AI Scientists Need a Theory of Scientific Value

Yunlong Xu
University of Chicago
dracoxu@uchicago.edu

September 8, 2026

Abstract

An AI scientist must decide which questions deserve an answer and what would count as an advance. Current systems learn these judgments from scientific work, expert evaluations, and task-specific measures of success. As they take greater responsibility for choosing research, the grounds for those judgments become part of the design problem. We argue that AI for science should make the development of a theory of scientific value a foundational research task. Such a theory should explain why a contribution is worth recognizing and support judgments that can be checked against explicit principles and evidence. New relationships and stronger grounds for existing conclusions offer concrete starting points. They also reveal why value depends on what is already known, why some contributions resist a common scale, and why scientific advances can demand new standards. Developing this theory would make the aims of AI research open to systematic justification, formal investigation, and empirical challenge.

1 Introduction

What are scientists being trained to pursue? Learning to do science includes learning to judge which questions matter, when an experiment is decisive, and whether a result changes how a subject should be understood. Scientists acquire much of this judgment through examples, criticism, and participation in communities of expertise [Polanyi, 1962]. A correct answer may be trivial; an investigation that fails to answer its question may reveal why the available evidence could never settle it. Recognizing these differences is part of scientific expertise. Explaining why they deserve different evaluations is a problem about the value of science itself.

For AI scientists, such evaluations help determine what research gets done. A Scientific Judge trained on citation-based comparisons supplies preferences for training a Scientific Thinker [Tong et al., 2026]. Co-Scientist develops hypotheses through automated criticism and tournament selection at inference time, with expert and experimental assessments of selected outputs [Gottweis et al., 2026]. The AI Scientist combines automated research and review with evaluation by human conference reviewers [Lu et al., 2026].

In materials discovery, the purpose of an evaluation determines which prediction errors matter. Matbench Discovery evaluates whether models can identify stable inorganic crystals, explaining why errors near a stability boundary matter differently from errors averaged across all materials [Riebesell et al., 2025]. DiscoveryBench asks whether an agent can recover verified hypotheses from data, comparing their context, variables, and relationships [Majumder et al., 2025].

Problem selection also shaped the research process behind OpenAI's proposed proof of the Navier–Stokes Millennium Prize Problem [OpenAI, 2026]. OpenAI reports that, after its agents produced a result on the related Euler equations, which describe fluid motion without viscosity, the team judged Navier–Stokes more promising and redirected agents from other Millennium problems toward it. An agent taking on such decisions would need to assess both what a result has established and what makes the next investigation worth pursuing.

Learning from experts offers one way forward, but a choice among judgments can still require a choice among scientific values. An evaluator that rewards new answers and one that rewards stronger evidence may agree on everything a study did while assigning it different merit. More accurate imitation of either evaluator would reproduce the difference. Deciding between them requires reasons for valuing the achievements each recognizes. Human evaluation of AI research, AI evaluation of human research, and AI evaluation of other AI systems share this problem.

AI for science should make the development of a theory of scientific value a foundational research task. Greater ability to pursue an objective makes the choice of that objective more consequential. A theory would explain why particular contributions matter and provide principles for comparing them, including contributions that existing evaluation tasks did not anticipate. Its comparisons could be criticized, revised, and, where possible, supported by independently checkable arguments. An assessment can be justified within its stated scope while capturing only part of a contribution's value, or depending on what is already known. Its use in choosing further research therefore needs justification of its own. Explicit principles and arguments would let us examine an evaluation's grounds even when the research exceeds what one person can retrace. The importance of different questions, the reach of formal guarantees, and the revision of scientific standards would become research problems in the design of AI scientists.

2 What Should a Theory of Scientific Value Explain?

The Hodgkin–Huxley account of the nerve impulse illustrates how a study can transform what is known about a system. Voltage-clamp experiments held the voltage across a nerve membrane fixed and measured the current needed to maintain it. These measurements supported estimates of membrane conductances: how readily ionic current passes through the membrane. Equations assembled from those relationships reproduced the action potential, the electrical pulse of a nerve impulse, and its propagation [Hodgkin and Huxley, 1952]. The model connected conductances to phenomena beyond the conditions used to estimate them, making these connections available for calculation and investigation while leaving their microscopic basis open.

We take finding out what is true of a subject, and how its properties depend on one another, to be a central aim of science. From this commitment, we defend two connected sources of scientific value: establishing previously unavailable answers or relationships, and improving the reasons for accepting them. A newly established relation reveals a dependency. An improved experiment can resolve an alternative explanation that previously undermined a conclusion. A method that distinguishes previously confounded mechanisms expands what research can find out. Each makes a claim about the subject more informative, better supported, or newly answerable, giving it scientific value before practical application or widespread approval.

Improving the grounds for an existing answer therefore deserves credit even when the answer itself is unchanged. A target counting only new answers would overlook that achievement. Whether stronger justification itself constitutes scientific progress or helps bring it about is disputed [Bird, 2016, Dellsén, 2021]. We regard a genuine improvement in protection against error as scientifically valuable in its own right. A study can earn credit for that improvement even if its particular conclusion later proves false; the credit belongs to the improvement, while the false claim loses its standing as established content. If the supposed improvement itself depended on a false premise, its assessment must also change. The distinction between actual achievement and our grounds for judging it therefore remains essential [Niiniluoto, 2025].

What is already available changes the contribution. Constructing the first accessible proof of a conjecture can be a major advance even if the conjecture was always a logical consequence of accepted axioms. Copying that proof adds neither a mathematical relation nor independent support. The relevant background includes evidence, proofs, and methods available to the community whose advance is being assessed. Making an existing result usable by another community can add capability without establishing its content anew. Treating either community as already possessing every logical consequence of its knowledge would erase much of the work science requires.

Research can also establish a limit rather than the answer it set out to obtain. Showing that an experiment cannot distinguish two proposed mechanisms tells us what a new design must overcome. Such a result is an achieved contribution to inquiry. A mistaken conjecture may instead stimulate a later discovery: the later contribution deserves its own assessment, while its success does not

retrospectively answer the original question. Valuing a research process requires identifying what it actually established along the way. Maieusis, for example, records data and design obstacles before analysis begins, making reasons to revise or stop a proposed investigation available for assessment [Xu and Doiron, 2026].

2.1 Scientific content and its representation

Relationships can be obscured by the form in which a result is summarized. In a model of a rhythm-generating neural circuit in crustaceans, an approximate joint posterior described plausible combinations of 31 conductance parameters [Gonçalves et al., 2020]. This probability distribution was inferred from recorded rhythm features, conditional on the model and prior assumptions. It allowed investigation of how parameters can vary together while sustaining the rhythm. A list of individual parameter ranges can miss precisely these dependencies. An evaluation that treats the list and the joint account as equivalent would overlook part of the scientific content at issue.

Even agreement on the full distribution of every pair of variables can hide a relation among all three. For $a, b, c \in \{-1, +1\}$, the uniform distribution over states satisfying $abc = 1$ has the same pairwise marginals as the uniform distribution over all eight states, where $abc = 1$ occurs half the time (Appendix C). Any two variables look independent in both distributions, but in the restricted one, knowing two fixes the third. Sampling a, b independently and uniformly and setting $c = ab$ also preserves the restricted distribution. The scientific question is which relations a representation preserves, with empirical evidence establishing whether they apply to the target.

An AI-generated model may preserve such relations without admitting an intuitively transparent decomposition. Cao and Yamins defend mechanistic explanations that depart from familiar forms of intelligibility [Cao and Yamins, 2024a,b]. The explanatory claim depends on an appropriate correspondence between model and target, including their organization and activities. Explaining an established relationship in a form people can readily grasp can add understanding and make the result easier to use. The relationship deserves scientific credit even before that accessible explanation is available.

2.2 From contribution to importance

Recognizing a contribution still leaves its importance to explain. A theory that counts new true statements would reward trivial additions and miss the force of unification. Relating surface gravity and orbital periods through a common gravitational parameter for the same central body in a Newtonian regime, for example, constrains how previously separate phenomena can vary together. Renaming that relationship as ten questions would not multiply its content. The relevant gain lies in the connection it establishes across the questions.

Connections of this kind provide reasons to value explanatory scope, and complementary discoveries may jointly establish something neither establishes alone. Independent questions present a harder problem. Why should resolving a mathematical conjecture take priority over explaining spatial representation? Their difficulty, popularity, or number of consequences cannot settle this by itself. A theory must explain which differences deserve weight and whether some contributions remain incomparable. It must also distinguish credit for an established contribution from a reason to explore a promising direction: research selection involves uncertainty, cost, and possible achievements as well as existing ones.

3 From Scientific Contributions to Checkable Judgments

Closing an experimental loophole can strengthen the grounds for a conclusion that earlier studies already supported. Bell tests make it possible to specify both the scientific change and a reason to value it.

3.1 What improves when an experimental loophole is closed?

Bell tests ask whether correlations between separated systems can be explained by local mechanisms: each outcome depends only on its laboratory's setting and shared hidden variables independent of the setting choices. Quantum theory predicts correlations that can exceed the limits of such

explanations. In the CHSH form of the test, each laboratory chooses between two measurement settings and records an outcome of $+1$ or -1. Each correlation C_{xy} is the mean product of the two outcomes for settings x, y: agreement contributes $+1$, disagreement -1. Local models obey $|S| \leq 2$ for $S = C_{00} + C_{01} + C_{10} - C_{11}$. Counting only detected trials can select different shared hidden states at different settings, however [Larsson, 2014]. The resulting correlations can exceed that bound while retaining a local explanation.

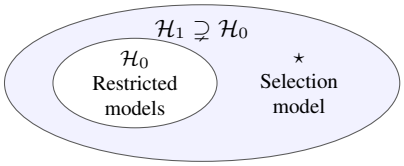
(a) Correct score, surviving local explanation

Settings xy	00	01	10	11
Retention probability	$1/4$	$1/4$	$1/4$	$1/4$
Retained correlation	$3/5$	$3/5$	$3/5$	$-3/5$

$S = \frac{3}{5} + \frac{3}{5} + \frac{3}{5} - (-\frac{3}{5}) = \frac{12}{5} > 2$

Local detection selects the retained trials.

(b) Broader evidential coverage



The broader guarantee covers a live alternative omitted by the restricted guarantee.

Figure 1: A correctly computed score and a broader evidential guarantee. (a) An illustrative local model gives the displayed correlations exactly after detection selects the retained trials. Its score exceeds 2 while a local explanation survives. (b) In a separate comparison of two valid, informative studies of the same target, a guarantee over $\mathcal{H}_1$ covers a live alternative omitted by a guarantee restricted to $\mathcal{H}_0$. The gain concerns this wider coverage at a common error level; other aspects of the studies may differ.

Figure 1a illustrates the problem: every setting pair retains one quarter of the trials, yet a wholly local model produces a score of 2.4. In this construction, each laboratory reports a detection only when its local setting matches a label carried by a shared hidden variable. Each setting pair therefore selects a different part of the hidden population; no communication between the laboratories is needed (Appendix A). These are exact model correlations. Finite experimental records fluctuate, so a real test must also control the probability of a spurious violation.

The experiment of Hensen et al. [2015] addressed detection and locality loopholes together. A signal of successful state preparation defined valid trials without being causally influenced by the setting choices. Each valid trial received a binary outcome at each laboratory, with the choices and readouts timed to exclude communication between the laboratories during measurement. Their observed correlations could then count against a wider range of local explanations.

Consider two studies of the same target, each with informative evidence and a valid argument at the same declared error level. The first guarantee limits how often its test wrongly rejects mechanisms in a restricted class $\mathcal{H}_0$; the second covers a wider class $\mathcal{H}_1$ that includes a relevant sampling mechanism omitted by the first (Figure 1b). Each protocol induces its own distributions of possible records, for which its guarantee must hold. A test that rejects on a data-independent coin flip can satisfy a false-positive bound while distinguishing nothing, so informative evidence is an essential additional requirement. Appendix A gives a sufficient discrimination condition.

Against a common scientific background, the second study now supplies previously unavailable grounds for the conclusion: its inference survives a live alternative that could have misled the first. We regard overcoming that vulnerability as a scientific gain because it removes a substantive reason to doubt the conclusion. A broader guarantee addressing a genuine threat earns credit for its evidential coverage; deleting an irrelevant premise does not. Costs, precision, and other achievements remain relevant to the value of the projects as a whole.

An evaluator must preserve this distinction if it is to recognize improvements in the grounds of a result. DiscoveryBench deliberately evaluates the recovery of verified relationships and gives reasons for assessing outcomes within that task [Majumder et al., 2025]. Suppose an agent receives an analogous answer-recovery reward for two successful investigations, one retaining a live sampling assumption and one overcoming it. Both add independent evidence, yet returning the same relation gives no extra reward for the latter's wider coverage. If that study is costlier, a policy optimizing the reward minus cost could favor the first. Giving the coverage gain explicit credit would allow it to be weighed against cost.

3.2 A judgment that carries its grounds

A model could submit an argument together with its evaluation. An independent checker could then examine whether the judgment follows from specified principles and premises. Proof-carrying code establishes this separation between generating a result and checking an accompanying proof against a specification [Necula, 1997]; scientific verification and structured arguments offer related resources for research claims [Cornelio et al., 2025, Ma, 2026]. Evaluating scientific value adds a further responsibility: justifying why the property being checked constitutes an achievement worth recognizing.

One possible form for such a judgment is

$$\Gamma; F; K \vdash \pi : \ A \succeq_{d,Q} B. \tag{1}$$

Given scientific background K, premises Γ, and value principles F, a derivation π establishes that study A is no worse than B for questions Q on aspect d. In the Bell case, the proposed derivation would certify the wider error guarantee and apply the principle that removing the specified vulnerability merits credit for its coverage. Experiments and modeling must establish whether the premises apply; scientific and philosophical argument must justify the principle. Putting it into F makes that dependence explicit and open to challenge. A complete proof would establish the stated comparison conditional on these commitments.

An evaluator might dispute a sampling assumption, find an error in the derivation, or reject the proposed reason for crediting the difference. Formal checking can give a strict, conditional guarantee while leaving the empirical premises and value principles open to revision. Building such judgments requires a scientifically faithful representation of the study and evidence for its premises. Obtaining these inputs and constructing an argument may cost more than checking it.

4 Building a Theory for Open-Ended Research

4.1 Comparisons before a universal scale

The Bell comparison fixes a question and an aspect of achievement. Broader comparisons need not begin with a common numerical score. Blackwell's theory compares experiments by the information they can provide [Blackwell, 1953, Fritz et al., 2023]. If a single, possibly randomized transformation of observations from A reproduces B's observation distribution for every possible true state, then A permits every decision procedure available after B. For finite experiments, this transformation can be specified by a matrix of probabilities and checked directly. It proves that the smallest attainable average loss with A is no worse for any finite decision problem and any distribution over the possible states, setting aside acquisition and processing costs (Appendix B).

For questions about the same target, preserving all existing inferential capabilities while adding a relevant one gives a reason to recognize an advance in that respect. Different aspects of a target can resist this comparison. For two independent binary quantities x, y, an experiment revealing only x and one revealing only y cannot simulate each other; one revealing both can simulate either. Their information relation is rigorous even though it supplies no exchange rate between learning x and learning y. Such a reflexive, transitive relation is a preorder; grouping equivalent experiments produces a partial order. Proven incomparability is distinct from equal value and from a comparison still awaiting an answer.

A theory of scientific importance must ask whether further reasons can complete these comparisons. A real-valued score orders every pair of assigned numbers, so it cannot exactly represent an incomplete comparison. Quantitative theories such as Fanelli's account of knowledge in terms of information gained, explanatory resources, and description costs provide substantive candidates to investigate [Fanelli, 2019]. The value choices embodied in such measures must be examined along with their mathematical properties.

Consider a study unifying several well-supported relationships and another removing a serious vulnerability in one foundational result. A principle favoring breadth of supported dependencies might prefer the first; one prioritizing the security of widely relied-on inferences might prefer the second. Investigating the disagreement requires specifying what the unification connects and which inferences the vulnerability threatens. It also requires defending why those differences deserve their

proposed weights. Inflating the number of restated consequences or invoking an irrelevant threat would undermine the respective arguments. Where both contributions remain substantial, a theory may support the competing reasons without yet settling their balance. Comparing rival principles on such disputed cases tests more than agreement on obviously valuable work.

The possibility of a general theory also depends on what we require of it. Clear meanings, computability, checkable proofs, coverage of every research contribution, and a total ranking are different demands. Both impossibility results and successful constructions would advance this research. Approximation offers a further route: errors in simulating one experiment from another can be translated into bounds on decision performance, with model error adding further conditions (Appendix B). A guarantee for approximate scientific importance would likewise need to specify the value being approximated and the domain over which its error is controlled.

4.2 Standards that scientific work can revise

Scientific work can challenge what an evaluation recognizes. Suppose an evaluator treats an intuitively transparent decomposition as necessary for crediting any model-based scientific contribution. A model that establishes supported dependencies without such a decomposition challenges that criterion, as discussed in Section 2. Recognizing those dependencies as scientific content expands what earns credit while preserving the value of illuminating human explanations.

Adding a justified dimension can revise an overall judgment while preserving valid local comparisons. Suppose two studies score 2 and 1 on dimension a, but 0 and 3 on a newly admitted dimension b. A rule requiring no loss on any included dimension previously ranked the first above the second; now they are incomparable. Its superiority on a survives. Formal rules can specify which judgments a revision retains or retracts, drawing on the logic of theory change [Alchourrón et al., 1985]. A framework for scientific value should allow justified additions without requiring every future dimension to be listed in advance. Admitting a new dimension requires a reason to value the scientific difference it captures and an account of which earlier conclusions remain valid.

4.3 From assessments to research objectives

A partial comparison can constrain research choices while leaving several options open. Decisions among incomparable contributions may also depend on cost, risk, or a commitment to exploration; an allocation decision need not imply that one contribution has greater scientific value. Optimizing a single scalar reward requires a further commitment to how the relevant scientific differences should be traded off. Multiobjective learning provides a precedent for retaining several policies under partially ordered returns, with methods depending on the allowed trade-offs and policy class [Roijers et al., 2013]. The number of alternatives and uncertainty about their consequences can make this computationally demanding.

Even with an agreed scalar objective, correct judgments can misdirect selection. If two studies have values 100 and 2, while certificates establish only lower bounds of 1 and 2, maximizing the certified bound chooses the less valuable study. Unequal tightness of the certified bounds has reversed the value ranking, illustrating a problem of reward optimization within formal evaluation itself [Skalse et al., 2022].

A theory should seek guarantees connecting a score to the value sought. An approximation within ϵ for every outcome being compared ensures that a score gain exceeding 2ϵ improves the specified value. Lower and upper bounds establish a preference when $L(B) > U(A)$: even the lower bound for B exceeds the upper bound for A. Overlapping intervals leave the comparison unresolved (Appendix D). Work beyond the reach of current certification may be valuable, so allocation policies also need ways to support its exploration.

What an investigation adds changes as research proceeds. Suppose establishing x has initial value 1 and establishing y has value 0.9, with value 1.9 for obtaining both. Reusing the initial scores would reward establishing x twice more than establishing both, even if the second occurrence merely copies the same result and evidence. Independent replication earns different consideration because it can add support. Evaluation needs to track the change in what the research program has established, including contributions that acquire value in combination.

Materials discovery already confronts a related dependence. Matbench Discovery discusses how updating the reference hull—the stability boundary defined by reference structures—can change the labels of earlier candidates; its fixed initial hull makes evaluation tractable at the cost of this limitation [Riebesell et al., 2025]. Where a scalar value for a scientific background is justified, one candidate is to reward each study's addition to it. Over a finite undiscounted sequence under a fixed theory, these increments sum to the change in assessed value (Appendix D). Potential-based rewards have an established theory [Ng et al., 1999]; defining the scientific value and representing its background are the substantive tasks here. Revising the value theory can change an old assessment, which is a different event from producing new research.

Even a justified account of cumulative value must become usable feedback for training and search. Previously established comparisons could supply training examples or search constraints where their stated scope, background, principles, and premises still apply. Choosing unfinished investigations involves uncertainty about their contributions; evidence may arrive too late to guide the decisions that produced it. A usable system must also decide when to spend resources obtaining a stronger assessment and when to pursue further research. Cheaper learned estimates offer one route, but need testing under the optimization they guide. Developing losses, verifiers, and selection rules that preserve scientific distinctions at workable cost is part of the research agenda.

Table 1: Questions for a research program on scientific value. Each comparison tests a claim about what deserves credit or how that credit should guide further inquiry.

Research problem	Concrete comparison	What the comparison tests
Value of contributions	Broadening supported relationships versus strengthening a foundational inference.	How rival principles justify their weights, and where comparison remains unresolved.
Checkable evaluation	Judgments under exact, approximate, and empirically uncertain premises.	Which guarantees survive, and which assumptions require further investigation.
Research objectives	Selection under changing backgrounds, complementary discoveries, and unequal tightness of certified bounds.	When increasing the reward advances research and when it suppresses valuable exploration.

Cases such as these can test a value theory before it is entrusted with directing research (Table 1). Liu and Zhai already test novelty metrics under controlled changes to the reference background, using axiomatic expectations at the level of systems [Liu and Zhai, 2026]. A broader investigation should state rival value principles before applying them, establish the scientific differences between cases, and examine the reasons for divergent judgments. The resulting evaluations can then be tested in research selection against independently argued criteria, counting evaluation costs against the available research budget. Expert disagreement is material for this inquiry, rather than a label to average away: it can reveal a disputed fact, an unsupported value weight, or a consequence that a principle cannot defend.

5 Alternative Views

Scientific judgment may work because it exceeds explicit rules. An expert's sensitivity to an important problem can depend on experience that resists articulation. Polanyi describes scientific appraisal as distributed across overlapping specialties, sustained through professional standards and personal judgment [Polanyi, 1962]. A learned evaluator might inherit some of this sensitivity. A formal theory could instead favor the most easily stated parts of scientific judgment and give them disproportionate authority. If so, making the criteria explicit could impoverish the science they guide.

The Bell comparison shows how an explicit principle can constrain a judgment while leaving much of that expertise intact. Once the target and reliability requirements are fixed, the gain in the coverage of evidence can be examined independently of who endorses it. An evaluator that gives no credit for removing this vulnerability owes an account of what defeats that reason. Other considerations may defeat it in an overall comparison, but the gain remains a consideration to address. A partial theory can make some judgments answerable to reasons while leaving others dependent on expertise. Its practical contribution should be tested against capable expert and learned evaluators, including its costs and treatment of unfamiliar advances.

Scientific progress may transform the standards used to recognize it. A new concept can change what counts as an explanation or a fruitful question. A fixed evaluator may then reject the very achievement that warrants changing it. Permission to revise creates a dilemma: restrictive admission rules can reproduce old exclusions, while unrestricted revision lets a system redefine success to suit its output. Merely labeling work as unassessed will not ensure that promising research receives attention or resources.

Revision within a framework can follow its existing rules. Replacing those rules must justify a change in what the framework recognizes as valuable. The replacement should explain which achievements the old framework misses and which earlier judgments retain their force. Rival revisions can be compared on cases where they preserve, exclude, or spuriously reward contributions. Lee's account of critical engagement supports scrutiny of both evaluation purposes and standards of reliability [Lee, 2026]. Formal rules help establish the consequences of a change; the case for making it must come from scientific and philosophical argument.

The role of community understanding makes this disagreement concrete. Tao emphasizes the work required to digest mathematical results and incorporate them into collective knowledge [Tao, 2026]. A contribution can establish a relation before it has been fully assimilated, while the later work of explanation and dissemination also deserves credit. How they should be weighted, and how inquiry should support unfamiliar work while its value is disputed, bears directly on the allocation of research effort.

6 Conclusion

As AI scientists take on more of the decisions that shape research, their evaluations will help determine which questions are asked and which advances are pursued. Scientific value deserves sustained study alongside scientific capability. Understanding what makes a scientific contribution valuable, and how such judgments should change as science advances, must become part of the work of developing AI scientists.

Use of AI tools

The central ideas and supporting arguments in this paper originated with the human author, who directed the research and writing process. LLMs were used to critique the ideas and arguments, and coding agents assisted with the literature review and figure preparation. Coding agents and LLMs accessed through web interfaces provided writing assistance, helping express the author's written materials, ideas, and arguments in English and editing and revising the wording and paragraph structure. LLM agents also read completed versions from different disciplinary perspectives. Their feedback informed revisions. The author takes responsibility for the scientific content and the final text, figures, and references.

References

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985. doi: 10.2307/2274239.

Alexander Bird. Scientific progress. In Paul Humphreys, editor, *The Oxford Handbook of Philosophy of Science*, pages 544–563. Oxford University Press, 2016. doi: 10.1093/oxfordhb/9780199368815.013.29. URL `https://www.alexanderbird.org/Research/Papers/Scientific_Progress_2_REV.pdf`.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Rosa Cao and Daniel Yamins. Explanatory models in neuroscience, part 2: Functional intelligibility and the contravariance principle. *Cognitive Systems Research*, 85:101200, 2024a. doi: 10.1016/j.cogsys.2023.101200.

Rosa Cao and Daniel Yamins. Explanatory models in neuroscience, part 1: Taking mechanistic abstraction seriously. *Cognitive Systems Research*, 87:101244, 2024b. doi: 10.1016/j.cogsys.2024.101244.

Cristina Cornelio, Takuya Ito, Ryan Cory-Wright, Sanjeeb Dash, and Lior Horesh. The need for verification in AI-driven scientific discovery. arXiv:2509.01398v2, 2025. URL https://arxiv.org/abs/2509.01398v2. Preprint, revised December 17, 2025.

Finnur Dellsén. Understanding scientific progress: The noetic account. *Synthese*, 199:11249–11278, 2021. doi: 10.1007/s11229-021-03289-z.

Daniele Fanelli. A theory and methodology to quantify knowledge. *Royal Society Open Science*, 6(4):181055, 2019. doi: 10.1098/rsos.181055.

Tobias Fritz, Tomáš Gonda, Paolo Perrone, and Eigil Fjeldgren Rischel. Representable Markov categories and comparison of statistical experiments in categorical probability. *Theoretical Computer Science*, 961:113896, 2023. doi: 10.1016/j.tcs.2023.113896. URL https://arxiv.org/abs/2010.07416v3.

Pedro J. Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F. Podlaski, Sara A. Haddad, Tim P. Vogels, David S. Greenberg, and Jakob H. Macke. Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLife*, 9:e56261, 2020. doi: 10.7554/eLife.56261.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with Co-Scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y. URL https://arxiv.org/abs/2502.18864v2.

B. Hensen, H. Bernien, A. E. Dréau, A. Reiserer, N. Kalb, M. S. Blok, J. Ruitenberg, R. F. L. Vermeulen, R. N. Schouten, C. Abellán, W. Amaya, V. Pruneri, M. W. Mitchell, M. Markham, D. J. Twitchen, D. Elkouss, S. Wehner, T. H. Taminiau, and R. Hanson. Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature*, 526:682–686, 2015. doi: 10.1038/nature15759. URL https://arxiv.org/abs/1508.05949.

A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, 1952. doi: 10.1113/jphysiol.1952.sp004764.

Jan-Åke Larsson. Loopholes in Bell inequality tests of local realism. *Journal of Physics A: Mathematical and Theoretical*, 47(42):424003, 2014. doi: 10.1088/1751-8113/47/42/424003. URL https://arxiv.org/abs/1407.0363.

Carole J. Lee. Critically engaged pragmatism: Scientific norm and social, pragmatist epistemology for AI science evaluation tools. *Social Epistemology*, 2026. doi: 10.1080/02691728.2026.2669578. URL https://arxiv.org/abs/2601.09753v2. Published online June 9, 2026.

Miri Liu and ChengXiang Zhai. An axiomatic benchmark for evaluation of scientific novelty metrics. arXiv:2604.15145v2, 2026. URL https://arxiv.org/abs/2604.15145v2. Preprint, revised August 5, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651:914–919, 2026. doi: 10.1038/s41586-026-10265-5.

Jiaqi W. Ma. Toward an engineering of science: Rebalancing generation and verification in the age of AI. arXiv:2605.10425v1, 2026. URL https://arxiv.org/abs/2605.10425v1. Preprint.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. DiscoveryBench: Towards data-driven discovery with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/0d70af566e69f1dfb687791ecf955e28-Paper-Conference.pdf.

George C. Necula. Proof-carrying code. In *Proceedings of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pages 106–119. ACM, 1997. doi: 10.1145/263699.263712.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 278–287. Morgan Kaufmann, 1999. URL https://ai.stanford.edu/~ang/papers/shaping-icml99.pdf.

Ilkka Niiniluoto. Scientific progress with an institutional aim. *Asian Journal of Philosophy*, 4:58, 2025. doi: 10.1007/s44204-025-00282-y.

OpenAI. On the Navier–Stokes millennium prize problem, September 2026. URL https://openai.com/index/navier-stokes-solution/. Published September 8, 2026; accessed September 8, 2026.

Michael Polanyi. The republic of science: Its political and economic theory. *Minerva*, 1:54–73, 1962. doi: 10.1007/BF01101453. URL https://polanyisociety.org/primary-resources-the-republic-of-science/.

Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence*, 7:836–847, 2025. doi: 10.1038/s42256-025-01055-1. URL https://www.nature.com/articles/s42256-025-01055-1.

Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013. doi: 10.1613/jair.3987. URL https://www.cs.ox.ac.uk/people/shimon.whiteson/pubs/roijersjair13.pdf.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pages 9460–9471, 2022. doi: 10.52202/068431-0687. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d719fee332caa23d5038b8a90e81796-Abstract-Conference.html.

Terence Tao. Mathematics in the age of AI. arXiv:2608.16753v1, 2026. URL https://arxiv.org/abs/2608.16753v1. Preprint; submitted to the Proceedings of the ICM 2026.

Jingqi Tong, Mingzhe Li, Hangcheng Li, Yongzhuo Yang, Yurong Mou, Weijie Ma, Hongji Chen, Xiaoran Liu, Qinyuan Cheng, Ming Zhang, Qiguang Chen, Weifeng Ge, Qipeng Guo, Tianlei Ying, Tianxiang Sun, Yining Zheng, Zhiheng Xi, Xinchi Chen, Jun Zhao, Ning Ding, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. AI can learn scientific taste. arXiv:2603.14473v3, 2026. URL https://arxiv.org/abs/2603.14473v3. Preprint, revised August 19, 2026.

Yunlong Xu and Brent Doiron. Maieusis. Computer software, version 0.1.1. Zenodo, 2026. URL https://github.com/BeibaiDraco/maieusis/tree/v0.1.1.

A Bell correlations, selection, and the scope of a guarantee

The following finite construction illustrates the role of selection in the detection loophole [Larsson, 2014]. Its probabilities are exact model probabilities; the experiment of Hensen et al. [2015] is discussed separately in the main text.

A local model with a large retained-sample score. Let settings $x, y \in \{0, 1\}$ be independent and uniform, and encode the two outcomes as bits $a, b \in \{0, 1\}$. A trial wins the CHSH game when $a \oplus b = xy$, where $\oplus$ denotes addition modulo 2. Independently of the actual settings, the devices share $\lambda = (i, j, r, n)$, where i, j, r are independent uniform bits and n is an independent Bernoulli(e) bit, equal to 1 with probability e. Alice reports $a = r$ only if $x = i$; otherwise she reports no detection. Bob reports $b = r \oplus ij \oplus n$ only if $y = j$; otherwise he reports no detection. Each response uses only a local setting and the shared variable.

Let D denote joint detection. For every setting pair,

$$\Pr(D \mid x, y) = \tfrac{1}{4}, \qquad \Pr(a \oplus b = xy \mid x, y, D) = 1 - e. \tag{2}$$

The first equality follows from $i = x, j = y$; under these equalities the winning condition is precisely $n = 0$. With binary responses encoded as $(-1)^a, (-1)^b$, the retained correlation is $C_{xy} = (-1)^{xy}(1 - 2e)$. Hence

$$S_D = C_{00} + C_{01} + C_{10} - C_{11} = 4 - 8e. \tag{3}$$

For $e = 0$, $S_D = 4$; for $e = 1/5$, $S_D = 12/5$. The retained settings remain uniform, since the retention probability is constant across settings. Nevertheless, selection changes the hidden-variable distribution conditional on those settings. An exactly computed retained-sample score therefore need not exclude a local generating process.

Why the complete-trial bound differs. For a deterministic local response defined on every trial, write $A_0, A_1, B_0, B_1 \in \{-1, 1\}$. Then

$$A_0(B_0 + B_1) + A_1(B_0 - B_1) \in \{-2, 2\}, \tag{4}$$

because one parenthesis is zero and the other is ± 2. Averaging over a setting-independent hidden-variable distribution gives $|S| \leq 2$. Equivalently, the four equations $a_x \oplus b_y = xy$ cannot all hold: their exclusive-or would give $0 = 1$. Uniform settings thus imply a winning probability at most $3/4$. Local randomization can be incorporated in the hidden variable. These arguments do not apply unchanged to the selected responses in Equation 3.

What extending a guarantee establishes. To compare protocols without conflating their data distributions, let P_m^R be the record law induced by mechanism m under research protocol R. Let $\mathcal{H}_0 \subsetneq \mathcal{H}_1$ denote a restricted and an enlarged class of relevant null mechanisms for the same scientific target. Fix the same declared error tolerance $0 < \alpha < 1$ for both protocols. A test ϕ_1 satisfying

$$\sup_{m \in \mathcal{H}_1} P_m^{R_1}(\phi_1 = \text{reject}) \leq \alpha \tag{5}$$

also has that error bound on $\mathcal{H}_0$. If an old test has the same error guarantee only on $\mathcal{H}_0$, and some $m^* \in \mathcal{H}_1 \setminus \mathcal{H}_0$ violates its claimed extension, the additional coverage has a substantive witness. This is a comparison of guarantees, not proof that the old test actually failed on its reported experiment.

An achieved evidential contribution requires more than valid error control and a reported rejection. A test that ignores its record and rejects on an independent coin flip of probability α satisfies the bound on every mechanism, yet its rejection has no discriminating content. Excluding only a never-rejecting test would not exclude this example. One sufficient, limited condition is a scientifically relevant alternative m_{alt} specified before inspecting the record, with $P_{m_{\text{alt}}}^{R_1}(\phi_1 = \text{reject}) = \beta > \alpha$, together with an observed rejection and valid application of the test. For any null mechanism giving rejection positive probability, the likelihood ratio of this coarsened event in favor of that alternative is at least $\beta/\alpha > 1$. This supplies a nonempty evidential contrast under the stated models; it does not establish greater power than the old test, uniform power over all alternatives, or a total value comparison. The comparison presupposes that both protocols supply such scientifically discriminating evidence. Costs, setting independence, and trial definitions remain separate obligations. In particular, α is not the posterior probability that the null is true. The scientific-value principle recognizing the removal of a substantive evidential dependence is an additional normative premise, not a consequence of set inclusion alone.

B Finite experiment comparison and approximation

Let Θ be a finite state space. Experiments E and F are row-stochastic matrices with finite observation sets S and Z: entry $E_{\theta s}$ is the probability of observing s in state θ, and each row sums to 1. A task specifies a prior π, a finite action set $\mathcal{A}$, and a loss $\ell(\theta, a)$. Its optimal Bayes risk is

$$b_{\pi,\ell}(E) = \min_{\delta} \sum_{\theta,s,a} \pi_\theta E_{\theta s} \delta(a \mid s) \ell(\theta, a), \tag{6}$$

where δ ranges over stochastic decision rules. This evaluates an experiment's prospective capability, not the scientific value of every realized outcome.

A finite witness for an entire task class. Suppose a state-independent stochastic matrix K satisfies $F = EK$. For any rule δ_F, the rule $K\delta_F$ after E reproduces the same conditional action distribution at every state. Thus it has the same risk, and

$$F = EK \quad \Longrightarrow \quad b_{\pi,\ell}(E) \le b_{\pi,\ell}(F) \quad \text{for every finite task.} \tag{7}$$

This proves the direction used here. The finite Blackwell theorem supplies the converse when comparison ranges over all finite decision problems; any fixed full-support prior suffices [Blackwell, 1953, Fritz et al., 2023]. For finite rational matrices, stochasticity and $EK = F$ are directly checkable, and finding such a kernel is a linear feasibility problem. Applying the comparison to science requires an adequate representation of the object by the matrices and a reason to value the decision tasks. If acquisition, simulation, or decision costs matter, the comparison must also preserve the allowed decision procedures and their resource bounds.

Incomparability is not an unsuccessful proof search. Let $\Theta = \{00, 01, 10, 11\}$ with uniform prior. Experiments E_x, E_y reveal only the first or second bit; C reveals both. Discarding a bit shows that C simulates each. But E_x cannot simulate E_y: its observation distribution is identical in states 00 and 01, and any state-independent processing preserves that equality, whereas E_y distinguishes them. The reverse fails using states 00 and 10. The corresponding optimal zero–one risks are

Experiment	Guess x	Guess y
E_x	0	$1/2$
E_y	$1/2$	0
C	0	0

There is a proof of non-comparability in the specified relation. This differs from lacking sufficient evidence or computation to decide a comparison. It also does not prove that the scientific importance of learning x and learning y is intrinsically incomparable under every possible value theory.

Approximate witnesses. Use $\mathrm{TV}(p, q) = \frac{1}{2}\sum_z |p_z - q_z|$. If a stochastic K obeys

$$\max_{\theta} \mathrm{TV}((EK)_\theta, F_\theta) \le \varepsilon, \tag{8}$$

then, for every prior and every finite task with $0 \le \ell \le 1$,

$$b_{\pi,\ell}(E) \le b_{\pi,\ell}(F) + \varepsilon. \tag{9}$$

To prove this, fix an optimal rule for F. Its conditional expected loss after observation z is a function $f_\theta(z) \in [0, 1]$. Replacing F_θ by $(EK)_\theta$ changes the expectation of this function by at most their total variation distance. Average over π, use the transferred rule after E, and then minimize over all rules for E. If losses lie in $[a, b]$, the error becomes $(b - a)\varepsilon$; without a uniform loss bound there is no such uniform absolute guarantee.

If the witness instead concerns estimated matrices $\widehat{E}, \widehat{F}$, suppose the actual matrices satisfy uniform rowwise bounds $\mathrm{TV}(E^*_\theta, \widehat{E}_\theta) \le \rho_E$ and $\mathrm{TV}(F^*_\theta, \widehat{F}_\theta) \le \rho_F$. Contraction under K and the triangle inequality give

$$\max_{\theta} \mathrm{TV}((E^*K)_\theta, F^*_\theta) \le \min\{1, \rho_E + \varepsilon + \rho_F\}. \tag{10}$$

These model-error bounds are additional premises, not consequences of a good fit or of checking the kernel. When they hold only with stated confidence, the resulting guarantee is conditional on their joint validity.

C A joint relation omitted by all pairwise marginals

Let P be uniform on the four triples $(a, b, c) \in \{-1, 1\}^3$ satisfying $abc = 1$, and let Q be uniform on all eight triples. Under P, each fixed pair (a, b) has one allowed completion, so its probability is $1/4$. Under Q, it has two completions of probability $1/8$, giving the same pairwise marginal. By symmetry this holds for all pairs; the one-variable marginals are also identical. Nevertheless,

$$P(abc = 1) = 1, \qquad Q(abc = 1) = 1/2. \tag{11}$$

Consequently no function of those marginals alone can recover this joint-event probability for both distributions. A compact representation can preserve it: draw a, b independently and uniformly, then set $c = ab$. This specifies P exactly; the equation alone specifies its support, not its probabilities. The construction illustrates information loss in representing joint constraints. Learned neural posteriors and biological compensation mechanisms require their own validation [Gonçalves et al., 2020]; complexity alone gives no reason to regard a representation as scientifically valuable.

D Correct evaluations and incorrect optimization

With their value functions stipulated in advance, the following elementary examples show how correct evaluations can misdirect reward optimization [Skalse et al., 2022].

A correct lower bound can rank incorrectly. Let $v(A) = 100$, $v(B) = 2$, and certified bounds be $L(A) = 1$, $L(B) = 2$. Both satisfy $L \le v$, but maximizing L selects the lower-value result. If both bounds are available, $L(B) > U(A)$ with $L \le v \le U$ suffices for $v(B) > v(A)$. Alternatively, a uniform approximation $|r(X) - v(X)| \le \varepsilon$ on the relevant outcome domain gives

$$v(B) - v(A) \ge r(B) - r(A) - 2\varepsilon. \tag{12}$$

It follows by bounding the two errors separately. Neither a sound certificate nor an average prediction score supplies this uniform approximation.

Correct initial scores can reward repetition. Let $K_0 = \varnothing$ and $\Phi(\varnothing) = 0$, $\Phi(\{x\}) = 1$, $\Phi(\{y\}) = 9/10$, $\Phi(\{x, y\}) = 19/10$. Action A establishes x; B establishes y. Repeating A only copies the same content, evidence, and method; it is not an independent replication. Initial standalone scores $s(A) = 1$, $s(B) = 9/10$ are correct. Reusing them at both steps gives AA reward 2, exceeding AB reward $19/10$, although the terminal values are 1 and $19/10$, respectively. Initial evaluation correctness therefore does not establish a sequence objective.

For a fixed value theory, finite horizon, and undiscounted increments,

$$r_t = \Phi(K_{t+1}) - \Phi(K_t) \quad \Longrightarrow \quad \sum_{t=0}^{T-1} r_t = \Phi(K_T) - \Phi(K_0). \tag{13}$$

This is cancellation of intermediate terms. It permits a non-additive Φ but neither chooses that function nor guarantees that greedy actions maximize its terminal value. It is not a new policy-invariance result of the kind studied in potential-based reward shaping [Ng et al., 1999]. Discounting, costs, an insufficient state representation, or changes to the value theory require separate treatment. In particular, increasing the score of an unchanged scientific state by revising its evaluation must not be silently credited as new research achievement.