Toward a Science of AI Evaluation: What the Disciplines That Study Science Can Teach Us

Christopher Kelly

Abstract

A growing consensus holds that AI evaluation should become a science, yet the field has largely pursued this goal without engaging the disciplines that study how sciences form, mature, and self-correct. We draw on three overlapping traditions, the science of science, the meta-science and open science movement, and the economics of innovation, to address three questions critical to AI evaluation's maturation: How does the AI evaluation ecosystem work as a knowledge system, and how can we measure it? How can we make AI evaluation more reliable, transparent, cumulative, and self-correcting? What structural conditions produce or suppress evaluation quality, and how do we design institutions accordingly? We diagnose AI evaluation as a field in early mobilization whose dominant benchmark paradigm shows degenerative characteristics, apply frameworks from field formation theory, the philosophy of scientific research programs, and culture change models to specify where the field stands, and propose a staged maturation agenda organized by Nosek's culture change pyramid. We argue that building the capacity to study AI evaluation scientifically is both a marker of and a prerequisite for the field's maturation, and that doing so produces returns well beyond tracking progress, including accelerating methodological innovation, strengthening the evidence base for AI governance, and making the field's knowledge cumulative rather than ad hoc.

1 Introduction

There is a growing consensus that AI evaluation needs to become a science. Weidinger et al. (2025) call for an "evaluation science." The ai-evaluation.org programme is building what may become the first MSc in AI Evaluation. The EvalEval Coalition frames its mission around making evaluation "scientifically grounded." The December 2025 issue of the AI Evaluation Digest observed that "one of the key things this year has brought is collective realisation that some sort of science of evaluation is needed and is a sort of discipline in itself." Hardt (2025) has been developing what he calls "The Emerging Science of Machine Learning Benchmarks." Apollo Research (2024) issued a "call to arms" for a science of evals. These calls span AI safety, NLP, and governance communities.

But what does "becoming a science" actually mean for AI evaluation, and how does a community pursue that goal deliberately? The field is largely attempting to answer these questions without engaging the disciplines that have spent decades studying exactly how sciences form, how they mature, how they self-correct, and how they fail to self-correct. This situation parallels the AI evaluation field's recent but belated engagement with measurement theory (Salaudeen et al., 2025; Wallach et al., 2025). Just as AI evaluation once operated without engaging psychometrics, the field now pursues scientific maturation without engaging the science of science.

The study of how science works has developed over several decades into a rich interdisciplinary enterprise, spanning what is variously called science of science (Fortunato et al., 2018), meta-science or meta-research (Ioannidis, 2005), and economics of innovation (Jones, 2009; Azoulay et al., 2011). These traditions overlap substantially in personnel, methods, and findings. But together they address three questions that are critical to AI evaluation's maturation:

1. How does the AI evaluation ecosystem work as a knowledge system, and how can we measure it?

2. How can we make AI evaluation more reliable, transparent, cumulative, and self-correcting?

3. What structural conditions produce or suppress evaluation quality, and how do we de-

sign institutions accordingly?

A productive recursion sits at the heart of this argument. Making AI evaluation into a science requires the ability to study the field scientifically. Building that capacity requires specific infrastructure, norms, and institutions that these traditions have spent decades developing. And once built, that capacity produces returns well beyond tracking progress: it accelerates methodological innovation, strengthens governance evidence, and makes the field's knowledge cumulative rather than ad hoc. This process follows what Chang (2004) calls epistemic iteration, the bootstrapping mechanism by which measurement sciences have historically developed, using imperfect tools to build slightly better ones in successive cycles. AI evaluation need not wait for perfect infrastructure to begin; the iteration itself is the path to maturity.

Moreover, as AI increasingly handles execution-level research tasks, the selection layer, identifying what to study, what to fund, which evaluation approaches to pursue, becomes the binding constraint on field progress. The traditions drawn upon here provide empirical foundations for these selection decisions, moving research direction-setting from intuition toward evidence.

AI evaluation today occupies what Hambrick and Chen (2008) would call the differentiation and early mobilization stage of field formation. Apollo Research (2024) reach an equivalent conclusion through a practitioner lens, characterizing the field as nascent (the first of three stages: nascent, maturation phase, mature field), defined by the absence of agreed-upon best practices and the sense that evaluation remains more art than science. Both diagnoses point to the same condition. Its dominant paradigm, static benchmarks as proxies for capability, shows characteristics that Lakatos (1970) would identify as degenerative: increasingly ad hoc adjustments that protect core assumptions without generating novel predictions. Meanwhile, what we term evaluation debt, the compounding risk that accumulates when systems are deployed with inadequate measurement infrastructure, grows with each deployment decision based on evidence that would not survive systematic quality assessment. The infrastructure layer of Nosek's (2019) culture change pyramid is partially built; the norms, incentives, and policy layers are largely absent. This is a predictable state for a field at this stage of development, and the literature tells us what to do about it.

We do not attempt a comprehensive review of any one tradition. We identify the highest-value connections between these traditions and AI evaluation, and propose a concrete path forward that the community, including the workshop venue hosting this paper, can act on.

2 What "Becoming a Science" Means: Diagnostic Criteria

"Becoming a science" is not a binary threshold. The field needs concrete criteria for assessing progress; otherwise the aspiration remains vague. We draw from three complementary frameworks, each illuminating a different dimension of scientific maturity.

2.1 Field Formation Stages

Hambrick and Chen (2008) model the emergence of a new academic field as a social movement consisting of three stages: differentiation (claiming important problems that existing fields cannot solve), mobilization (assembling social infrastructure such as journals, conferences, departments, and funding), and legitimacy (achieving external validation through PhD programs, tenure tracks, and a shared methodological canon).[1]

AI evaluation has clearly differentiated itself. No one disputes that evaluating AI systems constitutes a distinct and important set of problems. Mobilization is underway: the EvalEval Coalition, the AI Evaluator Forum, the ai-evaluation.org programme, workshops at NeurIPS and ACL, and government-led international networks all represent mobilization activities. Legitimacy, however, remains nascent. There are no dedicated journals for AI evaluation methodology, no PhD programs specifically in evaluation science, no tenure-track positions in the area, and no shared methodological canon.

2.2 Progressive Versus Degenerating Research Programs

Lakatos (1970) distinguishes between progressive and degenerating research programs. A progressive

[1]The sociology of science offers more empirically grounded alternatives: Frickel and Gross (2005) develop a theory of scientific/intellectual movements identifying micropolitical, mesolevel, and macropolitical conditions for field formation; Whitley (1984) classifies fields by mutual dependence and task uncertainty, predicting fragmented, ad hoc structure for fields like AI evaluation.

program generates novel predictions that are subsequently confirmed. A degenerating program becomes increasingly ad hoc, with modifications that serve only to defend its core assumptions against anomalies.

The current benchmark paradigm assumes that static benchmarks on fixed datasets measure AI capability. Its *positive heuristic* directs practitioners to extend coverage along three axes: more comprehensive datasets, harder evaluation splits, and greater task diversity—treating deployment-behavior prediction as outside the program's scope rather than as an open empirical question. Anomalies continue to accumulate: data contamination, score saturation, construct validity failures, Goodhart's law dynamics, and strategic evaluation gaming. The responses have been largely ad hoc: harder test sets, contamination detection patches, formatting adjustments, and new data splits. These modifications protect the core assumption without predicting novel phenomena about AI systems. This pattern fits Lakatos's criteria for a degenerating program. Schaeffer et al. (2023) provide a paradigmatic illustration: apparent emergent capabilities in LLMs, among the benchmark paradigm's most celebrated claimed findings, turn out to be artifacts of metric choice rather than genuine discontinuities in model behavior. The anomaly was not predicted by the benchmark paradigm; it was discovered by examining the paradigm's own measurement assumptions.

The field has also made more structural responses that go beyond patch-style fixes: dynamic and adversarial benchmarks, agentic task-completion evaluations, and the widespread adoption of LLM-as-judge approaches in which AI systems evaluate AI outputs. Each addresses a genuine anomaly (saturation, rigid item format, the cost of human evaluation) but none constitutes a principled reconception of evaluation methodology. Dynamic benchmarks inherit construct validity and contamination problems in new forms. Agentic evaluations introduce non-determinism across runs that existing statistical frameworks do not address: outcomes vary with stochastic sampling and environmental variation, making single-run results unreliable, yet the field has not standardized how many runs are required or how to aggregate across them (Song et al., 2025). LLM-as-judge approaches, now widely adopted for preference and quality assessment (**?**), introduce their own systematic validity failures: positional bias, verbosity bias, and self-preference bias are well-documented across judge models (Zheng et al., 2023; **?**; **?**), and the field lacks principled criteria for when such judgments are valid or what construct they measure. The core assumption, that measurement instruments proxy genuine capability, persists across each of these shifts, even as the instrument grows more sophisticated. This is the Lakatosian pattern: the protective belt expands, the core assumption holds. A genuinely *transitional* program would revise the core assumption and generate novel predictions about which evaluation designs actually measure capability. The structural responses do not do this: each addresses a surface anomaly without reconceiving what the instrument is measuring.

2.3 The Culture Change Pyramid

Nosek (2019) proposes a staged model for how scientific culture changes, moving through five levels: Infrastructure ("make it possible"), Ease of use ("make it easy"), Norms ("make it normative"), Incentives ("make it rewarding"), and Policy ("make it required"). Each level builds on those below.

Applying this framework to AI evaluation reveals a specific pattern of partial progress. At the infrastructure level, reporting tools like EEE and execution frameworks like lm-eval and Inspect exist, but evidence quality metadata, pre-registration registries, and replication infrastructure remain absent. Ease of use remains low for evidence quality: EEE converters reduce reporting friction, but no tools exist for documenting validity or assessing evidence quality in a structured way. Emerging norms around transparency and contamination reporting are visible, but community standards for pre-registration, registered reports, statistical power analysis, or systematic replication do not yet exist. Incentives for evaluation quality are largely absent: no career paths reward evaluation methodology research, no funding streams support replication, and SOTA-chasing continues to be rewarded over rigor. Policy interventions remain premature. The EU AI Act references evaluation without specifying quality standards; the NeurIPS reproducibility checklist represents the closest existing conference requirement.

2.4 Synthesis

Taken together, these three frameworks place AI evaluation in a specific, diagnosable state. Hambrick and Chen say the field is mobilizing but not yet legitimate. Lakatos says its dominant paradigm

shows degenerative signs, though progressive alternatives are emerging. Nosek says the infrastructure is partially built, but the norms, incentives, and policies that would cement scientific practice are largely absent.

This is a precise enough diagnosis to prescribe interventions. The following sections identify what the interdisciplinary study of science tells us about those interventions.

3 Three Questions for Making AI Evaluation a Science

The study of how science works, spanning science of science, meta-science, open science, and economics of innovation, has over several decades developed into a substantial interdisciplinary enterprise. We organize our treatment around three questions this enterprise helps answer for AI evaluation, noting which traditions and methods are most relevant to each.

The AI evaluation community has not yet substantively engaged this enterprise, particularly its economics-of-innovation and science-of-science dimensions. Existing proposals for making AI evaluation more scientific draw from AI safety (Apollo Research, 2024), psychometrics (Burnell et al., 2023; Salaudeen et al., 2025), NLP methodology (Bowman and Dahl, 2021; Kiela et al., 2021), and critical AI/STS scholarship (Raji et al., 2021). Each contributes valuable insights — Raji et al. engage institutional critique, and Bowman and Dahl address community incentive structures — but none draws systematically on the economics of innovation, the science of science, or the sociology of scientific fields to address these dimensions at a field level.

3.1 How Does the AI Evaluation Ecosystem Work as a Knowledge System, and How Can We Measure It?

A science that cannot study itself cannot improve systematically. At present, discussions about what is wrong with AI evaluation rest on intuition, anecdote, and individual case studies. The interdisciplinary study of science transforms these into empirical questions with quantitative answers. Building this self-study capacity is both a marker of scientific maturity (mature fields study themselves) and a prerequisite for tracking progress toward maturity (one cannot know whether reform efforts are working without measuring their effects).

A direct precedent. Teplitskiy's program of "building a science of scientific evaluation" (Teplitskiy et al., 2018, 2022) provides the most comprehensive empirical investigation of how evaluation systems function in high-stakes intellectual contexts, and has never been cited in the AI evaluation literature. Two findings apply directly: professional connections between authors and reviewers bias validity judgments even when evaluators are instructed to assess only scientific merit; and institutional design (editorial structure, specifically) shapes what gets recognized more than individual evaluator quality does.

Key methods and their analogs in AI evaluation. Science-of-science methods map directly onto open empirical questions in AI evaluation; Table 1 summarizes the main approaches.

Building this self-study capacity produces returns well beyond tracking maturation. It allows the field to identify which evaluation methods actually work, to spot evaluation blind spots before they become governance failures, and to predict benchmark saturation before it becomes a crisis.

Infrastructure. The Every Eval Ever (EEE) database, as a structured repository of evaluation results across models and benchmarks, is the first infrastructure project capable of supporting this research program at scale. Combined with citation data from OpenAlex or Semantic Scholar, it enables benchmark lifecycle analysis, evaluator ecosystem mapping, and method diffusion tracking.

3.2 How Can We Make AI Evaluation More Reliable, Transparent, Cumulative, and Self-Correcting?

Reliable, transparent, cumulative, and self-correcting: these are the properties of a mature science. The open science and meta-science movement has spent roughly two decades developing and empirically testing interventions that move fields toward these properties. AI evaluation can draw on this tested playbook rather than reinventing it.

The reinvention problem. The AI evaluation field has repeatedly rediscovered concepts that exist in mature forms elsewhere. The term "uplift" in AI safety recapitulates treatment effect estimation from the potential outcomes framework in causal inference (Rubin, 1974). Benchmark "validity" concerns recapitulate construct validity from

Method	Source	AI evaluation question
Citation network analysis	Standard bibliometrics	How do evaluation methods spread? Which communities remain insular?
Three-parameter citation model	Wang et al. (2013)	Which evaluation approaches will have lasting influence?
Self-referentiality index	Frank et al. (2019)	Is AI evaluation becoming more insular over time?
CD disruption index	Park et al. (2023)	Are new benchmarks creating methodological advances or consolidating existing paradigms?
Attention concentration	Hao et al. (2026)	Is the field evaluating what matters, or what is easy to evaluate?
Team size and disruption	Wu et al. (2019)	Do small academic teams produce more novel evaluation approaches than large industry ones?

Table 1: Science-of-science methods and their analogs in AI evaluation.

psychometrics (Cronbach and Meehl, 1955). Evaluation quality discussions recapitulate GRADE and PROBAST from evidence-based medicine. The desire for evaluation registries recapitulates ClinicalTrials.gov. Each reinvention costs time, produces less mature versions of existing tools, and, critically, compounds: imprecise AI evaluation language enters the training corpus, AI systems trained on that literature internalize the imprecision, and AI-assisted research accelerates the cycle of publication without improving definitional clarity. This terminology degradation loop is a novel, self-reinforcing process that makes interdisciplinary engagement urgent.

What the reform playbook says works, with evidence. Prospective trial registration, mandated by ICMJE in 2005, measurably increased clinical trial registration rates. Soderberg et al. (2021) found, in a blinded evaluation by 353 researchers, that Registered Reports outperformed standard papers on all 19 quality criteria while remaining indistinguishable in novelty and creativity, directly countering the objection that pre-registration stifles innovation. AI evaluation has no pre-registration registry. Pineau's NeurIPS reproducibility checklist increased code sharing from under 50% to approximately 75%, showing that low-cost requirements change practice. Card et al. (2020) demonstrate that NLP experiments are systematically underpowered and call for 80% power as the minimum threshold, the standard convention since Cohen (1962), a baseline AI evaluation has not adopted. A companion paper submitted to this workshop develops one additional gap Card et al. do not address: evaluations adequately powered for average effects are routinely underpowered for subgroup and heterogeneity analyses on the same data. The TOP Guidelines (adopted by over 5,000 organizations), the Coalition for Evidence-Based Policy's tiered evidence framework, GRADE, and the i4R's AI Replication Games supply further models.

Several tools transfer with modest modification: pre-registration models, tiered evidence systems, and the TOP transparency framework. Others require adaptation for AI's unique features, addressed in Section 4. A companion paper submitted to this workshop operationalizes the most directly transferable tools as concrete schema extensions to EEE.

A nuance. Munger (2024) has argued that metascience privileges process-based evaluation (did they pre-register?) over outcome-based evaluation (did the finding prove important?). AI evaluation should build genuine epistemic infrastructure rather than compliance theater. We acknowledge the risk of compliance theater while arguing that evidence-informed institutional design remains preferable to design from scratch.

3.3 What Structural Conditions Produce or Suppress Evaluation Quality, and How Do We Design Institutions Accordingly?

Becoming a science requires more than good methods and transparent norms. It requires institutional and incentive structures that sustain quality over time. Everyone in AI evaluation is aware that benchmarks have validity issues, that contamination is pervasive, and that most evaluations go unreplicated. If awareness alone were sufficient, these problems would already be solved. They persist because they are structural features of the evaluation ecosystem's incentive landscape, not failures of individual judgment.

The structural diagnosis. Drawing primarily from economics of innovation and institutional economics, several mechanisms explain the current situation. Rigorous evaluation is a public good: its benefits accrue broadly, but the costs are borne privately. Classic underproduction follows. Informa-

tion cascades (Bikhchandani et al., 1992) explain benchmark lock-in: when a benchmark becomes popular, subsequent researchers adopt it not because they independently assessed its quality but because non-adoption is costly. Stein (2025) finds that competition degrades quality on the most important problems. In AI evaluation, competitive pressure leads to rushed, lower-quality safety evaluations precisely where careful evaluation matters most. This application warrants a caveat: Stein's mechanism assumes quality is observable ex-post, a condition that may not hold for safety evaluation, and developers may face incentives to pass evaluations rather than maximize scores, partially modifying the degradation dynamic.

We propose **evaluation debt** as a concept for the AI evaluation community. Analogous to technical debt, evaluation debt accumulates when AI systems are deployed with inadequate measurement infrastructure. It compounds through a specific mechanism: deployment decisions based on weak evidence create precedents; those precedents become baselines; future evaluations are judged against those baselines rather than against absolute evidence quality standards.

? demonstrate that what a field chooses to measure determines what gets developed: in clinical trials, the use of surrogate endpoints shaped which drugs were pursued. In AI, benchmark design determines what developers optimize for. This is the primary causal pathway through which evaluation shapes the AI ecosystem. Getting evaluation right is about shaping what AI becomes, not merely knowing what it can do.

The reinvention problem described in Section 3.2 also has structural roots. Smaldino and O'Connor (2022) showed that contact between scientific communities spreads superior methodology, but requires reduced self-preferential bias and increased competence to evaluate foreign methods. AI research is increasingly self-referential (Frank et al., 2019), and jargon is nearly always negatively correlated with interdisciplinary impact (Lucy et al., 2023).

The role of institutions. Institutions play specific, well-studied roles at each stage of field development. Funders seed infrastructure: Arnold Ventures' allocation of over $60 million to metascience built the Center for Open Science, the Reproducibility Project, and the transparency infrastructure the open science movement now runs on. The HHMI model, which funds people rather than projects while tolerating early failure, produces approximately twice as many top-1% cited papers as NSF-style project grants (Azoulay et al., 2011). Focused Research Organizations (Marblestone and Rodriques, 2022) offer a further model suited to infrastructure too coordinated for academia and too public-goods-oriented for industry. Professional associations build norms: the EvalEval Coalition, AI Evaluator Forum, and ai-evaluation.org are performing this role, and Haas's (1992) epistemic community framework specifies what they need to succeed. Publication venues enforce standards: the NeurIPS Datasets and Benchmarks Track and Pineau's reproducibility checklist demonstrate that requirements change practice; no dedicated journal for AI evaluation methodology yet exists. Government sets policy: the UK Metascience Unit (£49M) and the FDA's formalization of regulatory science (the recognition that evaluating novel products requires developing the science of evaluation itself) are the closest institutional precedents.

Institutional existence proof. The meta-science trajectory from Ioannidis (2005) through the Center for Open Science, the Reproducibility Project, the TOP Guidelines, and the Metascience Alliance (now over 25 organizations) demonstrates that deliberate scientific field-building works, scales, and attracts both philanthropic and government investment. The UK Government's Metascience Unit (£49M) is the most recent sign that this model is being taken seriously at a policy level. The Institute for Progress's Economics of Ideas course (Spring 2026) already includes AI evaluation in its curriculum. The intellectual bridge is being built from the other side.

4 What Is Genuinely New About AI Evaluation

Not everything transfers from existing disciplines. Evaluation evidence drives deployment decisions for systems that could cause societal-scale harm; the International AI Safety Report describes an "evidence dilemma" in which policymakers must act on weak evidence or wait indefinitely for stronger evidence that may never arrive. Several features of AI evaluation have no direct analog in prior scientific fields and require genuinely new development.

Speed of change. A common objection holds that rigorous evaluation infrastructure will always

arrive too late to matter for the models currently being deployed. This misidentifies what the 15 to 20 year meta-science maturation estimate describes: the time needed to build an institutional stack from zero, not to implement individual practices once they are known. The interventions are already identified, several carry effect estimates, and Nosek's pyramid supplies a sequencing theory rather than just a list of things to try. Partial infrastructure already exists in EEE, lm-eval, and the EvalEval Coalition, and the AI research community's capacity to move quickly, aided in part by the AI tools it studies, may accelerate norm diffusion in ways prior fields could not. Different evaluations also have different stakes: a benchmark tracking academic NLP progress can proceed at lower evidence tiers, while a safety evaluation informing a frontier model deployment warrants prospective registration and independent verification. These modes coexist within a tiered evidence framework, and rigorous evaluation of a model no longer at SOTA is not wasted effort; it contributes methodological knowledge that accumulates across generations. The argument is not that maturation will be fast, but that deliberate, evidence-informed field-building can make it meaningfully faster than organic emergence, while carrying real risks: urgency generates pressure for compliance theater, and premature institutionalization can entrench the wrong norms as readily as the right ones.

Reflexivity. AI is simultaneously the object of evaluation and increasingly the tool for conducting it, creating feedback loops with no parallel in traditional science-of-science work. LLM-as-judge approaches make this structural rather than incidental, with systematic positional, verbosity, and self-preference biases for which no correction standards yet exist (Zheng et al., 2023; **?**; **?**). Agentic evaluations introduce a related challenge: nondeterminism across runs, where multiple independent runs and pass@k aggregation exist but are not normalized (Song et al., 2025). Both require new methodological development that existing science-of-science tools do not provide.

Evaluation gaming. The subject of AI evaluation can, in some cases, strategically detect and defeat the evaluation instrument (Apollo Research, 2024; Hubinger et al., 2024). No other field faces subjects capable of gaming measurement in this way, creating a fundamental tension between transparency (needed for verification and self-correction) and opacity (needed for evaluation integrity).

Commercial concentration. Evaluation, development, and deployment are often vertically integrated within the same organizations, making evaluator independence structurally harder to achieve than in any prior science. Closed-weight models add a distinct replication challenge: models can change silently between measurements, API access can be revoked, and mechanistic claims cannot be verified without weight access. This is not fatal; open-weight models sidestep the constraint, behavioral evaluation remains tractable on the FDA regulatory science model, and structured transparency documentation can make evidentiary gaps legible as a precondition for closing them through regulatory mandate.

These features do not invalidate the contributions of Section 3, but they define the frontier of genuinely new work. Appendix A provides a structured summary.

5 A Maturation Agenda

The agenda below is organized by the levels of Nosek's (2019) culture change pyramid, with priority actions at each.

Infrastructure ("make it possible"). Four priorities stand out. First, extend existing reporting infrastructure (EEE) to capture evidence quality, including design soundness, transparency, and evidence strength, as structured, queryable metadata. A companion paper submitted to this workshop proposes a concrete schema for this extension. Second, create a pre-registration registry for AI evaluations equivalent to the AEA RCT Registry or ClinicalTrials.gov. Third, establish a replication fund and infrastructure equivalent to i4R, funded to systematically reproduce evaluation results. Fourth, initiate validity infrastructure for the reflexive case: calibration standards for LLM-as-judge specifying the constructs and conditions under which LLM judgments are credible proxies for human evaluation, and statistical norms for multi-run agentic evaluation analogous to power analysis conventions in human studies (Song et al., 2025).

Ease of use ("make it easy"). Evidence quality documentation, pre-registration, and transparency reporting must integrate with tools evaluators already use, including lm-eval, Inspect, and EEE

converters. COS's strategy of building tools that integrate with existing workflows provides the model.

Norms ("make it normative"). The AI evaluation equivalent would involve the EvalEval Coalition adopting evidence quality standards for its own projects, leading evaluation organizations publicly committing to transparency and pre-registration, and workshops requiring reproducibility documentation for submitted evaluations.

This workshop, EvalEval at ACL 2026, is itself performing the kind of norm-setting and community-building that the field-formation literature identifies as essential. It functions as what Galison (1997) calls a trading zone: a space where practitioners from distinct communities (ML researchers, social scientists, policymakers, ethicists) can coordinate joint activity even when they hold different methodological commitments.

Incentives ("make it rewarding"). No major funding stream currently supports replication of AI evaluation findings or research on evaluation quality as a topic in its own right. Coefficient Giving's meta-research funding model, J-PAL's Science for Progress Initiative, and the UK Metascience Unit's grants provide templates. An "evaluation FRO" (Focused Research Organization), a five-year, tightly coordinated team building shared evaluation infrastructure, could fill a gap that is too coordinated for academia and too public-goods-oriented for industry.

Policy ("make it required"). Policy interventions are premature without the lower levels in place, but the trajectory should be noted. Conference requirements could adapt TOP Guidelines for AI evaluation venues. Funder mandates could require pre-registration and data sharing for funded evaluation research. Regulatory standards could connect evaluation evidence quality to regulatory recognition under the EU AI Act and Frontier Safety Frameworks.

The EvalEval Coalition, the AI Evaluator Forum, the ai-evaluation.org programme, and several research programs represent real progress at the infrastructure and early mobilization stages; the binding constraints lie primarily at the norms, incentives, and institutional levels.

6 Discussion

Hardt (2025) is the most directly relevant existing work. He approaches the science of benchmarks from ML and statistics. Our paper is complementary: we add the economics and meta-science perspective on incentive structures, institutional design, field-building dynamics, and the political economy of evaluation that ML and statistical analysis cannot address. Weidinger et al. (2025) argue why evaluation science is needed; this paper proposes how to build it. Salaudeen et al. (2025) and Wallach et al. (2025) apply psychometric validity to benchmarks, one important tool among many. Hernandez-Orallo (2017) correctly argues that much of what AI evaluation needs already exists in measurement theory; our paper extends this point to the entire enterprise of studying how science works.

More broadly, the maturation of AI evaluation requires a shift in the field's epistemic culture, from benchmark-centric, speed-driven knowledge production toward the kind of cumulative, self-correcting, evidence-quality-conscious practice that characterizes mature measurement sciences. The epistemic iteration described in Section 1 is the mechanism for this synthesis: each round of building infrastructure, studying the field, and applying lessons produces a progressively more coherent and rigorous epistemic culture for AI evaluation.

Limitations

Each tradition we draw from has its own extensive literature; we have identified the highest-value connections rather than provided comprehensive reviews. Adapting methods to AI evaluation's unique features (Section 4) is non-trivial; speed, reflexivity, and evaluation gaming create genuine challenges that existing tools do not fully address. The meta-science movement took roughly 15 to 20 years to develop from Ioannidis (2005) to a globally funded ecosystem. AI evaluation may not have that much time. Compressing the maturation process introduces risks of premature institutionalization or top-down imposition of norms before community buy-in is achieved. Hwang's (2023) argument about the "political fragility" of meta-science reforms is a real concern. Furthermore, building evaluation infrastructure is a public good, but the costs of compliance fall on individual researchers and organizations.

References

Apollo Research. 2024. We need a science of evals. Apollo Research Blog.

Pierre Azoulay, Joshua S. Graff Zivin, and Gustavo Manso. 2011. Incentives and creativity: Evidence from the academic life sciences. *RAND Journal of Economics*, 42(3):527–554.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026.

Samuel R. Bowman and George E. Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of NAACL 2021*, pages 4843–4855.

Ryan Burnell, Wout Schellaert, John Burden, and 1 others. 2023. Rethink reporting of evaluation results in AI. In *Science*, volume 380, pages 136–138.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274.

Hasok Chang. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.

Jacob Cohen. 1962. The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65(3):145–153.

Lee J. Cronbach and Paul E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.

Santo Fortunato, Carl T. Bergstrom, Katy Börner, and 1 others. 2018. Science of science. *Science*, 359(6379):eaao0185.

Morgan R. Frank, Dashun Wang, Manuel Cebrian, and Iyad Rahwan. 2019. The evolution of citation graphs in artificial intelligence research. *Nature Machine Intelligence*, 1:79–85.

Scott Frickel and Neil Gross. 2005. A general theory of scientific/intellectual movements. *American Sociological Review*, 70(2):204–232.

Peter Galison. 1997. *Image and Logic: A Material Culture of Microphysics*. University of Chicago Press.

Peter M. Haas. 1992. Introduction: Epistemic communities and international policy coordination. *International Organization*, 46(1):1–35.

Donald C. Hambrick and Ming-Jer Chen. 2008. New academic fields as admittance-seeking social movements: The case of strategic management. *Academy of Management Review*, 33(1):32–54.

Yiling Hao and 1 others. 2026. Artificial intelligence and the contraction of scientific understanding. *Nature*.

Moritz Hardt. 2025. The emerging science of machine learning benchmarks. Online manuscript.

Evan Hubinger, Carson Denison, Jesse Mu, and 1 others. 2024. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*.

Tim Hwang. 2023. The political fragility of meta-science. Macroscience (Substack), July 20, 2023.

John P. A. Ioannidis. 2005. Why most published research findings are false. *PLoS Medicine*, 2(8):e124.

Benjamin F. Jones. 2009. The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *Review of Economic Studies*, 76(1):283–317.

Douwe Kiela, Max Bartolo, Yixin Nie, and 1 others. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of NAACL 2021*, pages 4110–4124.

Imre Lakatos. 1970. Falsification and the methodology of scientific research programmes. In Imre Lakatos and Alan Musgrave, editors, *Criticism and the Growth of Knowledge*. Cambridge University Press.

Li Lucy and 1 others. 2023. Words as gatekeepers: Measuring discipline-specific terms and meanings in scholarly publications. *Proceedings of EMNLP 2023*.

Adam Marblestone and Sam Rodriques. 2022. Focused research organizations to accelerate science, technology, and medicine. ArXiv:2201.11022.

Satyam Mukherjee and 1 others. 2017. The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology. *Science Advances*, 3(4):e1601315.

Kevin Munger. 2024. The incoherence of science reform. Never Met a Science (Substack), July 26, 2024.

Brian A. Nosek. 2019. Strategy for culture change. Center for Open Science.

Michael Park, Erin Leahey, and Russell J. Funk. 2023. Papers and patents are becoming less disruptive over time. *Nature*, 613:138–144.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. In *NeurIPS 2021 Datasets and Benchmarks Track*.

Donald B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Hammed Salaudeen, Siddharth Vasan, Sagnik Ray Choudhury, and Zeerak Talat. 2025. Measuring what satisfies: On the evaluation of LLMs with psychometric validation. *arXiv preprint arXiv:2501.09674*.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Roberta Sinatra and 1 others. 2016. Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239.

Paul E. Smaldino and Cailin O'Connor. 2022. Interdisciplinarity can aid the spread of better methods between scientific communities. *Collective Intelligence*, 1(1).

Courtney K. Soderberg and 1 others. 2021. Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour*, 5(8):990–997.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2025. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. In *Proceedings of NAACL 2025*. ArXiv:2407.10457.

Christoph Stein. 2025. Race to the bottom: Competition and quality. *Quarterly Journal of Economics*, 140(1):299–352.

Misha Teplitskiy and 1 others. 2018. The sociology of scientific validity: How professional networks shape judgement in peer review. *Research Policy*, 47(9):1825–1841.

Misha Teplitskiy and 1 others. 2022. Is novel research worth doing? evidence from peer review at 49 journals. *Proceedings of the National Academy of Sciences*, 119(47):e2118046119.

Hanna Wallach, Yanda Shen, and Dan Wallach. 2025. Toward rigorous evaluations of AI systems. *arXiv preprint arXiv:2501.10868*.

Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying long-term scientific impact. *Science*, 342(6154):127–132.

Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. 2025. Toward an evaluation science for generative AI systems. *arXiv preprint arXiv:2503.05336*. First two authors contributed equally.

Richard Whitley. 1984. *The Intellectual and Social Organization of the Sciences*. Oxford University Press.

Lingfei Wu, Dashun Wang, and James A. Evans. 2019. Large teams develop and small teams disrupt science and technology. *Nature*, 566:378–382.

Lianmin Zheng and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36. ArXiv:2306.05685.

A Novel Features of AI Evaluation

Feature	Core challenge
Speed of change	Model generations arrive in months; citation-based methods built for multi-year observation windows need adaptation; replication must happen on faster timescales
Reflexivity	AI is simultaneously the object and increasingly the instrument of evaluation, creating feedback loops between evaluation and development with no traditional parallel
LLM-as-judge biases	AI systems routinely evaluate AI outputs at scale, with systematic positional, verbosity, and self-preference biases that lack established correction standards (Zheng et al., 2023; **?**; **?**)
Agentic non-determinism	Agentic evaluations produce different outcomes across runs; multiple independent runs and pass@k aggregation exist but are not standardized; single-run results remain the norm (Song et al., 2025)
Evaluation gaming	Subjects can, in some cases, detect and defeat the evaluation instrument (Apollo Research, 2024; Hubinger et al., 2024); no other field faces subjects capable of gaming measurement in this way
Commercial concentration	Evaluation, development, and deployment are often vertically integrated; evaluator independence is structurally harder to achieve than in any prior science; Merton's norm of disinterestedness faces an extreme challenge
Closed-weight replication	Models can change silently between measurements; API access can be revoked; mechanistic claims cannot be verified without weight access; open-weight alternatives, behavioral evaluation on the FDA regulatory science model, and structured transparency documentation provide partial paths forward

Table 2: Novel features of AI evaluation with no close analog in prior scientific fields.

B The Returns to Building This

Building the self-study capacity and reform infrastructure described above produces returns at multiple levels.

Accelerating methodological innovation. The field currently has no systematic way to know which evaluation approaches work, which are being adopted, or which are reaching saturation. The self-study capacity described in Section 3.1 would allow the field to allocate attention more efficiently.

More cost-effective R&D. The reinvention problem outlined in Section 3.2 is expensive. Every hour spent rediscovering construct validity or reinventing pre-registration is an hour not spent on the genuinely novel challenges described in Section 4.

Better evidence for governance. Governance decisions about AI systems depend on evaluation evidence. That evidence is currently difficult to synthesize: there are no common effect sizes, no power conventions, no evidence tiers, and no aggregation methods. Building evidence synthesis infrastructure makes governance evidence-ready.

Cumulative rather than ad hoc knowledge. A mature science builds cumulatively. AI evaluation currently produces largely non-cumulative findings; each benchmark result is an isolated data point rather than a contribution to a growing body of evidence.

Developing research taste empirically. The science of science provides tools for understanding what makes research impactful. The finding that the most influential work combines recent and vintage knowledge (Mukherjee et al., 2017) and the finding that individual research ability is a persistent, measurable trait (Sinatra et al., 2016) provide a basis for understanding productive direction-setting. A forecasting platform for AI evaluation results, adapted from Vivalt and DellaVigna's Social Science Prediction Platform, could further develop this capacity.

Monitoring reform efforts. Without self-study capacity, the field cannot know whether its interventions are actually improving evaluation quality.