

---

# AI Review Won’t Deliver a Credibility Revolution Until We Treat It Like the Wicked Policy Problem It Is

---

**Jessica Hullman**

Northwestern University  
jhullman@northwestern.edu

## Abstract

AI-implemented checks for robustness, preregistration, claim scope, and other intended proxies for scientific credibility can extend human reviewers’ capabilities. However, treating metascientific heuristics as necessary and sufficient signals for filtering out bad science ignores the contested nature of such heuristics as indicators of good science. AI review is best approached as a wicked policy design problem: there is no definitive formulation or notion of solution correctness, and any deployed solution shapes downstream outcomes by changing authors’ incentives. To promote accountability and community deliberation, we present a decision theoretic standard that proposers of AI review tools can use to make their assumptions about the signal-to-decision pipeline explicit, including specifying the latent epistemic state a signal is meant to inform, the review decision it supports, the assumed utility function, and a plausible data-generating model under which the signal improves that decision. Drawing on recent metascientific modeling of a corpus of OSC replication studies, we demonstrate using the standard to quantify the value of several proposed proxies—exact replication status and statistical significance—for several review decisions, illustrating how the same signal can vary substantially in how informative it is depending on how the goal of review is conceived. We discuss how commonly proposed signals of research quality are insufficiently justified as general indicators of scientific value, and why relying on humans as the final arbiters of review decisions does not remove the risks of overreliance on proxies.

## 1 Introduction

Until now, human peer review has played the role of a quality control mechanism for scientific production, albeit an imperfect one subject to bias, poor incentives, and limited expertise [Mann et al., 2025, Shah, 2022]. LLMs increase scientific production [Kusumegi et al., 2025], threatening to stress the peer review system to the point of collapse. A “review death spiral” ensues when human reviewers, overwhelmed by the number of papers, become noisier, further incentivizing authors to submit low quality, mass produced work [Bergstrom and Gross, 2026, Liu and Tan, 2026].

Using LLMs to evaluate science at scale provides a promising way forward. While human reviewers lack sufficient time or incentives to fully interrogate a submission, AI reviewers can conduct extensive checks for rigor. This could reduce reviewer lottery variance [Liu and Tan, 2026], which is often substantial [Beygelzimer et al., 2023, Ward et al., 2026]. Science appears to be on the verge of a credibility revolution. Proponents argue that the verification of scientific quality can be handled at scale by machines, reserving human judgment for which science is most significant to society.

However, automated science evaluation is a high stakes reward design problem, with the properties of a wicked policy problem: it offers no definitive formulation, no notion of solution correctness nor easy way to test solutions, and any chosen solution changes the problem space [Rittel and Webber, 1973].

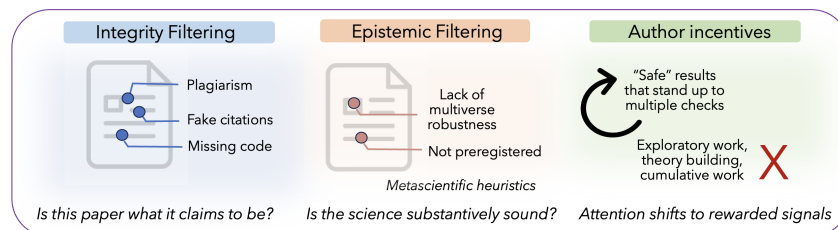


Figure 1: AI review research frequently blurs integrity filtering (is the paper what it claims to be?) and epistemic filtering (is the science good?) by relying on contestable metascientific heuristics.

Machine learning researchers are building the infrastructure that will structure which science receives attention in the future, but what checks should AI reviewers conduct?

One source of inspiration has been the reform discussions that have been unfolding more broadly in the sciences for over a decade. Spurred by a so-called replication crisis affecting many empirical fields, reformers have suggested guiding principles for producing a more robust research record. These include promoting reproducibility and replication as core tenets [Nosek et al., 2015, Collaboration, 2015], as well as preregistration of analysis plans [Nosek et al., 2018], sensitivity analyses [Steege et al., 2016, Simonsohn et al., 2020], and aligning claims with the evidence at hand [Simons et al., 2017, Yarkoni, 2022]. AI-based evaluation proposes to enact such reform principles by automating scoring according to rubrics for rigor and surfacing these for human reviewers, or filtering research based on precommitment to particular methods or experiments.

But taking proposed reform measures at face value as real indicators of quality institutionalizes heuristics that lack rigorous grounding as robust markers of scientific quality. Ultimately, scientific progress, and notions of scientific objectivity and theory choice, are imprecise, potentially conflicting, and dependent on socially situated standards and plural epistemic values [Douglas, 2009, Kuhn, 2012, Longino, 1990, 2002]. AI review has the potential to address known weaknesses of human peer review, such as limited attention and inconsistency, but we should expect any AI review system that uses machine-detectable proxies to allocate scientific attention to add bias and create an optimization target. Even with human oversight, AI-generated representations of papers come at the expense of correlated bias when reviewers trust those representations over their information [Liu et al., 2026b]. The question is what trade-offs to expect, and how they align with the community’s values.

We argue that **proposers of AI review tools should be accountable for specifying what review decision a proxy informs and under what assumptions it improves that decision**. To benefit from the useful functions of AI review while avoiding canonizing poorly understood heuristics, the community must move toward explicit specification of goals and trade-offs. We propose statistical decision theory as a framework for formalizing these goals. This path does not resolve the contestability of any proxy signal, but treats AI review as inherently subject to value-laden judgments.

Below, we first summarize proxies that recent research in AI review proposes to detect. We show how these proposals tend to blur a distinction between AI review implementing basic integrity checks, and epistemic filtering, where contestable criteria thought to be associated with rigor are used to reject or prioritize work as more scientifically valuable. We then present a decision-theoretic standard for formal specification of signal-to-decision pipelines to eliminate ambiguity about what is at stake. This standard asks proposers to state the proxy signal, the latent epistemic state it is meant to inform on, the review decision it is meant to support, including the preferred utility function(s), and a plausible generating model under which the proxy signal improves decision quality. Using a recent model fit to empirical study results [van Zwet et al., 2026], we demonstrate how the value of a proxy is highly conditional on the assumed decision problem, motivating the approach. We then discuss where the linking assumptions behind three proposed proxies borrowed from the broader scientific reform movement remain tenuous. Finally, we explain why keeping humans in the loop as final arbiters of decisions does not reduce the need for transparent justification.

## 2 What Are AI Reviewers Screening For?

Challenges faced by research AI reviewers like biases in human ground truth reviews [Checco et al., 2021, Tyser et al., 2024, Idahl and Ahmadi, 2025] and dependence of subjective ratings of AI reviews on style and author self-interest [Liang et al., 2023, D’Arcy et al., 2024] lead researchers to seek “first principles” that capture more fundamental aspects of rigor to base AI review on. A common assumption in research on AI review is that scientific assessment can be reduced to filtering on observable signals detected in paper text and supplemental materials. We characterize the kinds of signals AI reviewers are being asked to detect by surveying 40 recent papers at proposing AI tools for evaluating scientific work, identified via AI-assisted literature search and citation tracing. We retained 23 papers from these that explicitly connected detectable signals with dimensions of research quality. Table 1 displays proposed proxies for scientific quality that AI is argued or shown to be capable of detecting across these papers.

Authors make frequent reference to the broader scientific reform literature, implying that the epistemic significance of the proxies they implement via LLM review has already been established. Replication testing is referred to as “essential to academic and research integrity” [Nguyen et al., 2026]; robustness to data and model perturbations as “a core criterion for trustworthy conclusions” [Bertran et al., 2026]; and prespecification of analysis pipelines as necessary for preventing selective disclosure of prompts that favor specific hypotheses [Kosch and Feger, 2025, Bertran et al., 2026].

| Category                       | Proxy for scientific quality                                                                                                                                            | Definition                                                                                                                                                                                                                                            |
|--------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Verifying reported results     | <b>Computational reproducibility</b> [Xu and Yang, 2026, Bai et al., 2026, Xu et al., 2026, Starace et al., 2025, Hu et al., 2025, Shah et al., 2026, Wu et al., 2026]  | The ability to verify reported estimates such as regression coefficients by obtaining data, reconstructing the computational environment, and executing analysis protocols using the authors’ code.                                                   |
|                                | <b>Methods-only reproducibility</b> [Kohler et al., 2026, Starace et al., 2025]                                                                                         | The ability to verify reported estimates by relying only on the methods description.                                                                                                                                                                  |
|                                | <b>Specification completeness</b> [Liu and Tan, 2026]                                                                                                                   | Whether full details like hyperparameters, random seeds, and configuration are reported.                                                                                                                                                              |
|                                | <b>Claim traceability</b> [Bai et al., 2026, Dmonte et al., 2024, Xu et al., 2026, Javaji et al., 2025, Xi et al., 2025, Wei et al., 2025]                              | Whether a paper’s stated conclusions can be traced back to specific evidence, in the form of figures, tables, statistical analyses, or to the underlying data and code that generated them.                                                           |
| Stability and generalizability | <b>Replicability</b> [Nguyen et al., 2026, Bai et al., 2026]                                                                                                            | The ability to produce nearly similar results on new data from the same process.                                                                                                                                                                      |
|                                | <b>Robustness of effects to perturbing analysis choices</b> [Cummins, 2025, Bertran et al., 2026, Fishman and Sekeres, 2026, Cui and Alexander, 2026, Bai et al., 2026] | The ability to produce similar results when the analysis specification is changed slightly, such as by varying which variables are controlled for in a model, what stopping rule is used in model fitting, or various other routine analytic choices. |
|                                | <b>Robustness to stochastic perturbation</b> [Cui and Alexander, 2026]                                                                                                  | The ability to produce similar results among multiple repetitions of the same specification when using LLMs to analyze data.                                                                                                                          |
| Inferential validity           | <b>Pre-specification of analysis plans</b> [Cummins, 2025, Kosch and Feger, 2025, Fishman and Sekeres, 2026]                                                            | The declaration of an analysis specification (including the structure of a multiverse) prior to obtaining data and conducting the analyses.                                                                                                           |
|                                | <b>Statistical significance</b> [Bai et al., 2026]                                                                                                                      | Whether a paper’s results are statistically significant.                                                                                                                                                                                              |
| Novelty and Coverage           | <b>Novelty and contribution significance</b> [Biswas et al., 2026, Zhang et al., 2026]                                                                                  | Whether the research is novel and competitive in light of prior work.                                                                                                                                                                                 |
|                                | <b>Coverage of related literature</b> [Wei et al., 2025, Zhang et al., 2026]                                                                                            | Whether important related work is missed.                                                                                                                                                                                                             |

Table 1: Examples of proxies that recent work proposes to automate or assess with generative AI.

## 2.1 From Detection to Decision

We observe a blurring of a distinction between using AI for *integrity filtering* to using AI for *epistemic filtering* in emerging work on AI review. Integrity filtering requires basic, necessary but not sufficient signals that a work is what it claims to be, like the ability to reproduce the results reported in the paper from the data and code to rule out obvious errors, and the absence of plagiarism or fake citations. Epistemic filtering, on the other hand, seeks to assess the broader substantive value of the research contribution, for example, by checking whether the result can be replicated under perturbations of the data or modeling process, judging the size and statistical significance of effects, and noting whether analyses were specified in advance.

Authors go beyond promoting integrity filters to arguing that markers of epistemic value can be detected and used to allocate attention [Biswas et al., 2026, Liu and Tan, 2026, Liu et al., 2026a, Xi et al., 2025, Zhang and Abernethy, 2026]. For example, recent work distinguishes computational reproducibility and claim traceability as “level 1” requirements for submissions. However, additional layers of assessment (such as methodological rigor, falsifiability quality, and argument coherence) are also described as “objective” criteria with “epistemic soundness” that can be used to decide which papers get human review [Liu et al., 2026a]. Others propose making acceptance decisions dependent on whether authors have committed to particular approaches like preregistration or demonstrations of robustness [Fishman and Sekeres, 2026, Kosch and Feger, 2025].

The problem is that despite decades of metascience, we have yet to find a necessary and sufficient set of signals for credible science. Signals associated with good science can be in opposition and shift over time: Copernicus’ model of the solar system was simpler but initially inferior to the geocentric tradition in consistency and accuracy [Kuhn, 2012]. Epistemic filtering may sometimes aid scientific progress, but in other cases shift attention away from potentially more powerful yet less mature ideas.

We should expect whatever proxies we implement to structure attention in automated review to reshape science. This is the lesson of Goodhart’s Law [Goodhart, 1984]. Principal-agent problems describe how when a principal cares about something hard to observe—like scientific quality or long-term epistemic value—but must rely on observable proxy signals to judge agent outputs, they should expect agents to shift more effort toward improving exclusively on those proxies [Sappington, 1991]. Effort substitution is rational from the agent’s perspective, but reduces effort toward other valuable but unmeasured tasks [Holmstrom and Milgrom, 1991]. Emphasizing signals like preregistration, replicability, and robustness to perturbations shifts attention from harder-to-institutionalize but equally critical components like developing stronger theory, improving external validity, or identifying the right estimand for measuring a claimed capability. We should treat proxies proposed to structure review as the contestable policies they are, and develop standards for expressing the assumptions that are required for them to inform about a target latent state in the context of a review decision problem. These specifications should precede other forms of evaluation, such as controlled studies comparing AI-assisted review outcomes to unaided human review [Liu and Tan, 2026, Baumann et al., 2026], as they establish the estimands and tests of assumptions that empirical study should target.

## 3 Formalizing the Goals and Trade-offs of AI Review

There is no context-independent, non-conflicting set of detectable markers of good science. In light of this, we argue that AI review should be approached as a wicked policy problem [Rittel and Webber, 1973], where in the absence of ways to distinguish correct from incorrect solutions, the best we can do is clearly specify our commitments to particular framings of the problem. We propose decision theory as a natural formalism for specifying assumptions behind proxies implemented in AI review tools. The goal of such specification is not to make proxies non-gameable, nor to definitively validate that a given AI review tools will improve scientific outcomes, because neither are possible. Rather, the specifications become the inputs to an inherently value-driven deliberation process around what kinds of research the community wants to prioritize and what downstream incentives they are willing to tolerate. In this sense, specification is a mechanism for accountability. Those who wish to deploy new AI review tools are responsible for motivating them relative to a well-defined problem, and identifying what conditions must hold for the information they provide to be useful. Venues can enforce this by requiring signal-to-decision specifications before piloting AI-generated signals in review, and communities can contest the specifications they are given, such as by proposing alternative formulations of the problem or by building datasets that pair proxy signals with observed outcomes.

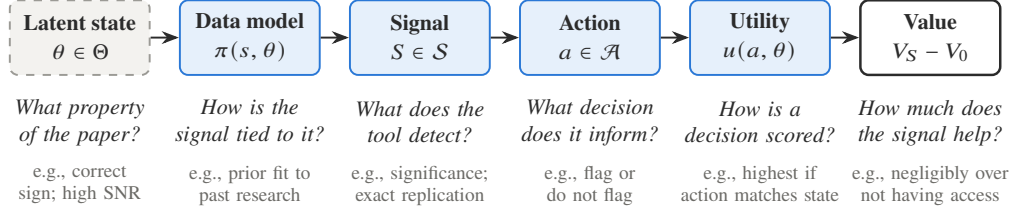


Figure 2: A decision-theoretic standard for justifying an AI review signal specifies the latent state the signal should inform, a data-generating model linking signal and state, the review action, and a utility function. These determine the value of the signal,  $V_S - V_0$ .

At the implementation level, reviewer training should accompany deployments of AI-assisted review flags. Reviewers should be prompted, at a minimum, with examples of failure cases where a proxy casts doubt on rigorous research and where it fails to flag questionable research.

### 3.1 Decision-theoretic specification of signal-to-decision pipelines

In proposing AI review tools, researchers implicitly claim that surfacing certain signals will help reviewers make better decisions about paper quality. We turn to decision theory to make such claims explicit (Figure 2) by identifying what the signal is, what latent property of the paper it is meant to inform, what decision the reviewer faces, how decisions are scored, and how the signal and latent property are assumed to be related. This approach enables computing the value of a proposed signal, which we measure as the improvement in the best case decision that can be made with access to that signal over not having it. This quantity, with the decision problem specification, can be assessed by the community and used to motivate downstream empirical evaluations against human review.

**Specifying the decision problem** Let  $S \in \mathcal{S}$  denote the *signal* surfaced by an AI review system. We use  $S$  to denote the random variable and  $s$  to denote a particular realization of the signal, where  $S$  could be a vector of multiple proxy outputs.

Let  $a \in \mathcal{A}$  denote the review action or *decision* we intend  $S$  to inform. For example, an initial review pass might return an action in  $\mathcal{A} = \{\text{do not flag, flag}\}$

The quality of the review decision depends on the true state of the paper, which is uncertain at the time of review. Let  $\theta \in \Theta$  denote the latent epistemic *state*, capturing the credibility of the scientific contribution along some dimension. The intention is that the proxies inform on this state. But how should proposers of AI review tools conceive of the latent state, given that the scientific quality of papers is unobservable, domain-specific, and often contested by experts? Rather than attempting to define a universal latent variable for quality, we advise using concrete, local definitions that can be mapped to empirical data. Some examples include downstream citations, whether a reported estimate exceeds a signal-to-noise threshold, whether a reported ranking would hold under repeated evaluation, whether the reported sign is correct, or whether evidence supports flagging a claim as fragile.

Given a signal space and state space, we define a *data-generating model*, which is the joint distribution over the latent states and signals on which we will evaluate decisions,  $\pi \in \Delta(\mathcal{S} \times \Theta)$ . This distribution assigns to each realization  $(s, \theta) \in \mathcal{S} \times \Theta$  a probability, denoted  $\pi(s, \theta)$ .  $\pi$  gives rise to a probability distribution over the latent state, which we denote as the *prior distribution*  $p \in \Delta(\Theta)$ :

$$p(\theta) = \sum_{s \in \mathcal{S}} \pi(s, \theta).$$

Review quality is dictated by a *utility function*  $U : \mathcal{A} \times \Theta \rightarrow \mathbb{R}$ , which assigns a real valued score  $u(a, \theta)$  to each combination of action and realized state.<sup>1</sup>

<sup>1</sup> If the action is instead a probabilistic assessment of paper quality, then  $U$  may be chosen as a strictly proper scoring rule over reports in  $\Delta(\Theta)$ .

**Optimizing the decision.** Once we have defined the decision problem and data generating model, we can characterize best-case use of the information, enabling us to quantify the ex-ante value of the signal to the decision problem.

The best possible decision-maker is represented by the Bayesian rational agent, informed by prior knowledge  $p(\theta)$ . Upon viewing a paper, this agent updates their beliefs to posterior beliefs  $q_s(\theta)$ :

$$q_s(\theta) = \pi(\theta | s) = \frac{\pi(s, \theta)}{\sum_{\theta' \in \Theta} \pi(s, \theta')}.$$

When  $S$  is a vector, the decision-maker conditions on all of its components to compute the posterior, with no separate weighting scheme required. They then choose the score maximizing action  $a^*(s)$ :

$$a^*(s) = \arg \max_{a \in \mathcal{A}} \mathbb{E}_{\theta \sim q_s} [u(a, \theta)]$$

**Estimating the value of the signal.** Following Wu et al. [2023], define the rational *benchmark* as the ex-ante expected score of the Bayesian agent who observes the signal and then chooses the score-maximizing action for each signal realization:

$$V_S = \mathbb{E}_S \left[ \max_{a \in \mathcal{A}} \mathbb{E}_{\theta \sim q_S} [u(a, \theta)] \right]. \quad (1)$$

Similarly, the *baseline* is the expected score of decisions made using only the prior:

$$V_0 = \max_{a \in \mathcal{A}} \mathbb{E}_{\theta \sim p} [u(a, \theta)]. \quad (2)$$

The value of signal  $S$  to the decision problem is the difference between the benchmark and baseline,  $V_S - V_0$ .

By normalizing  $U$  to  $[0, 1]$ , authors can report  $V_S - V_0$  directly as the expected improvement in decision performance (utility) from access to the signal.

These quantities can be estimated from data where both the signal and a measurement of the latent state are observed. When the signal is low dimensional, such as a binary replication outcome or significance indicator,  $\pi$  can be estimated directly from empirical frequencies, and  $V_S$  and  $V_0$  computed by plugging in the empirical posterior for each signal realization. When the signal is continuous or high-dimensional, such as an AI-generated text review, a predictive model can instead be used to predict the state from the signal to provide the posterior distribution, subject to hold-out testing and calibration checks, since calibration error bounds how far the estimated value can fall from that of the optimal decision rule [Guo et al., 2025].

Contrast the value of the signal above with the treatment effect of an AI review tool as estimated by running a randomized controlled experiment comparing the tool to unaided human reviewers. While deployment claims should ultimately be grounded by empirical evaluation against relevant controls, the specification standard above enables reasoning about the value of signals, both individually and jointly, prior to running costly human experiments. If a proxy cannot be shown to provide value in theory, then it is difficult to motivate more elaborate empirical studies with humans.

### 3.2 Applying the framework: Example using empirical model from van Zwet et al. [2026]

The above formalization makes clear that signals do not carry value in an unconditional sense; we can only reason about the value of their information with respect to a specified decision problem, including a particular formulation of the latent state, and a data-generating model that links the proxies and latent state. Thus, in proposing particular proxies for AI review, authors should think carefully about what a plausible generating process (or class thereof) looks like, and show that the signal has non-negligible value under that process.

Showing that a signal is valuable under an assumed data-generating model cannot resolve the contestability of that signal as a filter. However, proposed signal-to-decision pipelines can be made more credible by drawing on empirical metascientific analyses to motivate their plausibility. Human-subject studies of review with AI assistance can also inform on possible misinterpretations of

proxies and how they may complement or override human expertise (using, e.g., tools for measuring complementarities [Guo et al., 2025]).

We demonstrate using a data-generating model informed by empirical research to quantify the value of two proxies—an exact replication result and whether the result was statistically significant at a specified level—for informing on several latent epistemic states. One is whether the latent signal-to-noise ratio is sufficiently high, and the other is whether the reported sign of the effect is correct. We use van Zwet et al. [2026]’s model of how signal-to-noise ratios are related to statistical significance and replication success. The authors use maximum likelihood to estimate a latent distribution of signal-to-noise ratios from samples of absolute  $z$ -values across multiple disciplines. We use their results for the OSC corpus of 100 psychology replication studies as an illustrative prior, and simulate signed original and replication  $z$ -values under their assumed model. In this model, significance and whether an exact replication was successful are noisy functions of the underlying signal-to-noise ratio. We consider the value of these signals for review decision-making, considering two possible definitions of the latent epistemic state on which decision quality depends.<sup>2</sup>

**Action, general latent state, utility function.** The reviewer faces a binary action,  $\mathcal{A} = \{\text{do not flag, flag}\}$ . The intended use of the signal is to decide whether to flag a claim for evidential fragility. Let  $\theta \in \{0, 1\}$  denote the latent epistemic state, where  $\theta = 1$  means the claim satisfies the epistemic property of interest and  $\theta = 0$  means it does not. Below we define two different candidate epistemic states of whether the study has sufficiently high signal-to-noise and whether the observed effect has the correct sign.

We use a simple symmetric utility function which awards 1 if the action matches the true state (i.e.,  $a = \text{do not flag}$  and  $\theta = 1$  or  $a = \text{flag}$  and  $\theta = 0$ ) and 0 otherwise.

**Specific latent state, signals, and data-generating model.** Let  $\Lambda$  denote the (unobservable) signal-to-noise ratio of a reported estimate. Following the signal-to-noise model assumed by [van Zwet et al., 2026],  $\Lambda \sim H$ , where  $H$  is a corpus-level distribution of signal-to-noise ratios. We use the OSC-fitted distribution  $H$  as a prior over signal-to-noise ratios, where conditional on a draw  $\Lambda \sim H$ , we simulate a reported study statistic  $Z$  and an exact replication statistic  $Z_{\text{rep}}$ , assuming  $z$ -values are equivalent to signal-to-noise ratios plus independent standard normal errors, with an exact replication produced the same way:

$$Z = \Lambda + \epsilon, \quad Z_{\text{rep}} = \Lambda + \epsilon', \quad \epsilon, \epsilon' \sim \mathcal{N}(0, 1), \quad Z \perp\!\!\!\perp Z_{\text{rep}} \mid \Lambda.$$

We consider two latent epistemic states under these assumptions. The first is whether the study has a **sufficiently high signal-to-noise ratio** for the intended decision. The second is whether the observed effect has the **correct sign**:

$$\theta_{\text{snr}} = \mathbb{I}(|\Lambda| \geq \lambda_0), \quad \theta_{\text{sign}} = \mathbb{I}(\Lambda Z > 0),$$

where  $\lambda_0 > 0$  denotes a threshold chosen for the decision context. We use  $\lambda_0 = 2.8$ , which is the signal-to-noise ratio corresponding to 80% power at the 5% two-sided level under the assumed data-generating model.

We also consider two observable signals. The first is a **statistical significance** indicator calculated from the original study  $z$ -value, and the second is **exact replication success**, which indicates a statistically-significant exact replication in the same direction as the original result:

$$S_{\text{sig}} = \mathbb{I}(|Z| \geq 1.96), \quad S_{\text{rep}} = \mathbb{I}(ZZ_{\text{rep}} > 0, |Z_{\text{rep}}| \geq 1.96).$$

Figure 3 depicts the value of all signals for the two latent states. Neither signal improves on the prior for  $\theta_{\text{sign}}$ : the Bayes-optimal reviewer makes the correct decision with probability 0.783 with or without them. For  $\theta_{\text{snr}}$ , statistical significance raises the probability of a correct decision from 0.794 to 0.843, exact replication success raises it to 0.862, and observing both raises it to 0.925.

This example shows that the same signal can have very different value depending on the focal latent state, with neither proxy offering value over the prior for identifying results that get the direction correct, while both provide useful information for decisions based on the true signal-to-noise ratio.

<sup>2</sup>Code is available at [https://github.com/jhullman/AI\\_metascience\\_spec](https://github.com/jhullman/AI_metascience_spec).

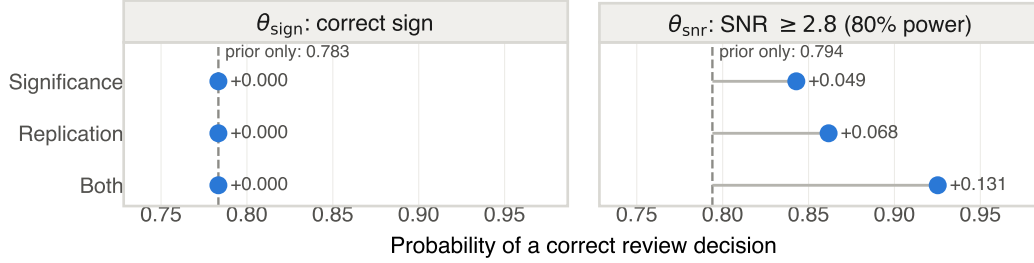


Figure 3: Expected utility of the Bayes-optimal decision rule under each signal (quantifying the ex-ante best case value of the signal), for two latent epistemic states,  $\theta_{\text{sign}}$  (left) and  $\theta_{\text{snr}}$  (right). Each dot shows the expected utility given the signal listed on the left. The dashed line is the baseline  $V_0$ , representing the utility of the best constant action without any signal. Under the symmetric 0/1 utility function, utility is the probability of making the correct review decision.

### 3.3 Modeling adaptation

Specifying a decision problem and plausible data-generating model is a first step toward building understanding of the epistemic commitments encoded in AI review tools. In addition to empirical study of how AI review tools change review outcomes, further theoretical and empirical exploration of how deploying a signal changes the data-generating model under which it was validated is warranted. The above approach to evaluating the value-of-information treats the joint distribution of proxies and latent states as fixed, but once a proxy is used to allocate attention, the authors’ best responses alter that distribution. This is the formal counterpart of Goodhart’s Law, which the ML community has explored through strategic classification and performative prediction, including characterizing when a decision rule induces agents to invest in improving the underlying quality rather than the observable signal [Kleinberg and Raghavan, 2020], or when a decision rule remains optimal under the distribution shift it induces [Perdomo et al., 2020]. Closer to the question of scientific review, recent modeling of editorial screening illustrates such an approach [Fishman and Sekeres, 2026].

## 4 Proxy problems

We use three conceptually distinct proxies of scientific rigor that AI review research has targeted to show why components of the specification in Section 3 matter.

### 4.1 Promise: Screening for multiverse robustness selects credible results

Studies of multiple analysts facing the same high level question and dataset routinely show high variability in conclusions [Silberzahn and Uhlmann, 2015, Silberzahn et al., 2018, Breznau et al., 2022]. The “garden of forking paths” has become a common trope for explaining how underspecified research questions create analytical flexibility that, when combined with data-dependent decision making, can undermine inferential guarantees [Gelman and Loken, 2013].

Recent work frames AI as amplifying these concerns. AI versions of many-analyst studies can produce substantial variation [Cui and Alexander, 2026], even if their results are less dispersed than those of human analysts [Huang et al., 2026, Grundl, 2026]. Small changes in prompting can steer the model toward different results [Bertran et al., 2026]. This variability presents a loophole for selective reporting, making it easier to search a large space of specifications and report only those that align with one’s preferred interpretation [Kosch and Feger, 2025, Fishman and Sekeres, 2026, Bertran et al., 2026, Cummins, 2025].

Consequently, authors propose that “multiverse analysis” must become the new abstraction for scientific communication [Bertran et al., 2026, Liu et al., 2026a, Fishman and Sekeres, 2026]. Instead of presenting the result from a single analysis specification, a multiverse multiplexes over all reasonable decisions that could have been made in analyzing the data and presents the resulting combinatorial set [Steege et al., 2016]. Readers can examine which analytical choices most impact estimates. Going a step further, consistency in the results is framed as enabling readers to assess

the credibility of the claim, with the idea that “the greater the variation in results, the greater the uncertainty in answering the research question and, accordingly, the lower the credibility of any individual research finding” [Auspurg and Brüderl, 2021].

**Problematic assumptions: Multiverse composition and unwarranted frequency interpretation**

Using result consistency as a signal of credibility treats the multiverse as an inferential tool. Yet as methodologists have stressed, a multiverse is most defensible as a tool for reflection and critique, exposing how analytic choices can matter [Rohrer et al., 2026]. It is harder to justify as an inferential summary of “how credible” a result is [Rohrer et al., 2026, Auspurg and Brüderl, 2021]. Variation across specifications is only interpretable relative to justified inclusion choices and a coherent estimand [Del Giudice and Gangestad, 2021]. Which specifications belong is a matter of domain expertise and statistical judgment, and experts within the same area regularly disagree.

Additionally, it’s tempting to assume that a claim that holds for a minority of specifications is not credible, and one that holds for a majority is. However, the paths in a multiverse are not random samples from a well-defined population. Consequently the frequency with which a result appears cannot be interpreted as the probability that it is correct. These issues suggest problems with the blanket linking assumption that good science will always show multiverse robustness.

## **4.2 Promise: Screening for replicability selects credible results**

If a paper’s results cannot be reproduced using the author-provided data and code, the paper fails a basic legibility standard that published results can be verified and built upon. But replicability is harder to judge. Imagine one paper has a low success rate across repeated replications on new datasets sampled from the same process, while another has a high success rate. Can we conclude that the paper with the high rate is more likely to represent a “true” result? Contrary to popular reform narratives in which replicability is framed as a requisite property of valid science, drawing conclusions about scientific credibility from observed replication rates faces several challenges.

**Problematic assumptions: Arbitrary definition, identifiability, insufficiency for true results** By definition, we do not expect perfect reproduction of all original point estimates when replicating. We must therefore decide on a criterion for judging replication success. Common proposals like “significance in the same direction as the original result” are difficult to rigorously motivate; we must grapple with questions of statistical power and hypothesized effect size [Gelman et al., 2026]. More fundamentally, the dichotomy between “true” and “false” effects that replication conversations assume is itself a crude simplifying assumption.

Further challenges arise given that the true replication rate of a study—the relative frequency of reproduced results upon repetitions of the associated probability experiment [Buzbas et al., 2023]—is not directly observable. It is a parameter of the idealized version of the experiment and true data generating model. Even in the best case, we’d need many replications are needed to infer it with confidence. Even modest heterogeneity among experiment-specific replicability rates induces an irreducible variance floor on the estimated mean replicability rate, so additional replications do not eliminate uncertainty [Devezer and Buzbas, 2026].

Finally, even if we could estimate an experiment’s long-run replication rate precisely, it would still not be a straightforward signal of whether the reported claim is true. A real effect or genuine methodological improvement can have a low replicability rate conditional on a particular evaluation design, because the design is confounded, underpowered, or misaligned with the claim. In ML a method may genuinely improve performance in some way, but the benchmark may be too small, saturated, or sensitive to implementation details to reproduce the improvement reliably.

Conversely, false effects can replicate at high rates. Devezer et al. [2021] demonstrates through simulation how when measurement error variability is much larger than sampling error variability, a “false” effect replicates with high frequency. Thus, as a signal replicability suffers from arbitrary definition and an underidentified estimand, with its value depending on which latent state is targeted, as Section 3.2 illustrates. Institutionalizing replicability risks scaling a “replication paradox” [Meng, 2020] where a focus on replication leads the community to overlook larger challenges to inference, such as unwarranted causal inferences, underspecified estimands, and weak links between theories of capability and evidence.

### 4.3 Promise: Requiring preregistration selects for credible results

Because AI reduces the cost of many tests, some authors argue that preregistration—where authors submit a timestamped analysis prior to collecting data [Nosek et al., 2018]—should be required [Cummins, 2025, Kosch and Feger, 2025, Fishman and Sekeres, 2026]. At first glance, this appears uncontroversial. Preregistration helps readers to distinguish analyses that follow from prediction from those that follow from postdiction, making it easier to interpret reported error rates and claims.

**Problematic assumptions: Oversimplifying the problem, penalizing exploration** However, the epistemic significance of preregistration is shaky. Some reform rhetoric treats “double dipping” or “data peeking” as if using data to decide analyses and report results necessarily invalidates inference [Devezer et al., 2021]. But inference is not invalidated simply because the data are used more than once. What matters is whether the reference distribution used to interpret the result properly accounts for the procedure that produced it, including data-dependent choices [Devezer et al., 2021, Lindley, 2000]. Taken too literally as policy, preregistration encourages the view that post-hoc adaptation is presumptively suspect. Many valid procedures routinely reuse data [Devezer et al., 2021], and alternative approaches to correcting for conditioning exist, such as multiplicity adjustment, reusable holdouts [Dwork et al., 2015], hierarchical modeling [Gelman and Hill, 2007], multiverse reporting [Steege et al., 2016], or simply treating the result as exploratory or descriptive.

Embracing preregistration as policy threatens exploratory and descriptive aspects of science by equating making a scientific contribution with error-controlled confirmatory claims. This is a poor fit for much ML research, where progress often comes through iterative benchmark construction and error analysis. Decades of progress suggest that exploration can also be disciplined, albeit also imperfectly, by using shared datasets, code release, leaderboards, ablations, and reproducibility checklists [Donoho, 2024, Pineau et al., 2021]. Thus, the utility function implied by heavily rewarding favors questions that are easy to formalize ex-ante over theory-building and measurement and diagnostic refinement.

## 5 Alternative views

A commonly proposed safeguard against unanticipated consequences of AI review is to have human reviewers decide how to weigh AI-generated flags [Liang et al., 2023, Goldberg et al., 2024, Xi et al., 2025, Wei et al., 2025, Liu et al., 2026a, Tan and Liu, 2026, Xu et al., 2026, Zhang and Abernethy, 2026]. We argue that human oversight does not relieve authors of the responsibility to rigorously motivate the evidential meaning of any proxies that shape downstream decisions, because the flags still encode controversial assumptions about what variation means. If the authors cannot define appropriate use of the proxies in review, why should we expect the average human reviewer to know how to use them? Empirical studies of AI-assisted decisions offer considerable evidence of overreliance on surfaced information (e.g., [Chen et al., 2023, Rathi et al., 2025, Romeo and Conti, 2026]). Even when reviewers do not blindly accept AI recommendations, systems that surface, rank, or summarize issues reframe papers by these salience cues, reshaping scientific production by rewarding work that is legible to the system.

Baumann et al. [2026] argue for a moratorium on AI-generated peer reviews until the practice meets basic requirements established through rigorous empirical evaluation. They present experimental evidence on ICLR 2026 reviews showing that AI-generated reviews collapse review diversity relative to humans and can be gamed by rewriting the paper using AI. They propose non-gameability and review diversity as necessary conditions any automated review system must satisfy. We agree on the need for a more rigorous science of AI review, but argue that empirical evaluation alone is insufficient. Baumann et al. [2026] assert that AI review can be used once it is shown to be non-gameable; we point out that *no* AI review system that relies on proxies for scientific quality will be non-gameable in the strong sense, because scientific quality is not reducible to a small set of stable proxies. While alternatives like submission quotas or fees are worth exploring, our position is that AI review cannot be ruled out as obviously better or worse than human review, and community deliberation about its value is best supported by explicit characterization of what the proxy measures, what it misses, how it can be optimized, and what tradeoffs follow from using it.

## Acknowledgments and Disclosure of Funding

We thank Kirill Skobelev for comments.

## References

- Katrin Auspurg and Josef Brüderl. Has the credibility of the social sciences been credibly destroyed? reanalyzing the "many analysts, one data set" project. *Socius: Sociological Research for a Dynamic World*, 7, January 2021. doi: 10.1177/23780231211024421. URL <http://dx.doi.org/10.1177/23780231211024421>.
- Xiaoyan Bai, Alexander Baumgartner, Haojia Sun, Ari Holtzman, and Chenhao Tan. The story is not the science: Execution-grounded evaluation of mechanistic interpretability research, 2026. URL <https://arxiv.org/abs/2602.18458>.
- Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. Stop automating peer review without rigorous evaluation. In *Forty-third International Conference on Machine Learning*, 2026. Spotlight.
- Carl T Bergstrom and Kevin Gross. Screening, sorting, and the feedback cycles that imperil peer review. *PLoS Biology*, 24(2):e3003650, 2026.
- Martin Bertran, Riccardo Fogliato, and Zhiwei Steven Wu. Many ai analysts, one dataset: Navigating the agentic data science multiverse, 2026. URL <https://arxiv.org/abs/2602.18710>.
- Alina Beygelzimer, Yann N Dauphin, Percy Liang, and Jennifer Wortman Vaughan. Has the machine learning review process become more arbitrary as the field has grown? the neurips 2021 consistency experiment. *arXiv preprint arXiv:2306.03262*, 2023.
- Joydeep Biswas, Sheila Schoepp, Gautham Vasan, Anthony Pipari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, Kiri L. Wagstaff, Matthew E. Taylor, and Odest Chadwicke Jenkins. Ai-assisted peer review at scale: The aaai-26 ai review pilot, 2026. URL <https://arxiv.org/abs/2604.13940>.
- Nate Breznau, Eike Mark Rinke, Alexander Wuttke, et al. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, 119(44):e2203150119, 2022. doi: 10.1073/pnas.2203150119.
- Erkan O. Buzbas, Berna Devezer, and Bert Baumgaertner. The logical structure of experiments lays the foundation for a theory of reproducibility. *Royal Society Open Science*, 10(3), March 2023. doi: 10.1098/rsos.221042. URL <http://dx.doi.org/10.1098/rsos.221042>.
- Alessandro Checco, Lorenzo Bracciale, Pierpaolo Loreti, Stephen Pinfield, and Giuseppe Bianchi. Ai-assisted peer review. *Humanities and Social Sciences Communications*, 8(1):25, Jan 2021. ISSN 2662-9992. doi: 10.1057/s41599-020-00703-8. URL <https://doi.org/10.1057/s41599-020-00703-8>.
- Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction*, 7(CSCW2):1–32, 2023.
- Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.
- Jiaxin Cui and Rohan Alexander. Same prompt, different outcomes: Evaluating the reproducibility of data analysis by llms, 2026. URL <https://arxiv.org/abs/2602.14349>.
- Jamie Cummins. The threat of analytic flexibility in using large language models to simulate human data: A call to attention. *arXiv preprint arXiv:2509.13397*, 2025.
- Mike D’Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. Marg: Multi-agent review generation for scientific papers, 2024. URL <https://arxiv.org/abs/2401.04259>.

- Marco Del Giudice and Steven W. Gangestad. A traveler’s guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, 4(1), January 2021. doi: 10.1177/2515245920954925. URL <http://dx.doi.org/10.1177/2515245920954925>.
- Berna Devezer and Erkan O. Buzbas. The difference between "replicable" and "not replicable" is not itself scientifically replicable, 2026. URL <https://arxiv.org/abs/2604.26268>.
- Berna Devezer, Danielle J. Navarro, Joachim Vandekerckhove, and Erkan Ozge Buzbas. The case for formal methodology in scientific reform. *Royal Society Open Science*, 8(3), March 2021. doi: 10.1098/rsos.200805. URL <http://dx.doi.org/10.1098/rsos.200805>.
- Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. Claim verification in the age of large language models: A survey, 2024. URL <https://arxiv.org/abs/2408.14317>.
- David Donoho. Data science at the singularity. *Harvard Data Science Review*, 6(1), 2024.
- Heather E Douglas. *Science, policy, and the value-free ideal*. University of Pittsburgh Press, 2009.
- Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.
- Nic Fishman and Gabriel Sekeres. Editorial screening when science is cheap. Working paper. Department of Statistics, Harvard University and Department of Economics, Cornell University, 2026.
- Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2007.
- Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 2013.
- Andrew Gelman, Amy Krefman, Lauren Kennedy, and Jessica Hullman. Hypothesizing an effect size by considering individual variation. *arXiv preprint arXiv:2604.08421*, 2026.
- Alexander Goldberg, Ihsan Ullah, Thanh Gia Hieu Khuong, Benedictus Kent Rachmat, Zhen Xu, Isabelle Guyon, and Nihar B. Shah. Usefulness of llms as an author checklist assistant for scientific papers: Neurips’24 experiment, 2024. URL <https://arxiv.org/abs/2411.03417>.
- Charles AE Goodhart. Problems of monetary management: the uk experience. In *Monetary theory and practice: The UK experience*, pages 91–121. Springer, 1984.
- Serafin Grundl. A comparison of agentic ai systems and human economists. *Available at SSRN 6551598*, 2026.
- Ziyang Guo, Yifan Wu, Jason Hartline, and Jessica Hullman. Explaining and improving information complementarities in multi-agent decision-making. In *The Fourteenth International Conference on Learning Representations*, 2025.
- Bengt Holmstrom and Paul Milgrom. Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, and Organization*, 7(special\_issue): 24–52, 1991.
- Chuxuan Hu, Liyun Zhang, Yeji Lim, Aum Wadhwani, Austin Peters, and Daniel Kang. Repro-bench: Can agentic ai systems assess the reproducibility of social science research?, 2025. URL <https://arxiv.org/abs/2507.18901>.
- Wenqian Huang, Albert J Menkveld, and Shihao Yu. Ai “errors”. Technical report, Working paper, Bank for International Settlements, Vrije Universiteit, 2026.

- Maximilian Idahl and Zahra Ahmadi. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In Nouha Dziri, Sean (Xiang) Ren, and Shizhe Diao, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 550–562, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-191-9. doi: 10.18653/v1/2025.naacl-demo.44. URL <https://aclanthology.org/2025.naacl-demo.44/>.
- Shashidhar Reddy Javaji, Yupeng Cao, Haohang Li, Yangyang Yu, Nikhil Muralidhar, and Zining Zhu. Can ai validate science? benchmarking llms for accurate scientific claim → evidence reasoning, 2025. URL <https://arxiv.org/abs/2506.08235>.
- Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation*, 8(4):1–23, 2020.
- Benjamin Kohler, David Zollikofer, Johanna Einsiedler, Alexander Hoyle, and Elliott Ash. Read the paper, write the code: Agentic reproduction of social-science results. *arXiv preprint arXiv:2604.21965*, 2026.
- Thomas Kosch and Sebastian Feger. Prompt-hacking: The new p-hacking?, 2025. URL <https://arxiv.org/abs/2504.14571>.
- Thomas S Kuhn. Objectivity, value judgment, and theory choice. In *Arguing about science*, pages 74–86. Routledge, 2012.
- Keigo Kusumegi, Xinyu Yang, Paul Ginsparg, Mathijs de Vaan, Toby Stuart, and Yian Yin. Scientific production in the era of large language models. *Science*, 390(6779):1240–1243, 2025.
- Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis, 2023. URL <https://arxiv.org/abs/2310.01783>.
- Dennis V Lindley. The philosophy of statistics. *Journal of the Royal Statistical Society Series D: The Statistician*, 49(3):293–337, 2000.
- Haokun Liu and Chenhao Tan. Ai-assisted reviewing is necessary for avoiding the review death spiral. Preprint. March 5, 2026, 2026.
- Jiachen Liu, Jiaxin Pei, Jintao Huang, Chenglei Si, Ao Qu, Xiangru Tang, Runyu Lu, Lichang Chen, Xiaoyan Bai, Haizhong Zheng, Carl Chen, Zhiyang Chen, Haojie Ye, Yujuan Fu, Zexue He, Zijian Jin, Zhenyu Zhang, Shangquan Sun, Maestro Harmon, John Dianzhuo Wang, Jianqiao Zeng, Jiachen Sun, Mingyuan Wu, Baoyu Zhou, Yuchen You, Shijian Lu, Yiming Qiu, Fan Lai, Yuan Yuan, Yao Li, Junyuan Hong, Ruihao Zhu, Beidi Chen, Alex Pentland, Ang Chen, Mosharaf Chowdhury, and Zechen Zhang. The last human-written paper: Agent-native research artifacts, 2026a. URL <https://arxiv.org/abs/2604.24658>.
- Xuanyou Liu, Ziyang Guo, Jason Hartline, and Jessica Hullman. Learning checklists to reduce lottery effects in scientific peer review, 2026b. Under review.
- Helen E Longino. *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton university press, 1990.
- Helen E Longino. *The fate of knowledge*. Princeton University Press, 2002.
- Sebastian Porsdam Mann, Mateo Aboy, Joel Jiehaoh Seah, Zhicheng Lin, Xufei Luo, Daniel Rodger, Hazem Zohny, Timo Minssen, Julian Savulescu, and Brian D Earp. Ai and the future of academic peer review. *arXiv preprint arXiv:2509.14189*, 2025.
- Xiao-Li Meng. Reproducibility, replicability, and reliability. *Harvard Data Science Review*, 2(4):10, 2020.

- Bang Nguyen, Dominik Soós, Qian Ma, Rochana R. Obadage, Zack Ranjan, Sai Koneru, Anna Szabelska, Adam Gill, Timothy M. Errington, Shakhlo Nematova, Sarah Rajtmajer, Jian Wu, and Meng Jiang. Replicatorbench: Benchmarking llm agents for replicability in social and behavioral sciences, 2026. URL <https://arxiv.org/abs/2602.11354>.
- Brian A Nosek, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen, et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015.
- Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.
- Juan Perdomo, Tijana Zrnic, Celestine Mandler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609, 2020.
- Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d’Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of machine learning research*, 22(164):1–20, 2021.
- Neil Rathi, Dan Jurafsky, and Kaitlyn Zhou. Humans overrely on overconfident language models, across languages. *arXiv preprint arXiv:2507.06306*, 2025.
- Horst W J Rittel and Melvin M Webber. Dilemmas in a general theory of planning. *Policy Sciences*, 4(2):155–169, 1973.
- Julia M. Rohrer, Jessica Hullman, and Andrew Gelman. What’s a multiverse good for anyway?, February 2026. URL [http://dx.doi.org/10.31234/osf.io/37g29\\_v1](http://dx.doi.org/10.31234/osf.io/37g29_v1).
- Giuseppe Romeo and Daniela Conti. Exploring automation bias in human–ai collaboration: a review and implications for explainable ai. *Ai & Society*, 41(1):259–278, 2026.
- David E M Sappington. Incentives in principal-agent relationships. *Journal of economic Perspectives*, 5(2):45–66, 1991.
- Nihar B Shah. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87, 2022.
- Syed Mehtab Hussain Shah, Frank Hopfgartner, and Arnim Bleier. Automating computational reproducibility in social science: Comparing prompt-based and agent-based approaches, 2026. URL <https://arxiv.org/abs/2602.08561>.
- Raphael Silberzahn and Eric L. Uhlmann. Crowdsourced research: Many hands make tight work. *Nature*, 526(7572):189–191, Oct 2015. ISSN 1476-4687. doi: 10.1038/526189a. URL <https://doi.org/10.1038/526189a>.
- Raphael Silberzahn, Eric L. Uhlmann, Daniel P. Martin, Pasquale Anselmi, Frederik Aust, Eli Awtrey, et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 2018. doi: 10.1177/2515245917747646.
- Daniel J Simons, Yuichi Shoda, and D Stephen Lindsay. Constraints on generality (cog): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6):1123–1128, 2017.
- Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. Specification curve analysis. *Nature human behaviour*, 4(11):1208–1214, 2020.
- Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. Paperbench: Evaluating ai’s ability to replicate ai research, 2025. URL <https://arxiv.org/abs/2504.01848>.
- Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016.

- Chenhao Tan and Haokun Liu. The mirage of autonomous ai scientists. Preprint. February 2, 2026, 2026.
- Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, Dov Te’eni, and Iddo Drori. Ai-driven review systems: Evaluating llms in scalable and bias-aware academic reviews, 2024. URL <https://arxiv.org/abs/2408.10365>.
- Erik van Zwet, Andrew Gelman, and Witold Więcek. A statistical case for qualified scientific optimism. Unpublished manuscript. Available at [https://sites.stat.columbia.edu/gelman/research/unpublished/A\\_statistical\\_case\\_for\\_qualified\\_scientific\\_optimism.pdf](https://sites.stat.columbia.edu/gelman/research/unpublished/A_statistical_case_for_qualified_scientific_optimism.pdf), 2026.
- Robert Ward, Alex Jones, and Lutz Bornmann. Judges matter more than papers in post-publication research assessment. *arXiv preprint arXiv:2607.09783*, 2026.
- Qiyao Wei, Samuel Holt, Jing Yang, Markus Wulfmeier, and Mihaela van der Schaar. The ai imperative: Scaling high-quality peer review in machine learning, 2025. URL <https://arxiv.org/abs/2506.08134>.
- Wenqing Wu, Yi Zhao, Yuzhuo Wang, Siyou Li, Juexi Shao, Yunfei Long, and Chengzhi Zhang. Novbench: Evaluating large language models on academic paper novelty assessment, 2026. URL <https://arxiv.org/abs/2604.11543>.
- Yifan Wu, Ziyang Guo, Michalis Mamakos, Jason Hartline, and Jessica Hullman. The rational agent benchmark for data visualization. *IEEE transactions on visualization and computer graphics*, 30(1):338–347, 2023.
- Sarina Xi, Vishisht Rao, Justin Payan, and Nihar B. Shah. Flaws: A benchmark for error identification and localization in scientific papers, 2025. URL <https://arxiv.org/abs/2511.21843>.
- Hang Xu, Ling Yue, Chaoqian Ouyang, Yuchen Liu, Libin Zheng, Shaowu Pan, Shimin Di, and Min-Ling Zhang. Factreview: Evidence-grounded reviews with literature positioning and execution-based claim verification, 2026. URL <https://arxiv.org/abs/2604.04074>.
- Yiqing Xu and Leo Yang Yang. Scaling reproducibility: An ai-assisted workflow for large-scale replication and reanalysis, 2026. URL <https://arxiv.org/abs/2602.16733>.
- Tal Yarkoni. The generalizability crisis. *Behavioral and Brain Sciences*, 45:e1, 2022.
- Ming Zhang, Kexin Tan, Yueyuan Huang, Yujiong Shen, Chunchun Ma, Li Ju, Xinran Zhang, Yuhui Wang, Wenqing Jing, Jingyi Deng, Huayu Sha, Binze Hu, Jingqi Tong, Changhao Jiang, Yage Geng, Yuankai Ying, Yue Zhang, Zhangyue Yin, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. Opennovelty: An llm-powered agentic system for verifiable scholarly novelty assessment, 2026. URL <https://arxiv.org/abs/2601.01576>.
- Tianmai M. Zhang and Neil F. Abernethy. Reviewing scientific papers for critical problems with reasoning llms: Baseline approaches and automatic evaluation, 2026. URL <https://arxiv.org/abs/2505.23824>.

## A A signal-to-decision specification using significance and replication signals

We illustrate decision theoretic signal-to-decision specification using specific proxies and states implied by the mixture model of van Zwet et al. [2026]. The authors use maximum likelihood to estimate a latent distribution of signal-to-noise ratios from samples of absolute  $z$ -values across multiple disciplines. We use their results for the OSC corpus of 100 psychology replication studies as an illustrative prior (after simplifying to remove the most diffuse of their four mixture components due to instability, which contributed only 1% of the total mass), and simulate signed original and replication  $z$ -values under their assumed model. In this model, significance and whether an exact replication was successful are noisy functions of the underlying signal-to-noise ratio. We consider the value of these signals for review decision-making, considering two possible definitions of the latent epistemic state on which decision quality depends.<sup>3</sup>

<sup>3</sup>Code is available at [https://github.com/jhullman/AI\\_metascience\\_spec](https://github.com/jhullman/AI_metascience_spec).

### A.1 Decision problem setup

**Action, general latent state, utility function.** The reviewer faces a binary action,  $\mathcal{A} = \{\text{do not flag, flag}\}$ . The intended use of the signal is to decide whether to flag a claim for evidential fragility. Let  $\theta \in \{0, 1\}$  denote the latent epistemic state, where  $\theta = 1$  means the claim satisfies the epistemic property of interest and  $\theta = 0$  means it does not. Below we define two different candidate epistemic states of whether the study has sufficiently high signal-to-noise and whether the observed effect has the correct sign.

We use a simple symmetric utility function:

$$u(a, \theta) = \begin{cases} 1, & a = \text{do not flag and } \theta = 1, \\ 1, & a = \text{flag and } \theta = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Note that under this utility function, the Bayes-optimal reviewer flags whenever

$$P(\theta = 0 \mid S) > \frac{1}{2},$$

or equivalently does not flag whenever

$$P(\theta = 1 \mid S) > \frac{1}{2}.$$

**Specific latent state, signals, and data-generating model.** Let  $\Lambda$  denote the (unobservable) signal-to-noise ratio of a reported estimate. Following the signal-to-noise model assumed by [van Zwet et al., 2026],  $\Lambda \sim H$ , where  $H$  is a corpus-level distribution of signal-to-noise ratios. We use the OSC-fitted distribution  $H$  as a prior over signal-to-noise ratios, where conditional on a draw  $\Lambda \sim H$ , we simulate a reported study statistic  $Z$  and an exact replication statistic  $Z_{\text{rep}}$ , assuming  $z$ -values are equivalent to signal-to-noise ratios plus independent standard normal errors,

$$Z = \Lambda + \epsilon, \epsilon \sim \mathcal{N}(0, 1).$$

Similarly, we assume an exact replication produces

$$Z_{\text{rep}} = \Lambda + \epsilon', \epsilon' \sim \mathcal{N}(0, 1), \text{ where } Z \perp\!\!\!\perp Z_{\text{rep}} \mid \Lambda.$$

We consider two latent epistemic states under these assumptions. The first is whether the study has a **sufficiently high signal-to-noise ratio** for the intended decision:

$$\theta_{\text{snr}} = \mathbb{I}(|\Lambda| \geq \lambda_0).$$

where  $\lambda_0 > 0$  denotes a threshold chosen for the decision context. We use  $\lambda_0 = 2.8$ , which is the signal-to-noise ratio corresponding to 80% power at the 5% two-sided level under the assumed data-generating model.

The second is whether the observed effect has the **correct sign**:

$$\theta_{\text{sign}} = \mathbb{I}(\Lambda Z > 0).$$

We also consider two observable signals. The first is a **statistical significance** indicator calculated from the original study  $z$ -value,

$$S_{\text{sig}} = \mathbb{I}(|Z| \geq 1.96).$$

The second is **exact replication success**:

$$S_{\text{rep}} = \mathbb{I}(ZZ_{\text{rep}} > 0, |Z_{\text{rep}}| \geq 1.96),$$

which indicates that the exact replication is statistically significant in the same direction as the original result.

We compute posteriors over  $\Lambda$  conditioned on the signed  $z$ -value  $Z = z$  as intermediate quantities, then marginalize over  $z$  within the significant or non-significant region to obtain posteriors conditioned on  $S_{\text{sig}}$ . Although van Zwet et al. use absolute  $z$ -values for model fitting ( $H$  is symmetric), conditioning on the sign of  $z$  is necessary for  $\theta_{\text{sign}}$ , and produces equivalent results for  $\theta_{\text{snr}}$ .

## A.2 Posterior distributions

**Posterior over signal-to-noise ratio.** We first define the posterior distribution over the signal-to-noise ratio. Posterior distributions for both latent epistemic states depend on this posterior, as they express the fraction of the signal-to-noise ratio posterior mass satisfying the condition given by  $\theta$ .

After observing  $Z = z$ , the posterior density evaluated at  $\Lambda = \lambda$  is

$$p_{\Lambda|Z}(\lambda | z) = \frac{\phi(z - \lambda)h(\lambda)}{\int \phi(z - \lambda')h(\lambda') d\lambda'},$$

where  $h$  is the density of  $H$  and  $\phi$  is the standard normal density.

If the reviewer also observes replication success  $S_{\text{rep}} = s_{\text{rep}}$ , then

$$p_{\Lambda|Z, S_{\text{rep}}}(\lambda | z, s_{\text{rep}}) \propto \phi(z - \lambda)P(S_{\text{rep}} = s_{\text{rep}} | z, \lambda)h(\lambda).$$

For  $z > 0$ ,

$$P(S_{\text{rep}} = 1 | z, \lambda) = 1 - \Phi(1.96 - \lambda),$$

and for  $z < 0$ ,

$$P(S_{\text{rep}} = 1 | z, \lambda) = \Phi(-1.96 - \lambda).$$

In either case,

$$P(S_{\text{rep}} = 0 | z, \lambda) = 1 - P(S_{\text{rep}} = 1 | z, \lambda).$$

**Posterior over sufficient signal-to-noise ratio.** Let  $p_Z(z) = \int \phi(z - \lambda)h(\lambda) d\lambda$  be the marginal density of  $Z$ . We define the intermediate quantity

$$P(\theta_{\text{snr}} = 1 | z) = \int \mathbb{I}(|\lambda| \geq \lambda_0) p_{\Lambda|Z}(\lambda | z) d\lambda.$$

After observing  $S_{\text{sig}} = s$ , we get the posterior by marginalizing over  $z$ :

$$P(\theta_{\text{snr}} = 1 | S_{\text{sig}} = s) = \frac{\int_{\{z: \mathbb{I}(|z| \geq 1.96) = s\}} P(\theta_{\text{snr}} = 1 | z) p_Z(z) dz}{P(S_{\text{sig}} = s)}.$$

After also observing  $S_{\text{rep}} = s_{\text{rep}}$ , define

$$P(\theta_{\text{snr}} = 1 | z, s_{\text{rep}}) = \int \mathbb{I}(|\lambda| \geq \lambda_0) p_{\Lambda|Z, S_{\text{rep}}}(\lambda | z, s_{\text{rep}}) d\lambda.$$

The joint posterior is

$$P(\theta_{\text{snr}} = 1 | S_{\text{sig}} = s, S_{\text{rep}} = s_{\text{rep}}) = \frac{\int_{\{z: \mathbb{I}(|z| \geq 1.96) = s\}} P(\theta_{\text{snr}} = 1 | z, s_{\text{rep}}) p_Z(z) P(S_{\text{rep}} = s_{\text{rep}} | z) dz}{\int_{\{z: \mathbb{I}(|z| \geq 1.96) = s\}} p_Z(z) P(S_{\text{rep}} = s_{\text{rep}} | z) dz},$$

where  $P(S_{\text{rep}} = s_{\text{rep}} | z) = \int P(S_{\text{rep}} = s_{\text{rep}} | z, \lambda) p_{\Lambda|Z}(\lambda | z) d\lambda$ .

**Posterior over correct sign.** We apply the same marginalization applies, first calculating

$$P(\theta_{\text{sign}} = 1 | z) = \int \mathbb{I}(\lambda z > 0) p_{\Lambda|Z}(\lambda | z) d\lambda$$

and

$$P(\theta_{\text{sign}} = 1 | z, s_{\text{rep}}) = \int \mathbb{I}(\lambda z > 0) p_{\Lambda|Z, S_{\text{rep}}}(\lambda | z, s_{\text{rep}}) d\lambda.$$

We obtain the posteriors  $P(\theta_{\text{sign}} = 1 | S_{\text{sig}} = s)$  and  $P(\theta_{\text{sign}} = 1 | S_{\text{sig}} = s, S_{\text{rep}} = s_{\text{rep}})$  by the same weighted averages as for  $\theta_{\text{snr}}$  above, where we replace  $P(\theta_{\text{snr}} = 1 | \cdot)$  with  $P(\theta_{\text{sign}} = 1 | \cdot)$ .

### A.3 Quantifying the value of the signals

**Baseline value without signals.** For either target state  $\theta \in \{\theta_{\text{sign}}, \theta_{\text{snr}}\}$ , let  $p(\theta)$  denote the prior probability of the favorable state before observing any signal. The expected value of the best prior-only action is

$$V_0(\theta) = \max_{a \in \mathcal{A}} \mathbb{E}_{\theta}[u(a, \theta)].$$

For the psychology replication corpus, the baseline for sufficient signal-to-noise ratio is 0.794. Without any signal, the best constant action is always flag (since  $P(\theta_{\text{snr}} = 1) = 0.206 < 0.5$ ), so the rational agent baseline would be right 79.4% of the time.

The baseline for correct sign is 0.783. The best constant action is don't flag (since  $P(\theta_{\text{sign}} = 1) = 0.783 > 0.5$  before any signal).

**Benchmark and value of observing statistical significance.** The benchmark for observing only statistical significance is

$$V_{\text{sig}}(\theta) = \mathbb{E}_{S_{\text{sig}}} [\max\{P(\theta = 1 \mid S_{\text{sig}}), P(\theta = 0 \mid S_{\text{sig}})\}],$$

which is 0.783 for  $\theta_{\text{sign}}$ , and 0.843 for  $\theta_{\text{snr}}$ .

The value of the significance signal is

$$\Delta_{\text{sig}}(\theta) = V_{\text{sig}}(\theta) - V_0(\theta),$$

which is 0.000 for  $\theta_{\text{sign}}$  and 0.049 for  $\theta_{\text{snr}}$ .

**Benchmark and value of replication success.** The benchmark for observing only exact replication success is

$$V_{\text{rep}}(\theta) = \mathbb{E}_{S_{\text{rep}}} \left[ \max_{a \in \mathcal{A}} \mathbb{E}_{\theta|S_{\text{rep}}} [u(a, \theta)] \right],$$

which is 0.783 for  $\theta_{\text{sign}}$ , and 0.862 for  $\theta_{\text{snr}}$ .

The value of replication success alone is

$$\Delta_{\text{rep}}(\theta) = V_{\text{rep}}(\theta) - V_0(\theta),$$

or 0.000 for  $\theta_{\text{sign}}$ , and 0.068 for  $\theta_{\text{snr}}$ .

**Benchmark and value of both signals.** The benchmark for both signals ( $S_{\text{both}} = (S_{\text{sig}}, S_{\text{rep}})$ ) is

$$V_{\text{both}}(\theta) = \mathbb{E}_{S_{\text{both}}} \left[ \max_{a \in \mathcal{A}} \mathbb{E}_{\theta|S_{\text{both}}} [u(a, \theta)] \right].$$

For  $\theta_{\text{sign}}$ , this is 0.783. For  $\theta_{\text{snr}}$ , it is 0.925.

The value of observing both signals is

$$\Delta_{\text{both}}(\theta) = V_{\text{both}}(\theta) - V_0(\theta),$$

or 0.000 for  $\theta_{\text{sign}}$ , and 0.131 for  $\theta_{\text{snr}}$ .

Figure 4 depicts the value of all signals for the two latent states.

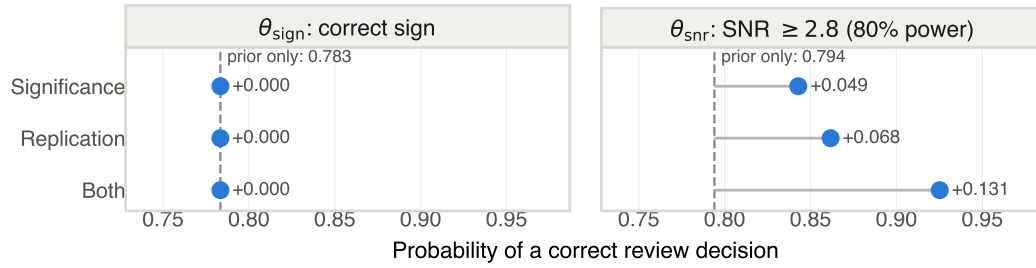


Figure 4: Expected utility of the Bayes-optimal decision rule under each signal (quantifying the ex-ante best case value of the signal), for two latent epistemic states,  $\theta_{\text{sign}}$  (left) and  $\theta_{\text{snr}}$  (right). Each dot shows the expected utility given the signal listed on the left. The dashed line is the baseline  $V_0$ , representing the utility of the best constant action without any signal. Under the symmetric 0/1 utility function, utility is the probability of making the correct review decision.