Stop Treating Metascientific Heuristics as Quality Filters in AI Review

Jessica Hullman
Northwestern University
jhullman@northwestern.edu

June 23, 2026

Abstract

AI-implemented checks for reproducibility, robustness, preregistration, claim scope, and other intended proxies for scientific credibility can extend human reviewers' capabilities. However, treating metascientific heuristics–whose theoretical grounding remains contested or incomplete–as necessary and sufficient signals for filtering out bad science is counterproductive to scientific progress. The emerging literature blurs the line between integrity filtering, based on necessary but insufficient signals of validity like reproducibility of stated results or lack of fake citations, and epistemic filtering, which uses machine-detectable signals to judge scientific quality. Drawing on critical metascience, we show that commonly proposed signals of research quality are insufficiently justified as general indicators of scientific value. The answer is not necessarily to ban AI in review, given the deluge of submissions venues are facing. Instead, in recognition of how any use of automated signals–even when deployed with human oversight–will shape attention and create incentives upstream, developers of AI review tools should explicitly specify their assumptions about how proxy signals inform on scientific quality in the context of specific review decisions. This approach treats AI review contributions as contestable decision policies that will shape future research, acknowledging the value-laden nature of scientific judgment and surfacing relevant tradeoffs.

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 peer reviewers lack sufficient time or incentives to fully interrogate a submission, AI reviewers can conduct extensive checks for rigor. 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 novel and significant to society.

However, automated science evaluation presents a high stakes reward design problem. Machine learning researchers are building the infrastructure that will structure which science receives attention in the future. What sort of 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 (Steegen et al., 2016; Simonsohn et al., 2020), and aligning claims with the evidence at hand (Simons et al., 2017; Yarkoni, 2022). AI science 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. This paper takes the position that **by deploying reform heuristics as decision criteria without specifying their presumed signal-to-decision role, the mechanization of metascientific criteria in AI science evaluation is counterproductive to scientific progress**. ML researchers are borrowing popular-–but unsubstantiated—narratives in which simple steps toward robustness restore credibility. A more discriminating read of contemporary metascience sees ideals such as transparency, robustness, replication, and multiverse reporting as heuristics, which may be useful prompts for deliberation but fall short of theoretically or empirically validated solutions. For example, a dataset-robust result may indicate a genuine capability, or it may represent a shared leakage pattern or an evaluation metric that rewards superficial behavior. Indeed, a growing critical metascience literature argues that many reform proposals–despite their good intentions–have been subject to the same kinds of overclaiming that they aim to critique, and can create perverse incentives when enacted as policy (Devezer et al., 2021; Bak-Coleman and Devezer, 2024; Gervais, 2021; Rubin, 2023; Center for Open Science, 2024).

Within machine learning, existing critiques emphasize that current AI reviewers are gameable, arguing that they should not be used until they can be empirically shown to resist gaming (Baumann et al., 2026). We agree that gaming is a serious concern, but disagree that striving for nongameable AI reviewers is a viable goal, as this implies that we may find a stable, non-conflicting and context-independent set of detectable criteria that distinguish high quality science. This assumption overlooks the ways in which scientific progress, and notions of scientific objectivity and theory choice, are imprecise, potentially conflicting, and dependent on socially situated standards and plural epistemic values (Kuhn, 2012; Longino, 1990, 2002). Any AI review system that uses machine-detectable proxies to allocate scientific attention–including criteria that borrow from popular reform narratives like replication, preregistration, and robustness to perturbations–should be expected to create an optimization target. The relevant question is instead what trade-offs to expect, and how they align with the community's values. Current AI systems are well-suited to implement basic integrity checks such as detecting plagiarism, fake citations, and missing materials for reproducing results. What proposers of tools must avoid is unwarranted leaps from integrity filtering to epistemic filtering, where criteria thought to be associated with rigor are used to rank, reject, or prioritize work as more scientifically valuable without specifying the auxiliary assumptions required to give them value. Without this, use of AI for science evaluation perpetuates myths about the legibility of scientific virtue and extent to which science can be reduced to a checklist.

Ultimately, to leverage the useful functions of AI review while avoiding canonizing poorly understood heuristics, the community must move toward explicitly engaging with the tradeoffs of particular filters, including what signals are being detected, what intermediate epistemic target they are claimed to inform and under what assumptions, how they should affect review decisions under those assumptions, how reviewers may misinterpret them, and how authors may adapt. This path forward cannot resolve the contestability of any proxy signal, but at least treats AI review as the policy evaluation problem that it is, one inherently subject to value-laden judgment.

In what follows, we identify common proxies and assumptions undergirding recent research on AI for scientific evaluation. We show how these assumptions ultimately imply that machine-detectable markers of good science are sufficiently reliable to be used for epistemic filtering, a perspective at odds with contemporary critical metascience. We discuss why a commonly proposed fallback plan–using AI to surface presumed flaws for human judgment–is not sufficient to prevent problems brought by undertheorized proxies. We then motivate signal-to-decision specification as a way to make assumptions, trade-offs, and failure modes explicit. We are cautiously optimistic that the increased urgency of scaling scientific judgment and ability to implement automated review policies could increase attention to metascientific concerns and catalyze a more thoughtful approach–if the community can resist turning heuristics into automated reward models.

2 From Integrity Filtering to Epistemic Significance

Many AI review papers motivate automation by pointing to weaknesses of peer review (Shah, 2022; Mann et al., 2025) and the prospect that AI-generated papers will further strain reviewer capacity (Bergstrom and Gross, 2026; Liu and Tan, 2026). However, using (potentially biased) human reviews as ground truth to evaluate AI reviewers is not ideal (Checco et al., 2021; Tyser et al., 2024; Idahl and Ahmadi, 2025), and

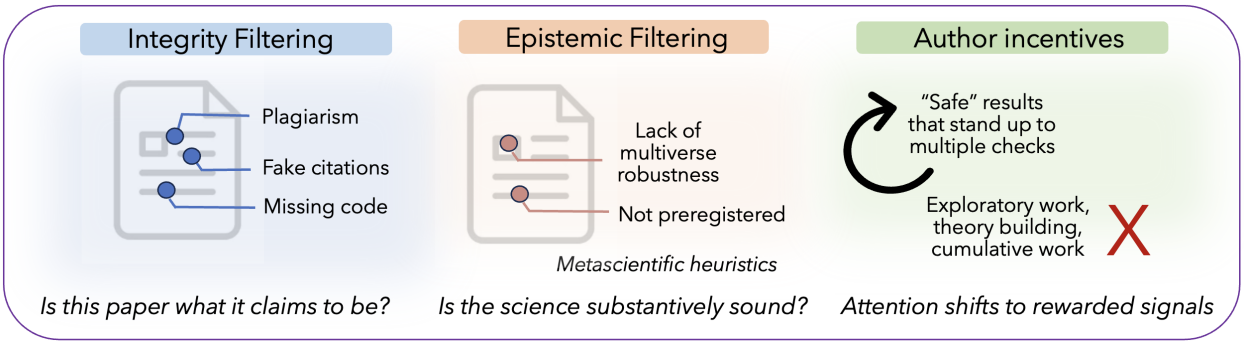

Figure 1: Research in AI review blurs the line between integrity filtering, which checks for necessary but insufficient signs that a piece of research is what it claims to be, like an absence of plagiarism or fake citations, and epistemic filtering, which aims to assess substantive credibility using contestable metascientific heuristics.

subjective ratings of AI reviews can be swayed by style or author self-interest (Liang et al., 2023; D'Arcy et al., 2024). These challenges, along with early evidence that LLM reviewers are better at finding flaws when prompted with more specific instructions (Liu et al., 2023; Xu et al., 2025) leads researchers to seek "first principles" that capture more fundamental aspects of rigor to base AI review on. As the site of significant deliberation about how to improve scientific credibility, metascience and science reform research seems an obvious place to look.

2.1 Proposed proxies for scientific quality

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 surveyed 40 papers at the intersection of AI and scientific evaluation from the last two years, identified via AI-assisted literature search and citation tracing. We filtered this set to 23 papers 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. For example, the Open Science movement's emphasis on reproducibility is cited as "central to scientific credibility" (Xu and Yang, 2026) and "essential for promoting rigor in research processes" (Hu et al., 2025). Beyond reproducibility, authors justify replication testing 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).

2.2 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. LLM review is implied to be useful for enabling automated filtering or structuring evidence to help human reviewers discriminate high-quality from low-quality papers (Biswas et al., 2026; Liu and Tan, 2026; Liu et al., 2026; 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., 2026). Others propose making acceptance decisions dependent on whether authors have committed to particular meta-scientific approaches, such as preregistration of analyses or showing consistency of a result across many perturbations (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 such signals for credible science. Signals associated with good science can be in opposition and shift over time. For example Copernicus' model of the solar system was simpler but initially inferior to the geocentric

Category	Proxy for scientific quality	Definition
Verifying report results	**Computational reproducibility** (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.
	Methods-only reproducibility (Kohler et al., 2026; Starace et al., 2025)	The ability to verify reported estimates by relying only on the methods description.
	Specification completeness (Liu and Tan, 2026)	Whether full details like hyperparameters, random seeds, and configuration are reported.
	Claim traceability (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	**Replicability** (Nguyen et al., 2026; Bai et al., 2026)	The ability to produce nearly similar results on new data from the same process.
	Robustness of effects to perturbing analysis choices (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.
	Robustness to stochastic perturbation (Cui and Alexander, 2026)	The ability to produce similar results among multiple repetitions of the same specification when using LLMs to analyze data.
Inferential valid	**Pre-specification of analysis plans** (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.
	Statistical significance (Bai et al., 2026)	Whether a paper's results are statistically significant.
Novelty and Coverage	**Novelty and contribution significance** (Biswas et al., 2026; Zhang et al., 2026)	Whether the research is novel and competitive in light of prior work.
	Coverage of related literature (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.

tradition in terms of consistency and accuracy (Kuhn, 2012); only with more development did it prove more accurate. Epistemic filtering may in some cases contribute to 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. Principal-agent problems describe how whenever 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. There is no non-gameable AI reviewer, but as we discuss in Section 4, the rational response is not to bar LLM use in review, as some have suggested (Baumann et al., 2026). Instead, we should treat proxies 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 specific operationalization of the target latent state in the context of a review decision problem.

3 Proxy problems

We illustrate overlooked issues that can arise from careless adoption of proxy-based filters by discussing three conceptually distinct proxies of scientific rigor that AI review research has targeted.

3.1 Promise: Screening for multiverse robustness selects credible results

Imagine giving many teams of analysts the same high level question–e.g., "Are soccer players with darker skin tones more likely to receive a red card than soccer players with lighter skin tones?"–and dataset, and asking them to produce results. Such studies routinely show high variability in conclusions depending on how teams interpret the question and make analytic decisions (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, such as persona prompting, can steer the model toward different results. Telling the analyst to support the hypothesis in all caps, for example, leads to aggressive p-hacking (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 adopting reforms from the broader scientific reform movement. One is "multiverse analysis," which several recent proposals argue must become the new abstraction for scientific communication (Bertran et al., 2026; Liu et al., 2026; 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 (Steegen 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).

Multiverse composition is a matter of art Using result consistency as a signal of credibility treats the multiverse as an inferential tool for answering the substantive question at hand. 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). One challenge is that 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, not a neutral or objectively determined set, and experts within the same area regularly disagree.

Consistency should not be confused for likelihood that a claim is credible A tempting intuition is that a claim that holds for only a minority of specifications in the multiverse is not credible, and one that holds for a majority is. However, this confuses what a multiverse represents: the paths 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. This kind of summarization is nonetheless tempting for authors and readers alike because it reduces the combinatorial explosion of complexity in making sense of results (Hall et al., 2022; Rohrer et al., 2026).

If AI scales up the size of the typical multiverse, these challenges become more formidable. Using a formal model in which "true" results are more likely to return statistically significant p-values, Fishman and Sekeres (2026) argue that the scale of sensitivity analyses commonly used in economic papers–roughly 50 checks per paper–is catastrophically underpowered given AI's dramatic cost reduction. Maintaining the same false discovery requires scaling up the number of specifications proportionally, putting the required number at roughly 7,000. In the absence of theory for how to interpret a multiverse beyond "think about the appropriateness of each choice in light of domain knowledge," such norms may not yield more rigorous science, but will certainly give reviewers much more information to navigate.

3.2 Promise: Screening for replicability selects credible results

If a paper's experiment 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. The importance of reproducibility becomes harder to interpret only when the computational environment is difficult to reconstruct or the process is non-deterministic, as in reinforcement learning, where stochastic variation must be accounted for.

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.

Replication success is arbitrarily defined 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. The difference between significant and non-significant is not itself significant (Gelman and Stern, 2006), and significance can often be achieved by cranking up the sample size. Instead 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. Two procedures (algorithms, models, etc.) rarely produce truly identical results, and we would not expect finite data to suffice to identify when they do.

Inferring replication rate faces identifiability issues 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. In practice, we should expect irreducible uncertainty to remain whenever replication sequences are not exact, such as when implementation details or server architecture might vary between replications. 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). This has led metascientists to argue that the common machinery for converting replication outcomes into verdicts is structurally inadequate for the demarcation task it is asked to perform (Devezer and Buzbas, 2026).

Replicability is neither necessary nor sufficient for true results Even if we could estimate an experiment's long-run replication rate precisely, that quantity would still not be a straightforward signal of whether the reported claim is true. On the one hand, 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. For example, replications may preserve the direction of an original effect estimate but not significance, suggesting a regular relationship measured under noise (van Zwet et al., 2026). Analogously, 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. Or the original evaluation might make an overly broad claim while providing valid narrower evidence, such as when a method said to "improve OOD generalization" only does so for a narrow benchmark distribution, prompting format, or hyperparameter regime. While the broad claim fails to replicate, the underlying capability is not an illusion; it is simply more narrowly scoped than stated.

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 (Devezer et al., 2021). ML analogues include models that consistently outperform baselines because of train-test leakage or poorly tuned baselines. In such cases, high replicability indicates the stability of the evaluation artifact rather than the credibility of the scientific claim.

Consequently, replication experiments are hardly a silver bullet from separating good from bad science. Experts disagree about whether non-replication is a useful source of insight for future work or evidence of poor rigor, and whether individual studies should be expected to replicate cleanly or be evaluated cumulatively (Meng, 2020; National Academies of Sciences, Engineering, and Medicine, 2019; Errington et al., 2021). 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 ablations not providing adequate conditions for inferring causality or underspecified estimands and weak links between theories of capability and evidence.

3.3 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 data collection (Nosek et al., 2018)–should be more strictly required (Cummins, 2025; Kosch and Feger, 2025; Fishman and Sekeres, 2026). At first glance, this appears to be 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 evidential claims. AI can also help address verification gaps with human reviewers. A recent analysis of 201 articles from PLOS journals finds that only 14% of articles had a review that mentioned accessing the preregistration, and only 5% of individual reviewers mentioned accessing the preregistration (Syed, 2023).

Privileging preregistration is statistically incoherent What enthusiasm for AI-mandated preregistration cannot resolve is the epistemic significance of preregistration. 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 any 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. Yet many standard, valid procedures routinely reuse data (Devezer et al., 2021), and many 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 (Steegen et al., 2016), or simply treating the result as exploratory or descriptive.

Requiring preregistration penalizes exploratory research Embracing preregistration as policy is especially threatening to exploratory and descriptive aspects of science, as it equates 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). When preregistration is cast as "solving" the problem of overlap between exploratory and confirmatory phases of research, procedural order becomes a substitute for the deeper reason we expect results to generalize at all: they are embedded in a plausible explanatory account (Szollosi and Donkin, 2021). A paradigm that heavily rewards preregistration may favor questions that are easy to formalize ex-ante over theory-building and measurement and diagnostic refinement.

4 What's missing: Specifying the proposed signal-to-decision pipeline

Science is a complex system, in which progress is supported by the scientific method as well as sociological factors like investement in training and a diversity of views and methods. There is no context-independent, non-conflicting set of detectable markers of good science. Currently it is too easy for authors working on AI review tools to act as if there is, by making bold claims about what criteria are necessary for good science while providing evidence only for the weaker claim that these proxies are detectable.

In light of how deploying AI to review research at scale *will* change incentives in ways that shape future research, we argue that AI review should be approached as a policy design problem. The first step toward this shift is to require more explicit specification of assumptions behind epistemic proposals. The goal of such specification is not to make proxies non-gameable, nor to definitively validate that a given AI review tools will improve science, 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.

We envision two levels of specification that the community can adopt as standards for system-oriented and theoretical work. A minimal standard, appropriate for work that is focused primarily on improving detection, asks those proposing review tools to define the review problem and a detected signal is designed to address, including a formulation of the target latent state the proxy is meant to inform on. This makes clear how the authors are thinking about the decision problem, making it easier for readers to consider failure modes, human interpretation risks, and potential shifts in incentives. A stronger theoretic validation standard formalizes the review decision problem, including specifying a plausible generating model under which the signals can be shown to improve expected decision quality.

In both cases, how to conceive of the target latent state that a proxy informs a reviewer about presents a challenge, because we can never directly observe scientific quality. We nonetheless argue for conceiving the latent epistemic state as more specific than simply "scientific quality," to avoid bundling many independent components (correctness, strength of evidence, external validity, etc.) that proxies may offer very varying (even conflicting) information for. Examples of latent states include the true signal-to-noise ratio of a reported estimate, the soundness of an idea, whether the sign of a reported effect is an artifact of the specific setting studied, or whether a reported ranking of techniques is an artifact of the specific setting studied.

4.1 Minimal standard: Pipeline specification

At a minimum, proposers of AI review tools should define core components of the signal-to-decision pipeline, including the proxy that is the target of detection (e.g., preregistration status, robustness to specification perturbation), the latent construct this proxy is thought to inform on (e.g., credibility of the capability claim the paper makes about a method), and the choice of action the reviewer faces (e.g., accept/reject/flag for human review). This ensures that readers (or users of a new review tool) have all the information they need to consider what conditions would need to be in place for the signal to improve the intended review decision, and what plausible failure cases of signal reliance could look like. Another natural point of reflection is how human reviewers might misinterpret the signals, and what sots of adaptation effects might arise as authors begin to optimize for these proxies (e.g., directing effort away from specific unrewarded procedures). Table 2 demonstrates a minimal specification for multiverse robustness consistency.

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.

Component	Multiverse robustness example
Target epistemic state	Credibility of paper's central claim that their technique is superior to prior techniques.
Signal	Fraction of specifications for which the result remains statistically significant and directionally consistent.
Intended decision use	Flag low-robustness claims for expert scrutiny and possible rejection.
Best-case linking assumption	Specifications are justified by available domain knowledge, estimand-preserving, and conditionally more likely to pass when the claim is credible.
Failure cases	Analysis paths do not share a single estimand, some paths are mutually exclusive (i.e., only one model specification can match true process), some paths are subject to shared measurement or benchmark artifacts.
Human interpretation risk	Reviewers treat the fraction of passing specifications as the probability that the claim is true.
Adaptation risk	Authors neglect exploring approaches with "jagged" gains depending on evaluation conditions. For example, early work on LLMs may have been discouraged under this filter.

Table 2: Example of a minimal signal-to-decision specification for a multiverse robustness signal.

4.2 Decision-theoretic motivation of the value of signals to decisions

Decision theory is a natural formalism for specifying assumptions behind proposed proxies and policies, because it makes explicit the latent state, observed signal, review action, utility function, and data-generating model under which a signal is claimed to improve decisions. By formalizing the decision problem in this way, authors can quantify the value of a particular signal under explicit assumptions about the data-generating model by comparing the expected utility of a Bayesian decision-maker who acts only on the prior with one who observes the signal, updates to the posterior, and chooses the utility-maximizing action. The resulting value-of-information quantity gives a best-case benchmark for whether the signal has decision value under the stated assumptions. We present a formal definition of a decision problem and optimal behavior within it in the appendix.

We emphasize that 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. Open review datasets provide outcomes from which some, albeit imperfect, latent states could be derived (e.g., reviewer soundness scores, AC decisions). Bibliometrics and expert annotation can supplement these sources, providing fodder for authors to empirically motivate assumptions they make about how specific proxies correlate with epistemic states. 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 this approach as a proof of concept in the appendix, using van Zwet et al. (2026)'s model of how signal-to-noise ratios are related to statistical significance and replication success, fitted to a corpus of psychology studies. We show how 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. This exercise shows that that the same signal can have dramatically 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. Full results are in the appendix.

4.3 Complementary efforts

Complementary efforts toward making implications more explicit include game theoretic analysis, which can be used to understand potential adaptation by authors. Recent modeling of editorial screening illustrates such an approach (Fishman and Sekeres, 2026). Ideally, the characterization should address how surfaced proxies should be interpreted and what sorts of negative externalities might arise, such as when greater levels of

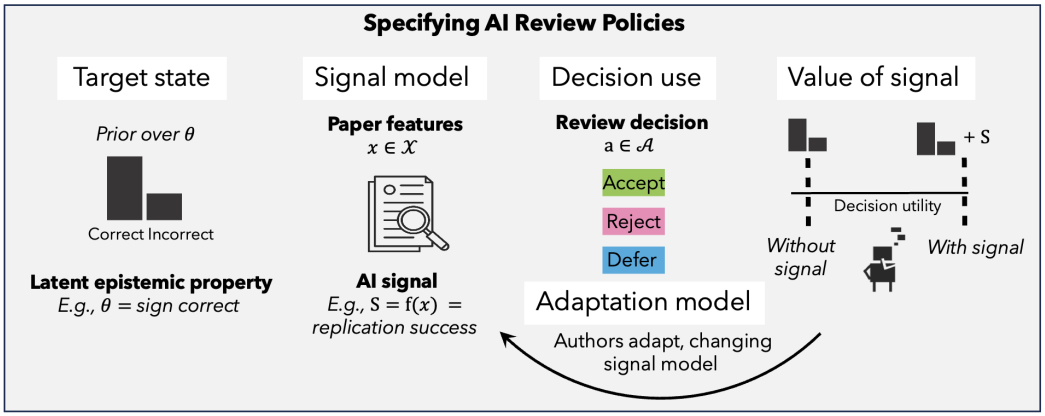

Figure 2: Decision theoretic approach to specifying signal-to-decision pipelines.

effort required from authors to respond and address longer lists of needed revisions from AI reviewers reduces the quality of feedback those individuals provide on papers they must review. Here, agentic simulations of author-driven optimizations may be useful for exploring possible dynamics.

Another direction is to study how human reviewers come to identify issues they consider important for a paper to address, and design AI-surfaced proxies to facilitate processes that are informative for this.

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., 2026; 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, especially when users lack the time, expertise, or incentives to independently verify them (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. The realistic goal is not non-gameability, but explicit characterization of what the proxy measures, what it misses, how it can be optimized, and what tradeoffs follow from using it.

6 Acknowledgment

We thank Kirill Skobelev for comments.

References

K. Auspurg and J. 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, Jan. 2021. doi: 10.1177/23780231211024421. URL http://dx.doi.org/10.1177/23780231211024421.

X. Bai, A. Baumgartner, H. Sun, A. Holtzman, and C. Tan. The story is not the science: Execution-grounded evaluation of mechanistic interpretability research, 2026. URL https://arxiv.org/abs/2602.18458.

J. Bak-Coleman and B. Devezer. Claims about scientific rigour require rigour. Nature Human Behaviour, 8 (10):1890–1891, 2024.

J. Baumann, J. Pei, S. Koyejo, and D. Hovy. Stop automating peer review without rigorous evaluation. In Forty-third International Conference on Machine Learning, 2026. Spotlight.

C. T. Bergstrom and K. Gross. Screening, sorting, and the feedback cycles that imperil peer review. PLoS Biology, 24(2):e3003650, 2026.

M. Bertran, R. Fogliato, and Z. S. Wu. Many ai analysts, one dataset: Navigating the agentic data science multiverse, 2026. URL https://arxiv.org/abs/2602.18710.

J. Biswas, S. Schoepp, G. Vasan, A. Opipari, A. Zhang, Z. Hu, S. Joseph, M. Lease, J. J. Li, P. Stone, K. L. Wagstaff, M. E. Taylor, and O. C. Jenkins. Ai-assisted peer review at scale: The aaai-26 ai review pilot, 2026. URL https://arxiv.org/abs/2604.13940.

N. Breznau, E. M. Rinke, A. 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.

E. O. Buzbas, B. Devezer, and B. Baumgaertner. The logical structure of experiments lays the foundation for a theory of reproducibility. Royal Society Open Science, 10(3), Mar. 2023. doi: 10.1098/rsos.221042. URL http://dx.doi.org/10.1098/rsos.221042.

Center for Open Science. Symposium: Critical perspectives on the metascience reform movement, Mar. 2024. URL https://www.cos.io/critical-perspectives-on-the-metascience-reform-movement. Online symposium, March 7, 2024.

A. Checco, L. Bracciale, P. Loreti, S. Pinfield, and G. 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.

V. Chen, Q. V. Liao, J. Wortman Vaughan, and G. 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.

O. S. Collaboration. Estimating the reproducibility of psychological science. Science, 349(6251):aac4716, 2015.

J. Cui and R. Alexander. Same prompt, different outcomes: Evaluating the reproducibility of data analysis by llms, 2026. URL https://arxiv.org/abs/2602.14349.

J. 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.

M. D'Arcy, T. Hope, L. Birnbaum, and D. Downey. Marg: Multi-agent review generation for scientific papers, 2024. URL https://arxiv.org/abs/2401.04259.

M. Del Giudice and S. 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), Jan. 2021. doi: 10.1177/2515245920954925. URL http://dx.doi.org/10.1177/2515245920954925.

B. Devezer and E. O. Buzbas. The difference between "replicable" and "not replicable" is not itself scientifically replicable, 2026. URL https://arxiv.org/abs/2604.26268.

B. Devezer, D. J. Navarro, J. Vandekerckhove, and E. Ozge Buzbas. The case for formal methodology in scientific reform. Royal Society Open Science, 8(3), Mar. 2021. doi: 10.1098/rsos.200805. URL http://dx.doi.org/10.1098/rsos.200805.

A. Dmonte, R. Oruche, M. Zampieri, P. Calyam, and I. Augenstein. Claim verification in the age of large language models: A survey, 2024. URL https://arxiv.org/abs/2408.14317.

D. Donoho. Data science at the singularity. Harvard Data Science Review, 6(1), 2024.

C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. The reusable holdout: Preserving validity in adaptive data analysis. Science, 349(6248):636–638, 2015.

T. M. Errington, M. Mathur, C. K. Soderberg, A. Denis, N. Perfito, E. Iorns, and B. A. Nosek. Investigating the replicability of preclinical cancer biology. eLife, 10:e71601, dec 2021. ISSN 2050-084X. doi: 10.7554/eLife.71601. URL https://doi.org/10.7554/eLife.71601.

N. Fishman and G. Sekeres. Editorial screening when science is cheap. Working paper. Department of Statistics, Harvard University and Department of Economics, Cornell University, 2026.

A. Gelman and J. Hill. Data analysis using regression and multilevel/hierarchical models. Cambridge university press, 2007.

A. Gelman and E. 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.

A. Gelman and H. Stern. The difference between "significant"' and "not significant" is not itself statistically significant. The American Statistician, 60(4):328–331, 2006.

A. Gelman, A. Krefman, L. Kennedy, and J. Hullman. Hypothesizing an effect size by considering individual variation. arXiv preprint arXiv:2604.08421, 2026.

W. M. Gervais. Practical methodological reform needs good theory. Perspectives on Psychological Science, 16(4):827–843, 2021.

A. Goldberg, I. Ullah, T. G. H. Khuong, B. K. Rachmat, Z. Xu, I. Guyon, and N. B. Shah. Usefulness of llms as an author checklist assistant for scientific papers: Neurips'24 experiment, 2024. URL https://arxiv.org/abs/2411.03417.

S. Grundl. A comparison of agentic ai systems and human economists. Available at SSRN 6551598, 2026.

Z. Guo, Y. Wu, J. Hartline, and J. Hullman. Explaining and improving information complementarities in multi-agent decision-making. In The Fourteenth International Conference on Learning Representations, 2025.

B. D. Hall, Y. Liu, Y. Jansen, P. Dragicevic, F. Chevalier, and M. Kay. A survey of tasks and visualizations in multiverse analysis reports. In Computer Graphics Forum, volume 41, pages 402–426. Wiley Online Library, 2022.

B. Holmstrom and P. 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.

C. Hu, L. Zhang, Y. Lim, A. Wadhwani, A. Peters, and D. Kang. Repro-bench: Can agentic ai systems assess the reproducibility of social science research?, 2025. URL https://arxiv.org/abs/2507.18901.

W. Huang, A. J. Menkveld, and S. Yu. Ai "errors". Technical report, Working paper, Bank for International Settlements, Vrije Universiteit, 2026.

M. Idahl and Z. Ahmadi. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In N. Dziri, S. X. Ren, and S. 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, Apr. 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/.

S. R. Javaji, Y. Cao, H. Li, Y. Yu, N. Muralidhar, and Z. Zhu. Can ai validate science? benchmarking llms for accurate scientific claim → evidence reasoning, 2025. URL https://arxiv.org/abs/2506.08235.

B. Kohler, D. Zollikofer, J. Einsiedler, A. Hoyle, and E. Ash. Read the paper, write the code: Agentic reproduction of social-science results. arXiv preprint arXiv:2604.21965, 2026.

T. Kosch and S. Feger. Prompt-hacking: The new p-hacking?, 2025. URL https://arxiv.org/abs/2504.14571.

T. S. Kuhn. Objectivity, value judgment, and theory choice. In Arguing about science, pages 74–86. Routledge, 2012.

K. Kusumegi, X. Yang, P. Ginsparg, M. de Vaan, T. Stuart, and Y. Yin. Scientific production in the era of large language models. Science, 390(6779):1240–1243, 2025.

W. Liang, Y. Zhang, H. Cao, B. Wang, D. Ding, X. Yang, K. Vodrahalli, S. He, D. Smith, Y. Yin, D. McFarland, and J. Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis, 2023. URL https://arxiv.org/abs/2310.01783.

D. V. Lindley. The philosophy of statistics. Journal of the Royal Statistical Society Series D: The Statistician, 49(3):293–337, 2000.

H. Liu and C. Tan. Ai-assisted reviewing is necessary for avoiding the review death spiral. Preprint. March 5, 2026, 2026.

J. Liu, J. Pei, J. Huang, C. Si, A. Qu, X. Tang, R. Lu, L. Chen, X. Bai, H. Zheng, C. Chen, Z. Chen, H. Ye, Y. Fu, Z. He, Z. Jin, Z. Zhang, S. Sun, M. Harmon, J. D. Wang, J. Zeng, J. Sun, M. Wu, B. Zhou, Y. You, S. Lu, Y. Qiu, F. Lai, Y. Yuan, Y. Li, J. Hong, R. Zhu, B. Chen, A. Pentland, A. Chen, M. Chowdhury, and Z. Zhang. The last human-written paper: Agent-native research artifacts, 2026. URL https://arxiv.org/abs/2604.24658.

R. Liu, S. Jecmen, V. Conitzer, F. Fang, and N. B. Shah. Testing for reviewer anchoring in peer review: A randomized controlled trial, 2023. URL https://arxiv.org/abs/2307.05443.

H. E. Longino. Science as social knowledge: Values and objectivity in scientific inquiry. 1990.

H. E. Longino. The fate of knowledge. 2002.

S. P. Mann, M. Aboy, J. J. Seah, Z. Lin, X. Luo, D. Rodger, H. Zohny, T. Minssen, J. Savulescu, and B. D. Earp. Ai and the future of academic peer review. arXiv preprint arXiv:2509.14189, 2025.

X.-L. Meng. Reproducibility, replicability, and reliability. Harvard Data Science Review, 2(4):10, 2020.

National Academies of Sciences, Engineering, and Medicine. Reproducibility and Replicability in Science. The National Academies Press, Washington, DC, 2019. doi: 10.17226/25303. URL https://nap.nationalacademies.org/catalog/25303.

B. Nguyen, D. Soós, Q. Ma, R. R. Obadage, Z. Ranjan, S. Koneru, A. Szabelska, A. Gill, T. M. Errington, S. Nematova, S. Rajtmajer, J. Wu, and M. Jiang. Replicatorbench: Benchmarking llm agents for replicability in social and behavioral sciences, 2026. URL https://arxiv.org/abs/2602.11354.

B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, et al. Promoting an open research culture. Science, 348(6242):1422–1425, 2015.

B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. The preregistration revolution. Proceedings of the National Academy of Sciences, 115(11):2600–2606, 2018.

J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, and H. 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.

N. Rathi, D. Jurafsky, and K. Zhou. Humans overrely on overconfident language models, across languages. arXiv preprint arXiv:2507.06306, 2025.

J. M. Rohrer, J. Hullman, and A. Gelman. What's a multiverse good for anyway?, Feb. 2026. URL http://dx.doi.org/10.31234/osf.io/37g29_v1.

G. Romeo and D. Conti. Exploring automation bias in human–ai collaboration: a review and implications for explainable ai. Ai & Society, 41(1):259–278, 2026.

M. Rubin. Questionable metascience practices. 2023.

D. E. M. Sappington. Incentives in principal-agent relationships. Journal of economic Perspectives, 5(2): 45–66, 1991.

N. B. Shah. Challenges, experiments, and computational solutions in peer review. Communications of the ACM, 65(6):76–87, 2022.

S. M. H. Shah, F. Hopfgartner, and A. Bleier. Automating computational reproducibility in social science: Comparing prompt-based and agent-based approaches, 2026. URL https://arxiv.org/abs/2602.08561.

R. Silberzahn and E. 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.

R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. 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.

D. J. Simons, Y. Shoda, and D. S. Lindsay. Constraints on generality (cog): A proposed addition to all empirical papers. Perspectives on Psychological Science, 12(6):1123–1128, 2017.

U. Simonsohn, J. P. Simmons, and L. D. Nelson. Specification curve analysis. Nature human behaviour, 4 (11):1208–1214, 2020.

G. Starace, O. Jaffe, D. Sherburn, J. Aung, J. S. Chan, L. Maksin, R. Dias, E. Mays, B. Kinsella, W. Thompson, J. Heidecke, A. Glaese, and T. Patwardhan. Paperbench: Evaluating ai's ability to replicate ai research, 2025. URL https://arxiv.org/abs/2504.01848.

S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel. Increasing transparency through a multiverse analysis. Perspectives on Psychological Science, 11(5):702–712, 2016.

M. Syed. Some data indicating that editors and reviewers do not check preregistrations during the review process. PsyArXiv Preprints, 2023.

A. Szollosi and C. Donkin. Arrested theory development: The misguided distinction between exploratory and confirmatory research. Perspectives on Psychological Science, 16(4):717–724, 2021.

C. Tan and H. Liu. The mirage of autonomous ai scientists. Preprint. February 2, 2026, 2026.

K. Tyser, B. Segev, G. Longhitano, X.-Y. Zhang, Z. Meeks, J. Lee, U. Garg, N. Belsten, A. Shporer, M. Udell, D. Te'eni, and I. Drori. Ai-driven review systems: Evaluating llms in scalable and bias-aware academic reviews, 2024. URL https://arxiv.org/abs/2408.10365.

E. van Zwet, A. Gelman, and W. 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, 2026.

Q. Wei, S. Holt, J. Yang, M. Wulfmeier, and M. van der Schaar. The ai imperative: Scaling high-quality peer review in machine learning, 2025. URL https://arxiv.org/abs/2506.08134.

W. Wu, Y. Zhao, Y. Wang, S. Li, J. Shao, Y. Long, and C. Zhang. Novbench: Evaluating large language models on academic paper novelty assessment, 2026. URL https://arxiv.org/abs/2604.11543.

Y. Wu, Z. Guo, M. Mamakos, J. Hartline, and J. Hullman. The rational agent benchmark for data visualization. IEEE transactions on visualization and computer graphics, 30(1):338–347, 2023.

S. Xi, V. Rao, J. Payan, and N. B. Shah. Flaws: A benchmark for error identification and localization in scientific papers, 2025. URL https://arxiv.org/abs/2511.21843.

H. Xu, L. Yue, C. Ouyang, Y. Liu, L. Zheng, S. Pan, S. Di, and M.-L. Zhang. Factreview: Evidence-grounded reviews with literature positioning and execution-based claim verification, 2026. URL https://arxiv.org/abs/2604.04074.

Y. Xu and L. Y. Yang. Scaling reproducibility: An ai-assisted workflow for large-scale replication and reanalysis, 2026. URL https://arxiv.org/abs/2602.16733.

Z. Xu, Y. Zhao, M. Patwardhan, L. Vig, and A. Cohan. Can llms identify critical limitations within scientific research? a systematic evaluation on ai research papers, 2025. URL https://arxiv.org/abs/2507.02694.

T. Yarkoni. The generalizability crisis. Behavioral and Brain Sciences, 45:e1, 2022.

M. Zhang, K. Tan, Y. Huang, Y. Shen, C. Ma, L. Ju, X. Zhang, Y. Wang, W. Jing, J. Deng, H. Sha, B. Hu, J. Tong, C. Jiang, Y. Geng, Y. Ying, Y. Zhang, Z. Yin, Z. Xi, S. Dou, T. Gui, Q. Zhang, and X. Huang. Opennovelty: An llm-powered agentic system for verifiable scholarly novelty assessment, 2026. URL https://arxiv.org/abs/2601.01576.

T. M. Zhang and N. 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 Decision-theoretic specification of signal-to-decision pipelines

We introduce statistical decision theory as a formalism for motivating use of particular proxies to inform review decisions.

Specifying the decision problem Let $\theta \in \Theta$ denote the latent epistemic *state* that we wish to inform on. The latent state captures the credibility of the scientific contribution along some dimension. 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.

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 $\mathcal{A} = \{\text{do not flag}, \text{flag}\}$, while papers that pass this initial filter become subject to individual reviewer scoring like $\mathcal{A} = \{1, 2, 3, 4, 5\}$, where > 3 means accept and ≤ 3 means reject, and ultimately an AC decision $\mathcal{A} = \{\text{unconditional accept}, \text{conditional accept}, \text{reject}\}$.

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 $\pi(s, \theta)$.

π 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* $U : \mathcal{A} \times \Theta \to \mathbb{R}$, which assigns a real valued score $u(a, \theta)$ to each combination of action and realized state.[1]

Optimizing the decision. One benefit of decision theoretic specification is that 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, who is informed by prior knowledge $p(\theta)$. Upon viewing a paper, this agent updates their beliefs to posterior beliefs $q_s(\theta)$:

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

They then choose the score maximizing action $a^*(s)$:

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

If the rational agent were not given access to the signals, the best they could do is to choose the best fixed action under the prior:

$$a_0^* \in \arg\max_{a \in \mathcal{A}} \mathbb{E}_{\theta \sim p}[u(a, \theta)].$$

Rational agent benchmark and baseline for decision problem. Following Wu et al. (2023), define as the *benchmark* the ex-ante expected score of the Bayesian agent who observes the signal and then chooses the score-maximizing action for each signal realization:

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

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

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

Value of the signal. The value of the signal S to the decision problem is the difference between the benchmark and baseline:

$$V_S - V_0 \tag{3}$$

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.

Using the framework. 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 and data-generating model. 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.

To inform what they propose as plausible processes, authors can benefit from empirical results from meta-analytic analyses. Below, we demonstrate a data-generating model informed by empirical research to link two proxies proposed for AI review–replicability and statistical significance–to the credibility of the direction of a claimed effect and its underlying signal-to-noise ratio.

[1] 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 $\Delta(\Theta)$.

B 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.[2]

B.1 Decision problem setup

Action, general latent state, utility function. The reviewer faces a binary action, $\mathcal{A} = \{\text{do not flag}, \text{flag}\}$. 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 Λ 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_{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.

[2] Code is available at `https://github.com/jhullman/AI_metascience_spec`.

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 Λ 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_{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 θ_{sign}, and produces equivalent results for θ_{snr}.

B.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 θ.

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

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

where h is the density of H and ϕ is the standard normal density.

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

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

For $z > 0$,

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

and for $z < 0$,

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

In either case,

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

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

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

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

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

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

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

The joint posterior is

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

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

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

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

and

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

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

B.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 $P(\theta_{snr} = 1) = 0.206 < 0.5$), 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 $P(\theta_{sign} = 1) = 0.783 > 0.5$ before any signal).

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

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

which is 0.783 for θ_{sign}, and 0.843 for θ_{snr}.

The value of the significance signal is

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

which is 0.000 for θ_{sign} and 0.049 for θ_{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 θ_{sign}, and 0.862 for θ_{snr}.

The value of replication success alone is

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

or 0.000 for θ_{sign}, and 0.068 for θ_{snr}.

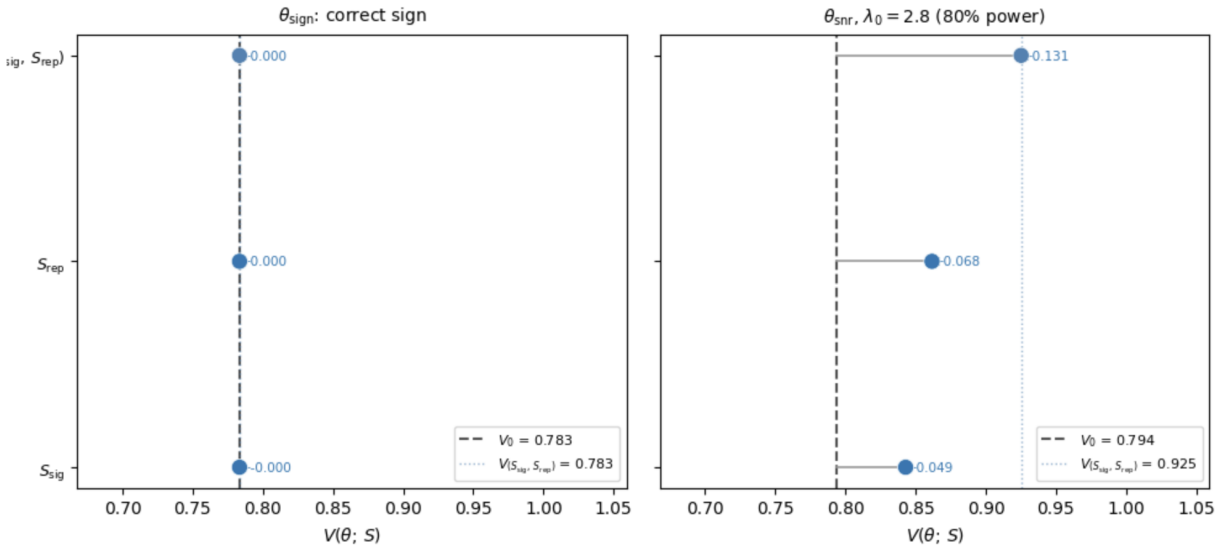

Figure 3: Expected utility of the Bayes-optimal decision rule under each signal (quantifying the ex-ante best case value of the signal), for two latent epistemic states, θ_{sign} (left) and θ_{snr} (right). Each dot shows the expected utility given the signal listed on the left. The dashed line is the 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.

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 θ_{sign}, this is 0.783. For θ_{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 θ_{sign}, and 0.131 for θ_{snr}.

Figure 3 depicts the value of all signals for the two latent states.