useless if it copies everything, as Borges saw, and useless if it copies nothing. It earns its keep in between, by preserving the relations that matter for a purpose while discarding the rest. A subway map is a triumph of distortion: distances are wrong, angles are wrong, rivers become cartoons. It keeps the order of stops and transfers, because that is what a rider needs.



The Tabula Peutingeriana is not trying to be a scaled geographic copy. Its distortion preserves what travelers needed: roads, stations, and sequence.

So a model is not a small true copy of the world. It is a **useful distortion**. Michael Weisberg's *Simulation and Similarity* gives the precise philosophical version: a model need not be identical to the target, or even perfectly isomorphic to it; it must be similar in the respects that matter, to the degree required by the job.

All models are wrong

George E. P. Box gave statistics its most portable sentence: "All models are wrong, but some are useful." The citation trail is a miniature hype-filter lesson. His 1976 paper *Science and Statistics* contains the seed: because all models are wrong, the scientist cannot obtain a correct one by endless elaboration. The polished aphorism appears as a section heading in a 1979 conference chapter, and the familiar "essentially" wording lands in the 1987 textbook he wrote with Norman Draper.

Three dates, one idea sharpening over a decade. The slogan is real, but the casual citation often attached to it is not. Check the receipts; we will do that at industrial scale on **Day 149**.

A joke older than Borges

Borges was not first to the one-to-one map. In Lewis Carroll's *Sylvie and Bruno Concluded* (1893), a character boasts of a national map on the scale of "a mile to the mile." Why was it never unrolled? The farmers objected, because it would block out the sun and kill the crops. So they use the country itself as its own map. The territory is always its own most accurate model, and also the least useful one.

THE CRAFT

The spherical cow, and why scientists lie on purpose

There is an old joke. A dairy farmer hires physicists to improve milk yield. Months later they return with a solution, but it works only for a **spherical cow in a vacuum**.

The joke is funny because it is true. Deliberately false assumptions are not a bug in science; they are one of its sharpest tools. Galileo's law of falling bodies is clean only for bodies in a vacuum, with no air resistance, no spin, and an ideal gravitational field. None of that is exactly true. Ernan McMullin called this *Galilean idealization*[1]: distort the problem to make the mechanism visible, then add detail back if the purpose demands it.

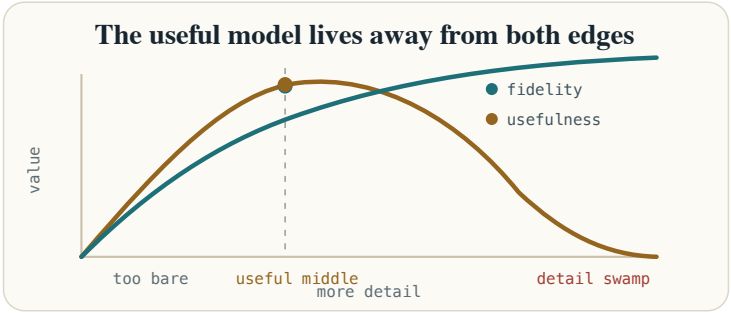
Sometimes the simplification is even deeper. A frictionless plane does not merely make the arithmetic easy; it reveals inertia by removing the thing that hides it. This *minimalist idealization* is not an approximation of the world so much as an argument about what matters. The spherical cow, the ideal gas, the point mass, the infinite population in genetics, and the frictionless pulley all say: ignore this noise and watch the structure that remains.

"All models are wrong" is therefore not a counsel of despair. It is a counsel of focus. A model that tried to be right about everything would be Borges's map: total, faithful, and inert. The art is choosing which falsehoods to commit to.

Static alternate · Idealization Dial

The live dial moves from severe abstraction to a one-to-one replica. Fidelity rises as detail increases, but usefulness peaks in the middle and collapses when the model becomes as complicated as the target.

[1] Galilean idealization deliberately distorts a problem, often by removing complicating factors, so the core relation becomes mathematically tractable.



A good model preserves the structure needed for a task rather than maximizing detail. Fidelity can keep rising; usefulness usually cannot.

REGION	EXAMPLE	LESSON
Too little detail	A point mass for a dairy cow	The simplification has removed the very property you needed.
Useful middle	A subway map or ideal gas	The model preserves the structure needed for the task.
Too much detail	Borges's one-to-one map	Maximum fidelity can destroy compression, clarity, and use.

THE DEBATE

Do our maps describe a real world, or just work?

Here the practical question becomes one of philosophy's deepest fights. Physics speaks of electrons, quarks, fields, and spacetime curvature: entities and structures no one sees directly. The theories work spectacularly. Does that success mean the unseen furniture is really there, more or less as described? Or are electrons entries in a powerful bookkeeping system, useful for predicting observations without telling us what reality is like underneath?

The *scientific realist* says success would be miraculous if our best theories were not latching onto reality. This is the no-miracles argument associated with Hilary Putnam and J. J. C. Smart: truth is the best explanation for sustained predictive success.

The anti-realist reply comes from history. Larry Laudan's 1981 paper *A Confutation of Convergent Realism* lists theory after theory that was successful in its day and false in ours:

- **Phlogiston**: the fire-stuff supposedly present in combustibles.

- **Caloric**: heat treated as an invisible fluid.
- **The luminiferous ether**: the medium light supposedly waved through.
- **Crystalline spheres**: the shells that carried the planets.
- **Vital forces**: the special spark of living matter.

The punch is brutal. Past success did not guarantee reference to real entities. Why assume our current success finally does? This is the *pessimistic meta-induction*[2]. Add *underdetermination*[3] and the realist has a serious problem: more than one map can fit the same ground.

Four ways to live with the tension

POSITION	OBSERVABLE PREDICTIONS	UNSEEN ENTITIES	MATHEMATICAL STRUCTURE
Scientific realism	True enough	Real, approximately as described	Tracks real relations
Structural realism	True enough	Hold loosely	The structure is what survives
Entity realism	True enough	Real when experimentally manipulated	Grand laws may hold only in models
Constructive empiricism	Empirically adequate	Stay agnostic	Accept the map, not its truth

Structural realism, associated with John Worrall, is the most map-like compromise. Look at the ether. Fresnel's ether disappeared; Maxwell's electromagnetic field replaced it. But parts of the mathematical structure survived the revolution. Maybe science accumulates structure more reliably than it accumulates furniture. Keep the equations; hold the entities loosely.

Entity realism, associated with Ian Hacking and Nancy Cartwright, moves from prediction to intervention. Hacking's slogan was that if you can spray electrons, they are real. Believe in entities you can manipulate to do work; be more skeptical of the grand theory around them. Cartwright drove the knife in from the other side in *How the Laws of Physics Lie*: fundamental laws govern idealized objects in models, not messy objects in the world.

Hold onto this table. Computers are about to force every row of it into the open.

[2]The pessimistic meta-induction argues that because many successful past theories later proved false, current successful theories may also turn out false.

[3]Underdetermination is the problem that more than one theory can fit the same evidence, so evidence alone may not select a unique true theory.

THE FRONTIER · 2026

The old fight, reborn in silicon

For most of history, science had two ways to learn: think about the world and poke the world. In the last few decades, a third has muscled in: **simulation**. A climate model runs a century in code. A jet engine's digital counterpart predicts a crack before it forms. A learned weather model out-forecasts a physics solver. What kind of knowledge is that?

Eric Winsberg argues that simulations have their own epistemology: they can produce genuine knowledge, justified by practices that are not reducible to either pencil-and-paper theory or laboratory experiment. Paul Humphreys named the unease *epistemic opacity*[4]. A process may be formally specified and still too large for any human to inspect step by step.

EDGE 01 · TWIN FRAMEWORK · HYPE CONTESTED

The digital twin: when does a model earn the name?

"Digital twin" is heavily marketed, so use the strict source: the 2024 U.S. National Academies report *Foundational Research Gaps and Future Directions for Digital Twins*. A real digital twin is not merely a simulation. It has a virtual representation, dynamic data updates from a physical counterpart, predictive capability, decision value, and a bidirectional virtual-physical loop. That feedback loop is the bar.

The hard problem is trust. The report emphasizes *VVUQ*[5] as a continual process, not a one-time stamp. The physical system changes, data change, decisions change, and the twin must be re-validated as it moves.

EDGE 02 · DESTINE LAUNCHED · SKILL GROWING · 2030 ROADMAP

A digital twin of the whole planet

The most ambitious version is Europe's **Destination Earth** program, built with ECMWF, ESA, and EUMETSAT. In June 2024, its first system release brought two twins online: Weather-Induced Extremes and Climate Change Adaptation. The extremes twin includes a global component at roughly 4.4 km resolution, with regional zooms for selected events.

That is an established launch milestone. It is not yet the science-fiction object implied by the phrase "digital twin of Earth." Higher resolution does not automatically mean a better forecast for every task, and the European goal of a full

[4] Epistemic opacity is a situation where no human can survey all the relevant steps by which a computational result was produced.

[5] VVUQ: verify code, validate model-world fit, and quantify uncertainty.

Earth-system twin by 2030 remains a roadmap.

EDGE 03 FFR-CT WORKS WHOLE PATIENT ASPIRATIONAL

Clinical digital twins: one real exception

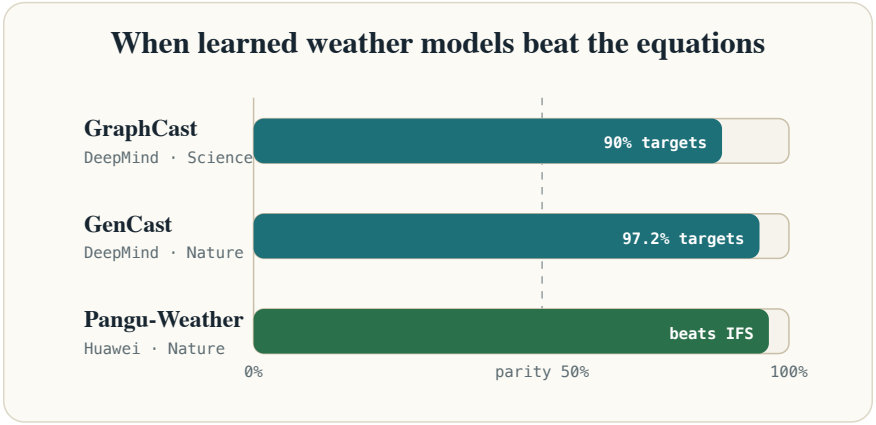
Medicine loves the idea of a digital twin of **you**: a running model of your body on which doctors could test interventions. The sober literature is far thinner. A 2025 scoping review in *npj Digital Medicine* found that only 18 of 149 healthcare studies using the label fully met the National Academies definition, and only two mentioned VVUQ.

The exception proves the rule. HeartFlow's FFR-CT takes a coronary CT scan and builds a patient-specific computational-fluid-dynamics model of blood flow to assess whether a blockage is starving the heart. It received FDA De Novo clearance in 2014. The scope is narrow, physical, and decision-specific. That is why it works.

EDGE 04 FORECAST SKILL PHYSICS CONTESTED

Forecasting weather without writing the equations

For decades, weather forecasting meant solving atmospheric physics on enormous computers. Then machine-learning models trained on past weather caught up with, and in several benchmarks surpassed, leading physics-based systems. GraphCast, Pangu-Weather, and GenCast are not press-release curiosities; their headline results were published in *Science* and *Nature*.



These are peer-reviewed hindcast targets. They establish forecast skill, not that the model's interior understands physics.

The philosophical sting is sharp. By the only measure a forecast user cares about, these learned systems can be more empirically adequate than the systems built from explicit physical equations. Yet their interiors are opaque weights, not named mechanisms. A constructive empiricist can shrug: prediction was the goal. A realist should worry: if prediction can arrive without interpretable representation, the no-miracles argument loses some force.

The structural realist points to the hybrid response. Google's NeuralGCM keeps a differentiable atmospheric dynamics core and uses machine learning for parts the physics handles badly. Keep the trusted structure; learn the closures. That is structural realism as engineering practice. But the caveat matters: like other learned systems, NeuralGCM is not expected to extrapolate far beyond its training regime. Mechanism may be exactly what you need at the edge of the map.

OPEN QUESTIONS

What's genuinely unsettled

- **Is simulation a third pillar of science, or applied math with better machines?** Winsberg and Margaret Morrison argue for a distinctive epistemology; critics still see expanded modeling.
- **Does predict-without-representing weaken scientific realism?** If a mechanism-free model can out-forecast a physical one, success alone looks less like evidence for truth.
- **Can learned models extrapolate?** The weather results are strongest inside regimes represented in training data. Novel climates and rare extremes are the empirical test.
- **When does a model deserve the name digital twin?** The National Academies drew a strict line. Many products using the phrase do not clear it.
- **What happens when the model is an AI system?** If a language model produces a true, well-supported claim, is it describing reality or only operating as an empirically adequate map? We return to this in the AI block.

The day in three sentences

Big idea

Every model is a deliberate, useful distortion: it preserves structure for a purpose while discarding nearly everything else. The hard question is whether our most successful maps describe a real world or merely save the phenomena.

Best analogy

Borges's one-to-one map and the spherical cow in a vacuum: perfect fidelity can be useless, while a false simplification can reveal exactly what matters.

Live controversy

Digital twins and AI weather models reopen scientific realism in silicon, especially when prediction succeeds without a human-readable mechanism. computation (simulation as a third mode of science) · information (models as lossy compression) · emergence (multiscale models and digital twins) · energy and evolution in the idealizations.

TOMORROW → DAY 11

Heuristics, Biases & Rationality

Today we watched science simplify the world on purpose. Tomorrow the lens turns inward: the mind itself runs on shortcuts. Meet Linda the bank teller, System 1 and System 2, and the rationality wars over whether heuristics are flaws to correct or adaptive bets under scarcity.

Sources & further reading

1. Korzybski, A. (1933). *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. Map-territory relation, p. 58. wikipedia.org/wiki/Map-territory_relation
2. Borges, J. L. (1946). "Del rigor en la ciencia"/ "On Exactitude in Science." wikipedia.org/wiki/On_Exactitude_in_Science
3. Carroll, L. (1893). *Sylvie and Bruno Concluded*, ch. XI. wikisource.org
4. Box, G. E. P. (1976). "Science and Statistics." *Journal of the American Statistical Association* 71(356): 791-799. doi:10.1080/01621459.1976.10480949
5. Box, G. E. P. (1979). "Robustness in the Strategy of Scientific Model Building," in Launer & Wilkinson (eds.), *Robustness in Statistics*, 201-236.
6. Box, G. E. P. & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley.
7. Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press.
8. McMullin, E. (1985). "Galilean Idealization." *Studies in History and Philosophy of Science* 16(3): 247-273.
9. Putnam, H. (1975); Smart, J. J. C. (1963). See *Stanford Encyclopedia of Philosophy*, "Scientific Realism." plato.stanford.edu/entries/scientific-realism
10. Laudan, L. (1981). "A Confutation of Convergent Realism." *Philosophy of Science* 48(1): 19-49. jstor.org/stable/187066
11. van Fraassen, B. C. (1980). *The Scientific Image*. Oxford University Press.
12. Worrall, J. (1989). "Structural Realism: The Best of Both Worlds?" *Dialectica* 43(1-2): 99-124.

13. Hacking, I. (1983). *Representing and Intervening.* Cambridge University Press.
14. Cartwright, N. (1983). *How the Laws of Physics Lie.* Oxford University Press.
15. Winsberg, E. (2010). *Science in the Age of Computer Simulation.* University of Chicago Press. Humphreys, P. (2009), on epistemic opacity.
16. National Academies of Sciences, Engineering, and Medicine. (2024). *Foundational Research Gaps and Future Directions for Digital Twins.* doi:10.17226/26894. nationalacademies.org/read/26894
17. European Commission / ECMWF. Destination Earth system launch, 10 June 2024. destination-earth.eu · destine.ecmwf.int
18. Tudor, Burton, et al. (2025). "A scoping review of human digital twins in healthcare applications and usage patterns." *npj Digital Medicine* 8: 587. doi:10.1038/s41746-025-01910-w
19. FDA. HeartFlow FFR-CT De Novo DEN130045; HeartFlow FFR-CT clearance materials, 2014.
20. Lam, R. et al. (2023). "Learning skillful medium-range global weather forecasting." *Science* 382(6677): 1416-1421. doi:10.1126/science.adi2336
21. Bi, K. et al. (2023). "Accurate medium-range global weather forecasting with 3D neural networks." *Nature* 619: 533-538. doi:10.1038/s41586-023-06185-3
22. Price, I. et al. (2024). "Probabilistic weather forecasting with machine learning." *Nature.* doi:10.1038/s41586-024-08252-9
23. Kochkov, D. et al. (2024). "Neural general circulation models for weather and climate." *Nature* 632: 1060-1066. doi:10.1038/s41586-024-07744-y

Source hygiene: future-dated arXiv identifiers surfaced during source research and were discarded as unreliable. The frontier claims above rely on peer-reviewed papers, official program pages, or institutional reports.

Optional appendix: The Cutting-Room Floor

The main descent ran a straight line: from Borges's empire-sized map, through the spherical cow, into the realism debate, and out the far end at a neural network that forecasts the weather without knowing a single law of physics. It was a clean arc, and it left an enormous amount on the cutting-room floor. The realism fight is four centuries older than van Fraassen. The craft of "lying usefully" has a precise grammar nobody mentioned. And the strangest question of all, **how do you trust a model whose innards no human can survey?**, loops all the way back to Day 1.

WHAT'S DOWN HERE

Six rooms the main tour skipped: **(1)** the older war between realism and instrumentalism; **(2)** the ether's one triumphant prediction, and how realism fights back with "divide and conquer"; **(3)** why the whole realism debate may be a base-rate fallacy, a straight callback to **Day 4**; **(4)** the real grammar of idealization, and Levins's "pick two" triangle; **(5)** what models actually are: mediators, fictions, and Maxwell's imaginary gears; and **(6)** the trust problem, from verification's 1994 bombshell to "computational reliabilism," which is just **Day 1**'s reliabilism wearing a lab coat.

ROOM I THE LONG PREHISTORY

Saving the phenomena: an argument four centuries deep

The main page framed realism versus instrumentalism as a twentieth-century affair: Putnam, Laudan, van Fraassen. But the fight is ancient, and for most of its life the instrumentalists were winning. The tradition even has a battle cry, handed down from Greek astronomy: *sozein ta phainomena*, "to save the phenomena." The job of a model, on this old view, was never to say what the heavens **are**, only to reproduce what they appear to do, to get the dots in the right places on the right nights.

Ptolemy's epicycles, wheels turning on wheels to explain the planets' loop-the-loops, were the original useful fiction. Many of his readers, and most medieval astronomers, treated them frankly as calculating devices. Nobody had to believe the sky was full of literal gears. The gears just had to work. That is instrumentalism in its cradle, a thousand years before the word existed.

The drama sharpened with Copernicus. When *On the Revolutions of the Heavenly Spheres* went to press in 1543, the Lutheran theologian Andreas Osiander slipped an unsigned preface into the front, without the dying Copernicus's blessing, assuring readers that the wild new sun-centered scheme need not be true, only a convenient hypothesis for computing positions. It was a peace treaty with the authorities,

written in pure instrumentalist ink: relax, it is just a better calculator.

The cardinal was an instrumentalist

The cleanest statement of the divide comes from Galileo's own prosecutor. In 1615, Cardinal Robert Bellarmine told the Copernican Paolo Foscarini that one could hold heliocentrism perfectly well *ex suppositione*, hypothetically, as a device that saves the appearances, but **not** as a claim about how the world physically is. Speak of it as a model, fine. Speak of it as the truth, and you have a problem. The Church's position was textbook instrumentalism. Galileo's heresy was realism: the insistence that the best model is telling you what is actually there. The same fault line runs straight through to the neural weather model on the main page.

By the early 1900s the physicist-philosopher Pierre Duhem, whom we met on **Day 2** for the "you can always blame an auxiliary" thesis, had turned this into a full doctrine. A physical theory, Duhem argued in *The Aim and Structure of Physical Theory* (1906), is not an explanation of hidden reality but an economical classification of experimental laws: a supremely compressed filing system for the phenomena. Truth about unseen causes was not the goal. Tidy, predictive bookkeeping was. Three hundred years of "save the phenomena," distilled. The modern instrumentalist owes Duhem the blueprint.

ROOM II REALISM STRIKES BACK

The ether's one perfect prophecy

The main page used the luminiferous ether as Exhibit A in Laudan's graveyard, a wildly successful theory whose central object turned out not to exist. But the ether's case is far more disturbing than a simple "it worked, then it died," and the reason is a single beautiful episode that the realist keeps in her back pocket.

Around 1818, the French Academy ran a prize competition on the nature of light. Augustin Fresnel submitted his wave theory: light as vibrations in the ether. On the judging panel sat Simeon Denis Poisson, a committed believer in the rival particle theory, who did some math with Fresnel's equations and triumphantly derived an apparently absurd consequence. If Fresnel were right, the shadow of a small round disc should have a **bright spot dead in its center**, where the waves bending around every edge arrive back in step. Ridiculous, said Poisson: a clear *reductio* that sinks the whole wave theory. Another judge, Francois Arago, went to the bench and actually did the experiment. The spot was **there**.

This is the realist's haymaker, and it has a name: a *novel prediction*[6]. The ether theory did not merely accommodate facts already known. Anyone can curve-fit the

[6] A novel prediction is a specific successful prediction of something not used to build or tune the model, especially when the result was surprising.

past. It predicted something nobody had ever seen and no one expected, a detail so specific and strange that its coming true looks like no coincidence at all. This is the no-miracles argument from the main page at full strength: surely a theory cannot pull a rabbit like that out of a hat unless it has hold of something real. And yet there is no ether. The single best evidence realism ever assembled was produced by a theory we now call false. The graveyard, it turns out, contains a genius.

Divide and conquer

So how does the realist survive her own best example? With a move sometimes called selective, or divide-and-conquer, realism, sharpened by Philip Kitcher and Stathis Psillos in the 1990s. The trick is to stop being a realist about whole theories and become one about parts. In any successful theory, only some posits actually do the predictive heavy lifting: the **working posits**. Others are **idle**, along for the ride. Be a realist only about the working parts.

Run the ether through this filter and watch it sort itself. What did the real work in deriving Poisson's spot? Not the substance "ether," not its density, its elasticity, or the question of what it was made of. The work was done by the **mathematical form of the wave equations**: the relationships, the structure. And those carried over into Maxwell's field theory and beyond. The idle metaphysical filler, the stuff of the ether, is exactly what got buried. So the pessimistic induction loses some of its teeth. Science did not keep the furniture and then throw it out; it kept the structure, which survived revolution, and quietly dropped the decoration. This is precisely the structural realism the main page ended on, now shown doing real work against its hardest case.

ROOM III THE CALLBACK TO DAY 4

Is the whole debate a base-rate fallacy?

Here is the room that should make a Day 4 graduate sit up. Recall the base-rate fallacy: a test that is 99% accurate for a disease that afflicts 1 in 10,000 still flags far more healthy people than sick ones, because you forgot to ask how common the disease is to begin with. Confusing P(positive | sick) with P(sick | positive) is the classic trap.

In 2004, the philosophers P. D. Magnus and Craig Callender pointed out that **both** sides of the realism war may be walking straight into it. Look again at the no-miracles argument. It wants to conclude that a successful theory is probably true, that is, that P(theory true | theory successful) is high. But by Bayes' theorem, you cannot get there from P(success | truth) alone. You also need the **base rate**: the prior proportion of true theories in the pool of candidate theories science draws from. And that number is not just unknown. Magnus and Callender argue that it is unknowable, because there is no fair way to count or sample all candidate theories that could have existed for mature sciences. No base rate, no inference.

The same blade cuts the pessimistic induction, which tries to push *P*(theory true | success) low from the historical record of failures, and trips over the same missing denominator. Magnus and Callender's diagnosis is almost clinical: framed probabilistically, the grand "wholesale" arguments are not just unresolved; they may be malformed, each a base-rate fallacy in a frock coat. Colin Howson made a closely related charge against no-miracles in 2000. The Day 4 machinery does not just describe this dispute. It may quietly dissolve it. LIVE PHILOSOPHICAL DISPUTE

ROOM IV THE GRAMMAR OF USEFUL LIES

Idealization has rules, and a famous tradeoff

The main page treated "deliberately false assumptions" as one thing. The literature is fussier, and the distinctions matter. First, separate two moves people lump together. *Abstraction* is leaving something out: modeling a planet's orbit without mentioning its color. True as far as it goes, just silent. Idealization is actively misrepresenting: declaring the plane frictionless when you know full well it is not. The frictionless plane is not a gap in the description. It is a fib. The two feel similar and behave very differently.

Then there are kinds of idealization, and the third one is the one the main page skipped. There is Galilean idealization, which distorts for tractability and promises to add detail back later, and minimalist idealization, the spherical cow strategy that keeps only the causal factors thought to matter. But there is also *multiple-models idealization*[7]: building several mutually incompatible models of the same thing, each false in a different direction, and triangulating among them. No single map is trusted. The agreement between differently wrong maps is.

> *"Our truth is the intersection of independent lies."* Richard Levins, "The Strategy of Model Building in Population Biology," 1966.

Levins's point, now called *robustness analysis*[8], is that if you attack a problem with several models built on different simplifying falsehoods, and they all spit out the same qualitative result, then that result probably depends on the real shared structure of the problem rather than on any one model's particular lie. The lies are independent. Their intersection is where truth is most likely hiding.

Levins is also the source of modeling's most famous impossibility claim: you cannot maximize **generality**, **realism**, and **precision** all at once. Push any two as far as they will go and the third has to give. The dial on the main page measured fidelity against usefulness. Levins's triangle is the same wisdom in three dimensions: pick your sacrifice.

[7] Multiple-models idealization deliberately uses several incompatible models of the same target, each simplified in a different way, to see what survives across them.

[8] Robustness analysis looks for results that persist across models with different simplifying assumptions.

Modeling tradeoff · Levins's triangle

No model maxes all three desiderata. The source version made this a click-through triangle; the static appendix keeps all three states visible for web, EPUB, and PDF readers.

SACRIFICE	KEEP	STRATEGY	EXAMPLE
Precision	Generality + realism	Robust qualitative models that capture the direction of change without decimal-point claims.	Theoretical ecology and toy economic or physical models that isolate a mechanism.
Realism	Generality + precision	Clean, exactly solvable models built on assumptions known to be false.	Frictionless planes, perfectly rational agents, infinite populations.
Generality	Realism + precision	High-resolution models of one specific target, accurate but narrow.	A digital twin of this aircraft, this lake, or this patient.

Orzack and Sober (1993) argued that the three-way tradeoff is not a strict theorem. Sometimes one desideratum can improve without another immediately falling. The disciplined version is therefore not "pick two forever." It is "know which dimension your current modeling move is spending."

ROOM V WHAT IS A MODEL, REALLY?

Mediators, fictions, and Maxwell's imaginary gears

Step back and ask the question the main page tiptoed around: when a scientist "has a model," what kind of thing is it? Three answers illuminate different corners.

A theory is a family of models. On the older syntactic picture, a scientific theory was a big set of sentences: axioms and their logical consequences. The modern *semantic view* says no. A theory is better understood as a collection of abstract structures, together with a claim that some real system is similar to one of them in specified respects. Newtonian mechanics is not fundamentally a wall of equations. It is a toolkit of idealized systems, the pendulum, the two-body orbit, the harmonic oscillator, plus the bet that bits of the world resemble them closely enough. Ronald Giere called the resulting stance *perspectival realism*[9]: our models are like instruments, each giving a partial, perspective-bound view, true to the world the way a map is true, never the way a mirror is.

[9] Perspectival realism says scientific representations can be genuinely world-directed while remaining partial, purpose-bound perspectives rather than mirror copies.

Models are mediators. In the influential 1999 collection *Models as Mediators*, Mary Morgan and Margaret Morrison argued that models are not simply read off from theory, nor simply summaries of data. They sit between, partly autonomous from both. You build a model the way you build an instrument, making pragmatic choices theory does not dictate; then, crucially, you learn by manipulating it, the way you learn by turning the knobs on an apparatus. A model is a tool you think with, not just a picture you look at. This is also why simulation, in the next room, can feel so much like experiment.

Models are fictions. The most provocative view takes the frictionless plane at its word: it is a fiction, an object that does not exist, described as if it did. The idea is old. Hans Vaihinger's 1911 *The Philosophy of "As If"* argued that whole swathes of science and mathematics run on useful falsehoods we knowingly treat as if true. Recall the Day 3 surprise that Sherlock Holmes's "deductions" are really abductions: here is its cousin. A scientific model may be closer to a character in a novel than to a photograph. The ideal gas is a bit like Sherlock: not real, never was, but reasoning about what it would do delivers real understanding.

The man who built a machine he knew was false

The patron example sits next to the main page's Fresnel-to-Maxwell story. To derive his immortal equations of electromagnetism, James Clerk Maxwell first built an outrageous mechanical contraption: a space packed with spinning "molecular vortices," separated by little rolling "idle wheels." He did not believe space was full of gears. He called it an illustration, a fiction to think with. Out the other end came Maxwell's equations: structure that was true, midwifed by a model that was false. His mentor Lord Kelvin took the creed to the hilt: "I never satisfy myself until I can make a mechanical model of a thing." The fiction was scaffolding. The structure was the keeper. That is the whole of Day 10 in one Victorian anecdote.

Cartwright's dappled world

One more turn of the screw deepens the entity realism from the main page. Nancy Cartwright, who gave us "the laws of physics lie," built that thought into a whole worldview in *The Dappled World* (1999). Her claim: fundamental laws are true only inside models, and they describe the real world only where we have painstakingly engineered, or luckily found, a *nomological machine*[10]: a shielded, stable, repeating arrangement of parts that produces lawlike regularity. A physics lab is such a machine. So is the solar system, roughly.

But most of the world is not shielded or stable, and there the tidy laws simply do

[10] A nomological machine is a stable arrangement that reliably produces lawlike behavior, such as an experimental setup or a well-isolated natural system.

not reach. They hold only *ceteris paribus*, all else equal, and all else is never equal. The world, for Cartwright, is not a pyramid with grand laws at the base but a patchwork: a dappled quilt of little domains where different models work and between which nothing universal runs. Read her way, "all models are wrong" stops being a remark about our maps and becomes a claim about reality: the laws themselves were idealizations all along. LIVE PHILOSOPHICAL POSITION

ROOM VI THE TRUST PROBLEM

Verifying the unsurveyable, and a bombshell from 1994

Engineers who stake lives on simulations drew a hard distinction long ago. *Verification*[11] asks: did we **solve the equations right?** Is the code a faithful numerical solution of the model we wrote down? *Validation*[12] asks the harder thing: did we **solve the right equations?** Does the model actually correspond to the world? Solve-it-right versus solve-the-right-thing. Almost every dispute about trusting a model is really a dispute about validation.

And then, in 1994, three researchers detonated a small bomb under the word itself. In *Science*, Naomi Oreskes, Kristin Shrader-Frechette, and Kenneth Belitz argued that for models of **open natural systems**, climate, groundwater, ecosystems, verification and validation are impossible in principle. Their reasoning lands on three course threads at once. Natural systems are never closed, so you can never control all the inputs. Model solutions are non-unique, so a good fit never proves your model is the right one, which is the underdetermination from the main page. And confirming a model by matching its predictions to data is, strictly, the fallacy of affirming the consequent from **Day 3**: if my model is right, I will see X; I see X; therefore my model is right. Their verdict on the word "validation" is severe: it falsely implies a legitimacy the model cannot earn. The primary value of models, they argued, is heuristic. They are instruments for thinking, not certificates of truth.
ESTABLISHED ARGUMENT

So how do we trust the black box?

This is where the main page's neural weather model comes back to haunt us. If you cannot validate a model of an open system, and you cannot even survey the billions of weights inside a learned emulator, Humphreys's epistemic opacity from the main page, then on what grounds could you ever trust its forecast?

Juan Duran and Nico Formanek proposed one answer in 2018: *computational reliabilism*[13]. Stop and feel the click. Reliabilism was Day 1: the idea that a belief is

[11] Verification asks whether the code or calculation correctly solves the model that was specified.

[12] Validation asks whether the model corresponds well enough to the real target system for the intended purpose.

[13] Computational reliabilism treats simulation output as justified when it is produced by a process with

justified not because you can recite an argument for it from the inside, but because it was produced by a reliable process, such as good vision or sound memory. Duran and Formanek point that externalist answer at simulations. You do not need to see inside the opaque model to trust its output; you need evidence that the process producing it is reliable.

That evidence looks like a track record of verification and validation where checking is possible, robustness across independent methods, a history of the technique succeeding where it could be checked, and the judgment of experts who know its failure modes. Trust migrates from transparency to reliability. The black box can be known the way a witness can be known: not by cracking open the skull, but by examining the record. The thing that defined a justified belief in a human brain on Day 1 turns out to define a trustworthy answer from a machine on Day 10.

● PRINCIPLED, STILL DEBATED

BONUS ROOM

Where "digital twin" actually came from

The main page leaned on the 2024 National Academies definition: bidirectional, continually updated, predictive. Where did that idea come from, and what makes the strict definition matter?

The ancestor flew on **Apollo**. NASA built physical twins of its spacecraft and kept them on the ground. When Apollo 13's oxygen tank ruptured 200,000 miles from home, engineers fed live telemetry into ground-based simulators and physical mockups to improvise fixes for a craft they could not touch. A model on Earth, updated with data from its endangered counterpart in space, used to make a life-or-death decision: that is the platonic form of a digital twin, decades before the phrase existed.

The modern concept was laid out by **Michael Grieves** in 2002, in a University of Michigan product-lifecycle lecture: a slide already showing real space, virtual space, and data flowing both ways between them. The catchy name came later. NASA's **John Vickers** coined "Digital Twin" around 2010, and it entered NASA's technology roadmap. From about 2014, Siemens, Dassault, GE, ANSYS, and the rest stamped it across their marketing, which is exactly why a sober definition became necessary, and why the National Academies drew the line where they did.

That line is best seen as a three-rung ladder, following Kritzinger and colleagues (2018): what earns the name "twin" is not how detailed the model is, but **how the data flows**.

The taxonomy is deliberately austere. A plain digital model can be enormous and beautiful; a true twin can be visually boring. The difference is the loop.

---

a strong reliability record, even if the full process is opaque.

Diagram in prose · what earns the name

The only thing that changes across the three is the data link between the physical object and its virtual copy. Dashed means manual or absent; solid means automatic.

RUNG	DATA LINK	WHAT IT IS
Digital model	No automatic link	A CAD drawing or stand-alone simulation. You update it by hand, if at all. Many marketed twins are really just this.
Digital shadow	One-way automatic link	Live sensor data flows from the physical object into the model, but the model cannot act back. Think live dashboard.
Digital twin	Bidirectional automatic link	Data flows both ways: the model mirrors the object and steers it. Only this rung earns the strict name.

BONUS ROOM

A model of a model of a model

One last vista reframes the whole frontier. The main page set up a contrast between clean physics-based models and messy black-box machine-learning emulators, sharper in the telling than in reality. Two facts blur it.

First: "physics-based" climate models are not pure first principles either. They cannot resolve a cloud or a thunderstorm, because those are far smaller than the grid, so they fake the small stuff with *parameterizations*[14]: simplified stand-ins whose knobs are tuned until the model reproduces the climate already observed. Climate scientists openly call this "the art and science of climate model tuning." A model hand-tuned to match known climate, then trusted to predict an unknown one, is leaning on exactly the kind of fit that Oreskes warned cannot validate anything. The principled model has fudge in it too.

Second: when a simulation is itself too expensive to run thousands of times, scientists build a cheap statistical stand-in for the simulation, called an *emulator*[15] or surrogate. Follow the chain: the world is approximated by a **simulation**, a model of the world, which is approximated by an **emulator**, a model of the model. Each rung is faster, cheaper, and a little more wrong. The AI weather models from the main page

[14] Parameterizations are simplified formulas that stand in for processes too small or complex to represent directly in a simulation grid.

[15] An emulator, or surrogate, is a cheaper model trained to approximate the outputs of a more expensive simulation.

are simply the newest, most powerful rung of that ladder: surrogates that learned to mimic decades of atmosphere directly. "All models are wrong" turns out to be recursive. We routinely build wrong models of our wrong models, and the trick, every time, is knowing which lie is good enough for the job in hand.

Zoom all the way out and the day sits on a four-step staircase that the computer scientist Jim Gray called the **four paradigms** of science.

The arc · how science has learned to know

Each paradigm did not replace the last. It stacked on top. Today's AI emulators are the fourth, and the realism-versus-instrumentalism question rides up the staircase with them.

PARADIGM	VERB	WHAT IT DOES
1. Empirical	describe	Observe and record the world: star charts, notebooks, instruments.
2. Theoretical	explain	Compress observations into laws and equations: Newton, Maxwell, and the birth of the realism debate.
3. Computational	simulate	When equations are unsolvable by hand, run them on a machine: the third mode, with its own epistemology.
4. Data-driven	learn	Let models learn patterns directly from oceans of data, often with no explicit theory inside: GraphCast, GenCast, NeuralGCM.

The fourth rung is precisely where an instrumentalist feels at home and a realist feels uneasy, which is why this thousand-year-old argument suddenly matters for a weather app.

CODA

The aphorism, audited

Let us give the day's slogan the same scrutiny we gave everything else. "All models are wrong, but some are useful" is true, but it has hardened into a thought-terminating cliche, a phrase people reach for to wave away the hard question instead of answering it. The real questions are the ones it skips: wrong **how**? Wrong **for what**? Some models are vastly wronger than others, and the difference between a useful map and a misleading one is the entire game. Statisticians like Andrew Gelman have pushed back precisely here. The aphorism, taken as a shrug, can excuse sloppy work and dull the urge to check.

George Box himself was no nihilist about it. The same papers that gave us "all models are wrong" warn just as hard against the opposite error: mistaking the model for the world, or "worrying about the wrong things" because you have fallen in love with your own elaboration. The discipline he was pointing at is not "models are wrong, so relax." It is the harder thing this course is built around: **calibrated judgment about the specific gap between this map and this purpose**. A good modeler is not someone who has stopped lying. It is someone who knows exactly which lie they are telling, and exactly what it is good for.

The appendix in three sentences

Big idea: the realism-versus-instrumentalism fight is four centuries old, may be a base-rate fallacy on both sides, and ultimately turns on a trust question answered not by seeing inside a model but by judging the reliability of the process that made it.
Best new analogy: Maxwell deriving true equations from a clockwork of imaginary gears he knew were false: fiction as scaffolding, structure as keeper, paired with Levins's "intersection of independent lies."
Sharpest callback: computational reliabilism is literally Day 1's reliabilism aimed at simulations: you trust an opaque model the way you trust a witness, by its record rather than by cracking it open.
computation, simulation, and emulation as modes three and four of science · information as compression and robustness as independent evidence · emergence in Cartwright's patchwork world and tuning across scales · all braided back through Days 1-4.

Sources & further reading

1. Duhem, P. (1906). *The Aim and Structure of Physical Theory*. Theory as economical classification, not explanation. See also Duhem, *To Save the Phenomena* (1908).
2. Osiander's anonymous preface to Copernicus, *De revolutionibus orbium coelestium* (1543); Cardinal Bellarmine's 1615 letter to Foscarini. See *Stanford Encyclopedia of Philosophy*, "Scientific Realism." plato.stanford.edu/entries/scientific-realism
3. The Arago/Poisson bright spot: French Academy prize competition, c. 1818-1819; Fresnel's wave theory and the predicted-then-observed spot in a circular shadow. en.wikipedia.org/wiki/Arago_spot
4. Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. Routledge. Divide-and-conquer realism; working versus idle posits. Kitcher, P. (1993). *The Advancement of Science*.
5. Magnus, P. D. & Callender, C. (2004). "Realist Ennui and the Base Rate Fallacy." *Philosophy of Science* 71(3): 320-338. philpapers.org/rec/MAGREA-2 See also Howson (2000), *Hume's Problem*.
6. Levins, R. (1966). "The Strategy of Model Building in Population Biology." *American Scientist* 54(4): 421-431. Generality/realism/precision tradeoff, robustness, and "our truth is the intersection of independent lies." Critique: Orzack & Sober (1993), *American Naturalist* 148:201.
7. Morgan, M. & Morrison, M., eds. (1999). *Models as Mediators*. Cambridge University Press.

8. Giere, R. N. (1988/2006). *Explaining Science*; *Scientific Perspectivism*. University of Chicago Press. The semantic/model-based view and perspectival realism.
9. Vaihinger, H. (1911). *Die Philosophie des Als Ob* (*The Philosophy of "As If"*, English trans. 1924). Useful fictions. See also Frigg & Hartmann, "Models in Science," *Stanford Encyclopedia of Philosophy*. plato.stanford.edu/entries/models-science
10. Maxwell, J. C. (1861-62). "On Physical Lines of Force." *Philosophical Magazine*. The mechanical vortex and idle-wheel model of the electromagnetic field.
11. Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge University Press. Nomological machines, ceteris paribus laws, and the patchwork world.
12. Oreskes, N., Shrader-Frechette, K. & Belitz, K. (1994). "Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences." *Science* 263(5147): 641-646. doi:10.1126/science.263.5147.641. science.org
13. Duran, J. M. & Formanek, N. (2018). "Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism." *Minds and Machines* 28(4): 645-666. doi:10.1007/s11023-018-9481-6. link.springer.com
14. Grieves, M. (2002 concept; 2014 white paper "Digital Twin: Manufacturing Excellence through Virtual Factory Replication"); John Vickers/NASA technology-roadmap usage around 2010; Grieves & Vickers, "Origins of the Digital Twin Concept"(2016).
15. Kritzinger, W. et al. (2018). "Digital Twin in manufacturing: A categorical literature review and classification." *IFAC-PapersOnLine* 51(11): 1016-1022. Digital model / digital shadow / digital twin by data-flow automation.
16. Kennedy, M. C. & O'Hagan, A. (2001). "Bayesian calibration of computer models." *Journal of the Royal Statistical Society B* 63(3): 425-464. Gaussian-process emulators and surrogates of expensive simulators.
17. Hourdin, F. et al. (2017). "The Art and Science of Climate Model Tuning." *Bulletin of the American Meteorological Society* 98(3): 589-602.
18. Hey, T., Tansley, S. & Tolle, K., eds. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research. Jim Gray's empirical -> theoretical -> computational -> data-driven framing.
19. On "all models are wrong" as cliche: Andrew Gelman, "Statistical Modeling, Causal Inference, and Social Science"(blog, various), and the nuance in Box's own 1976/1979 papers.

Source hygiene: as on the main page, several post-cutoff or future-dated arXiv identifiers surfaced during source research and were **not** relied upon; every source above is a dated publication, institutional page, or encyclopedia entry.

Optional appendix: The Edge of the Map

Old mapmakers had an honest habit: where surveys stopped and rumor began, they drew monsters and wrote *hic sunt dracones* — here be dragons. The main day and Appendix I mapped relatively settled country: Borges, Bellarmine, verification, reliabilism, and the old realism fight. This appendix sails off the edge. Most of what follows is from 2020 or later, much of it from the last few years, and a good share of it will turn out to be wrong. That is the point. The frontier is where calibration matters most.

READ THIS FIRST

Every claim below carries a tag. ESTABLISHED means peer-reviewed, reproduced, or hardware-validated. PROMISING means credible but early, narrow, or still preprint-shaped. CONTESTED OR HYPE-FRAMED means the evidence is thin, the framing is inflated, or independent review has already deflated the claim. The Day 10 skill — calibrated judgment about how wrong a model is, and for what — is exactly the skill needed to read this frontier.

THE SHIFT

From building a model to summoning one

The deepest post-2020 change is not any single result. It is a change in what a scientific model is made of. For four centuries, the usual recipe was explicit: choose idealizations, write equations, solve them. Since about 2022, a rival recipe has been spreading. Train a giant neural network on a staggering amount of domain data until it absorbs the statistical shape of the field, then fine-tune that one model for many concrete tasks.

These systems are *foundation models*[16], and importing them into science is one of the decade's major stories. Wang and 41 co-authors' 2023 *Nature* review, "Scientific discovery in the age of artificial intelligence," reads like a flag planted on new territory. INFLUENTIAL REVIEW

The flagship Earth-system instance arrived in May 2025: **Aurora**, a 1.3-billion-parameter model from Microsoft Research and collaborators, pretrained on more than a million hours of geophysical data. It is not tuned for one job; it is a general pretrained atmosphere that can be fine-tuned for air quality, ocean waves, tropical-cyclone tracks, and high-resolution weather. The paper reports that Aurora matches or beats numerical air-pollution simulations on 74% of targets, wave models on 86% of variables, and official five-day tropical-cyclone track

[16] A foundation model is a large pretrained model that can be adapted to many downstream tasks, rather than being trained from scratch for one narrow use.

forecasts across several agencies, including the U.S. National Hurricane Center basins. Its authors frame this as the first time a machine-learning model surpassed full operational tropical-cyclone forecasts up to five days. PEER-REVIEWED RESULT

Notice what this does to the main page's realism debate. Aurora is constructive empiricism industrialized: a model with no human-readable meteorology inside can still be empirically adequate for important forecasting tasks. The realist's discomfort — *a model that predicts without representing in a surveyable way* — is now a deployed research program, not a classroom puzzle.

The same pattern is spreading. NASA and IBM's **Prithvi WxC**, released in 2024 as open weights plus preprint, is a 2.3-billion-parameter weather-climate foundation model trained on MERRA-2 variables and tested across forecasting, downscaling, gravity-wave flux parameterization, and extremes. The gap between released and validated is the reason the hype filter exists. PROMISING PREPRINT

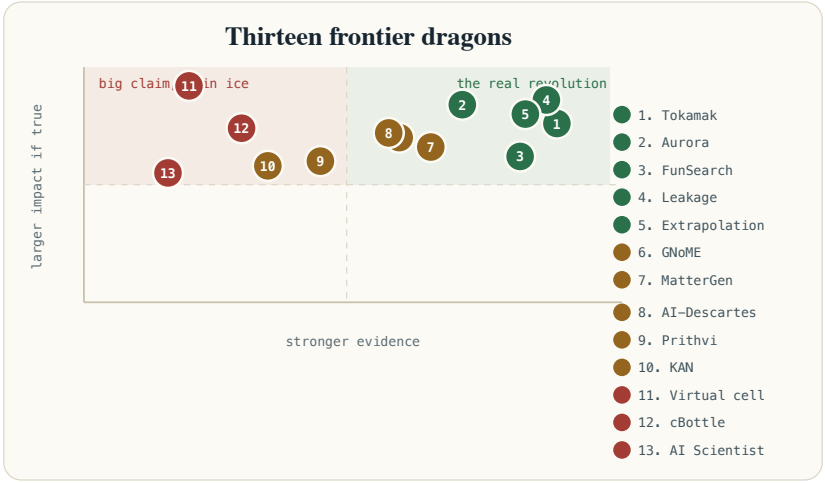
THE MAP OF THE FRONTIER

Thirteen dragons, sorted

The useful question is not "is this impressive?" The useful question has two axes: **how established is it**, and **how much would it matter if it held up?** The upper-right is the real revolution: big and solid. The upper-left is where the dragons live: enormous claims resting on thin ice.

Frontier map · impact if true x evidence

Positions are judgment calls, not measurements. The table keeps the source interactive's receipts visible for web, EPUB, and PDF readers.



The positions are a calibration exercise, not a measurement. Green work toward the right is comparatively solid; the upper-left claims are still dragons.

WORK	STATUS	WHY IT SITS THERE
Tokamak plasma control	established	Deep RL shaped plasma on TCV and reduced tearing-instability risk on DIII-D: real hardware, high impact.
Aurora Earth-system model	established	Peer-reviewed multi-task forecasting skill, including five-day operational cyclone tracks; whole-Earth-model rhetoric still needs restraint.
FunSearch	established	LLM plus evaluator found new verifiable mathematical constructions, making the output inspectable rather than opaque.
Leakage crisis	established	Kapoor and Narayanan found leakage failures across 17 fields and 294 papers; its impact is deflationary but field-shaping.
Extrapolation drift	established	NeuralGCM and ACE2 show the hard limit: skill inside the training climate is not guaranteed beyond it.

GNoME materials discovery	promising	The Nature result is real; the practical novelty, usefulness, and synthesizability of headline counts remain contested.
MatterGen	promising	Peer-reviewed generative materials work with a synthesis proof point, but broad materials-design payoff is early.
AI-Descartes / AI-Hilbert	promising	Strong realist direction: recover laws from data plus background theory. Most examples still rediscover known laws.
Prithvi WxC	promising	Open 2.3B-parameter weather-climate model; still preprint-era evidence across downstream tasks.
KANs	promising	A 2024 preprint and 2026 PRX follow-up argue for interpretable scientific discovery; broad superiority over ordinary networks remains unsettled.
AI Virtual Cell	contested	A serious Cell roadmap, not a functioning cell twin. The ambition is huge; the built artifact is not here.
cBottle / Earth-2	contested	A credible generative climate-emulator preprint wrapped in digital-planet and compression marketing.
The AI Scientist	contested	The autonomous-scientist framing was sharply deflated by independent evaluation: failed experiments, novelty errors, hallucinated results.

DISPATCH 1 · SURROGATES AT THE CONTROLS

The machine now runs the machine

Some frontier models are already load-bearing in the most literal sense: trusted with expensive hardware and unstable plasma. In 2022, DeepMind and EPFL published a reinforcement-learning controller that shaped the magnetic field of a real **tokamak fusion reactor**, holding plasma in target configurations on the TCV machine. In 2024, a follow-up used deep reinforcement learning to reduce the risk of tearing instabilities on the DIII-D tokamak. When a learned model is allowed to steer fusion plasma in real time, “just a model” is no longer a useful dismissal.

HARDWARE-VALIDATED

The same dispatch also carries a warning. *Physics-informed neural networks*[17] were pitched as a bridge between explicit equations and learned models: bake the differential equation into the loss function, then let the network learn a solution. The failure-mode literature shows the catch. A PINN can drive its residual loss down while the actual solution remains badly wrong. The model can satisfy every constraint you wrote down and still miss the world, because you did not write down all the constraints that matter. FAILURE MODES ESTABLISHED FIXES PROMISING

DISPATCH 2 · THE REALIST COUNTERATTACK

Discover the law — do not just fit it

If the black-box emulator is instrumentalism's triumph, a quieter counter-movement fights for the realist. Its demand is simple: do not settle for a model that predicts; make the machine hand you a law you can read. This is *symbolic regression*[18] and theory-guided discovery.

Three specimens matter. **AI-Descartes** combines symbolic regression with logical reasoning over background axioms and rederived laws such as Kepler's third law from data plus theory. **AI-Hilbert** pushes the idea through polynomial optimization and formal certificates. **FunSearch** pairs a language model with an automated evaluator and evolutionary search; in *Nature*, it found new constructions for the cap-set problem and new bin-packing heuristics. That last case matters philosophically because the output is not a weight vector. It is an object humans can inspect, test, and prove against. FUNSEARCH ESTABLISHED

The caveat keeps this from becoming a realist victory lap. Many systems still mostly rediscover laws we already had, which is powerful validation but thin discovery. The KAN line is similarly promising but unsettled: the 2024 Kolmogorov-Arnold Networks preprint proposed learnable functions on network edges as a more interpretable alternative to MLPs, and a 2026 *Physical Review X* follow-up argues for scientific-discovery uses. That is real momentum. It is still not proof that KANs generally beat well-tuned ordinary networks. PROMISING KAN ROLE DEBATED

Mind the gap · paper measured vs press implied

The frontier's biggest distortions live in the space between a careful result and its announcement.

CASE	THE PAPER MEASURED	THE PRESS IMPLIED

[17] Physics-informed neural networks train a neural model while penalizing violations of known equations or boundary conditions.

[18] Symbolic regression searches for compact mathematical expressions that fit data, ideally yielding an interpretable law rather than an opaque predictor.

Aurora	Multi-task forecast gains against operational systems on air quality, waves, cyclone tracks, and weather.	A foundation model of the whole Earth. It is better described as a powerful forecasting and emulation model.
GNoME	Large-scale predicted-stable crystal structures from graph-network screening.	An order-of-magnitude expansion of useful known materials. Novelty and synthesizability need harder auditing.
cBottle / Earth-2	A generative diffusion emulator for kilometer-scale global atmosphere fields, currently preprint-stage.	A digital twin of the planet, with vendor-framed compression claims doing extra rhetorical work.
AI Virtual Cell	A roadmap for priorities, data, evaluation, and collaboration needed to build AI virtual cells.	A functioning digital twin of a living cell already exists.
The AI Scientist	An autonomous-research preprint whose independent test found coding failures, poor novelty judgments, and hallucinated results.	A new era of cheap fully autonomous scientific discovery.

DISPATCH 3 · THE RESULT THAT SHOULD SCARE YOU

Leakage, and the reproducibility crisis in AI-for-science

The most important result in this appendix is a warning. In 2023, Sayash Kapoor and Arvind Narayanan published "Leakage and the reproducibility crisis in machine-learning-based science" in *Patterns*. Their survey found leakage errors across 17 fields, collectively affecting 294 papers. *Data leakage*[19] makes a model look brilliant on paper because the evaluation data have partially entered the training pipeline.

This is Day 2's replication crisis reborn for the AI age. When a model is evaluated on information it has effectively already seen, its "accuracy" is Borges's one-to-one map of the training set: perfect and useless. ● LANDSCAPE-SHAPING CAUTION

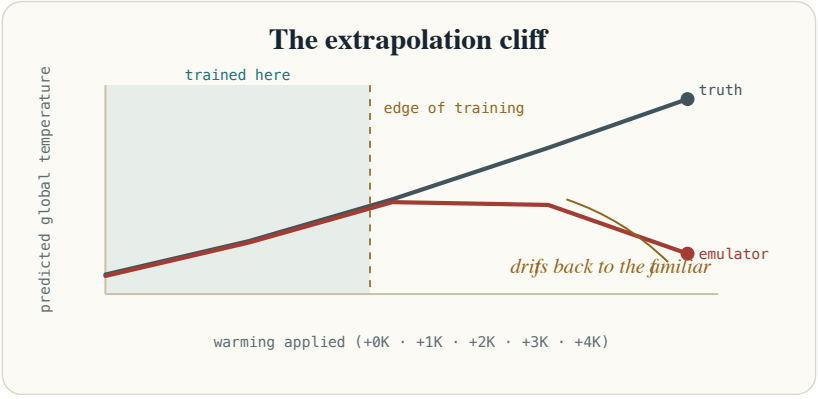
Leakage inflates performance inside the data you have. Its twin problem appears when the world moves outside the climate the model saw. NeuralGCM's authors report that the emulator stayed sensible at +1 K and +2 K sea-surface-temperature perturbations, but at +4 K its warming response diverged from expectations. ACE2, from the Allen Institute team, is explicit that sensitivity to separately changing sea surface temperature and carbon dioxide is not entirely realistic. This is the empiri-

[19] Data leakage occurs when information from the test target or future data sneaks into training, validation, preprocessing, or feature selection.

cal heart of the realism worry: a pattern learner can be a superb interpolator and an unreliable prophet. EXTRAPOLATION FAILURE ESTABLISHED

Diagram · predictive skill is not mechanism

A schematic of the failure mode: excellent inside the training climate, then drift toward the familiar as the model is pushed somewhere it has not seen.



This is a schematic, not a reproduction of a paper figure. It captures the failure mode flagged by NeuralGCM and ACE2: skill inside the training climate is not mechanism outside it.

DISPATCH 4 · THE NEW PHILOSOPHY

What can an unreadable model let you understand?

The philosophers have not been idle. Emily Sullivan's 2022 "Understanding from Machine Learning Models" gives the cleanest reframing: what blocks understanding is not black-box complexity by itself, but *link uncertainty*[20] — weak evidence that connects the model to the target it claims to be about. A transparent model of the wrong target gives you nothing; an opaque model with strong empirical links may still support understanding. Opacity is not the core problem. Evidence is.

Others divide the terrain differently. Florian Boge separates opacity of training from opacity of representation and warns that deep learning can open a gap between discovery and explanation. Ramon Alvarado argues that AI is an epistemic

[20] Link uncertainty is uncertainty about whether the model is empirically connected to the target system in the right way.

technology, a tool that operates on knowledge itself and deserves its own epistemology. Sabina Leonelli keeps attention on the data ecosystem: these models inherit the opacity, bias, gaps, and politics of the datasets that feed them. The upshot is that AI-for-science may be forcing a third epistemic category: not theory, not experiment, and not quite ordinary simulation either. ACTIVE, UNSETTLED PHILOSOPHY

DISPATCH 5 · THE AUDACIOUS DRAGONS

A virtual cell, a planet twin, and an AI that does science

Now the upper-left of the map: claims so large that, if they land, they redraw everything, and which today rest on thinner ice.

The AI Virtual Cell. In 2024, a large team spanning Stanford, the Chan Zuckerberg Initiative, Genentech, Google, and others published a *Cell* roadmap for building multi-scale AI models of living cells. The ambition is serious: a virtual cell that can support interpretable *in silico* experiments across biological scales. But the paper is a priorities-and-opportunities roadmap, not a built cell twin. VISION AND ROADMAP

The planet twin. NVIDIA's **cBottle** is a generative model for kilometer-scale global-atmosphere fields: a coarse global generator plus local super-resolution, trained over simulation and reanalysis data. The preprint is credible and useful. The "digital twin of the planet" aura around Earth-2 is marketing language doing work that the evidence has not earned. PREPRINT PLANET-TWIN FRAMING IS HYPE

The autonomous scientist. Sakana AI's 2024 "AI Scientist" preprint pitched an end-to-end automated researcher: idea, experiment, paper, and review. Independent evaluation in 2025 found a harsher reality: five of twelve proposed experiments failed because of coding errors, novelty judgments were poor, and some results were hallucinated or misleading. The result is not useless; the autonomous-scientist framing is exactly what the contested tag is for.

INDEPENDENTLY DEFLATED

The materials-discovery cautionary tale

One story compresses the whole appendix. In 2023, Google DeepMind's **GNoME** announced 2.2 million crystal structures, including hundreds of thousands predicted stable. The result was real and impressive; a peer-reviewed ACS perspective soon asked how many were genuinely new, useful, or synthesizable. In 2025, Microsoft's **MatterGen** shifted the goalposts productively from listing candidates to generating materials under target constraints, with a reported synthesis proof point. The arc is the frontier in one breath: real result, inflated headline, sober correction, progress underneath. Believe the papers; audit the press; wait for synthesis. REAL RESULTS SCOPE CONTESTED

OPEN QUESTIONS

The experiments that would settle it

The right posture is not belief or dismissal. It is knowing which result would change your mind.

- **Validated warming extrapolation.** If an ML climate emulator reproduces a genuine warming response against held-out high-emission simulation or paleo-climate, the instrumentalist case strengthens. Right now, +4 K drift says caution.
- **Leakage-audited replication.** Independent, leakage-checked reproduction of a flagship foundation-model science claim would tell us how much of Kapoor and Narayanan's warning applies at the top of the field.
- **A genuinely novel discovered law.** A symbolic-regression system that finds a new, subsequently confirmed law of nature would be a landscape shift. FunSearch's cap-set result is the closest analogue in pure math.
- **A validated clinical digital twin.** A peer-reviewed, independently validated patient or organ twin that clears the National Academies bar would move medicine's dragons out of the upper-left.

The appendix in three sentences

Big idea: since 2020, modeling has been reorganized around foundation models and learned surrogates that can predict well without human-readable representation, while symbolic discovery tries to make machines produce inspectable laws.
Sharpest caution: data leakage has already inflated published ML-for-science results, and even strong climate emulators can drift when pushed beyond their training regime; predictive skill is not mechanism.
Live controversy: whether deep-learning models can explain rather than merely predict, with the AI Virtual Cell, planet-twin claims, and autonomous-scientist systems as spectacular but still under-evidenced bets. computation as a modeling substrate · information as compression, leakage, and evidence links · emergence in multi-scale cell and climate models · Day 10's realism/instrumentalism divide stress-tested on 2020s hardware.

Sources & further reading

1. Wang, H. et al. (2023). "Scientific discovery in the age of artificial intelligence." *Nature* 620:47-60. doi:10.1038/s41586-023-06221-2
2. Bodnar, C. et al. (2025). "A foundation model for the Earth system"(Aurora). *Nature* 641:1180-1187. doi:10.1038/s41586-025-09005-y

3. Schmude, J. et al. (2024). “Prithvi WxC: Foundation Model for Weather and Climate.”[preprint] arXiv:2409.13598; NASA/IBM open model release.
4. Degrave, J. et al. (2022). “Magnetic control of tokamak plasmas through deep reinforcement learning.” *Nature* 602:414-419. doi:10.1038/s41586-021-04301-9
5. Seo, J. et al. (2024). “Avoiding fusion plasma tearing instability with deep reinforcement learning.” *Nature* 626:746-751. doi:10.1038/s41586-024-07024-9
6. Karniadakis, G. E. et al. (2021). “Physics-informed machine learning.” *Nature Reviews Physics* 3:422-440. See also Krishnapriyan, A. et al. (2021), “Characterizing possible failure modes in physics-informed neural networks,” *NeurIPS 2021*.
7. Romera-Paredes, B. et al. (2023). “Mathematical discoveries from program search with large language models”(FunSearch). *Nature* 625:468-475. doi:10.1038/s41586-023-06924-6
8. Cornelio, C. et al. (2023). “Combining data and theory for derivable scientific discovery with AI-Descartes.” *Nature Communications* 14:1777. doi:10.1038/s41467-023-37236-y
9. Cory-Wright, R. et al. (2024). “Evolving scientific discovery by unifying data and background knowledge with AI-Hilbert.” *Nature Communications* 15:5922. doi:10.1038/s41467-024-50074-w
10. Liu, Z. et al. (2024). “KAN: Kolmogorov-Arnold Networks.”[preprint] arXiv:2404.19756. See also “Kolmogorov-Arnold Networks Meet Science,” *Physical Review X* (2026).
11. Kapoor, S. & Narayanan, A. (2023). “Leakage and the reproducibility crisis in machine-learning-based science.” *Patterns* 4(9):100804. doi:10.1016/j.patter.2023.100804
12. Kochkov, D. et al. (2024). “Neural general circulation models for weather and climate”(NeuralGCM). *Nature* 632:1060-1066. doi:10.1038/s41586-024-07744-y
13. Watt-Meyer, O. et al. (2025). “ACE2: Accurately learning subseasonal to decadal atmospheric variability and forced responses.” *npj Climate and Atmospheric Science* 8:205. See also arXiv:2411.11268.
14. Sullivan, E. (2022). “Understanding from Machine Learning Models.” *British Journal for the Philosophy of Science* 73(1):109-133. See also Grote, T., Genin, K. & Sullivan, E. (2024), “Reliability in Machine Learning,” *Philosophy Compass* 19(5):e12974.
15. Boge, F. J. (2022). “Two Dimensions of Opacity and the Deep Learning Predicament.” *Minds and Machines* 32:43-75. See also Duede, E. (2023), “Deep Learning Opacity in Scientific Discovery,” *Philosophy of Science* 90(5):1089-1099.
16. Alvarado, R. (2023). “AI as an Epistemic Technology.” *Science and Engineering Ethics* 29:32. doi:10.1007/s11948-023-00451-3. See also Leonelli, S. (2023), *Philosophy of Open Science*, Cambridge University Press.
17. Bunne, C. et al. (2024). “How to build the virtual cell with artificial intelligence: Priorities and opportunities.” *Cell* 187(25):7045-7063. doi:10.1016/j.cell.2024.11.015
18. Brenowitz, N. et al. (2025). “Climate in a Bottle: Towards a Generative Foundation Model for the Kilometer-Scale Global Atmosphere.”[preprint] arXiv:2505.06474; NVIDIA Earth-2/cBottle release materials.
19. Lu, C. et al. (2024). “The AI Scientist.”[preprint] arXiv:2408.06292. Independent evaluation: Beel, J., Kan, M.-Y. & Baumgart, S. (2025), “Evaluating Sakana’s AI Scientist,” arXiv:2502.14297 / *ACM SIGIR Forum*.
20. Merchant, A. et al. (2023). “Scaling deep learning for materials discovery”(GNoME). *Nature* 624:80-85. doi:10.1038/s41586-023-06735-9. Zeni, C. et al. (2025). “A generative model for inorganic materials design”(MatterGen). *Nature* 639:624-632. doi:10.1038/s41586-025-08628-5

Source hygiene: the frontier is noisy, so the appendix uses dated publications, institutional model releases, and identifiable preprints rather than unsupported headline claims. Blog and vendor framing are treated as framing, not as evidence.