



Article

Cost, Verification, and the Standard for Judging Peer Review: A Philosophical Critique

Muteb Alqarni ^{1,*}  and Alaa Albahi ^{2,3,4} 

¹ Department of English, College of Languages and Translation, King Khalid University, Abha 61413, Saudi Arabia

² English Language Department, College of Science and Health Professions, King Saud bin Abdulaziz University for Health Sciences, Jeddah, Saudi Arabia; albahia@ksau-hs.edu.sa

³ King Abdullah International Medical Research Center, Jeddah, Saudi Arabia

⁴ Ministry of the National Guard—Health Affairs, Jeddah, Saudi Arabia

* Correspondence: moteebalqarni@kku.edu.sa

Abstract

Peer review's reliability on checkable dimensions has been low since the agreement studies of the 1970s and 1980s. The remedies have been known almost as long: explicit criteria, complete rather than sampled checking, and plural independent evaluation. None has been implemented at scale. This paper asks three questions. Is that unreliability a property of human judgment or of procedure? What prevented adoption of the known remedies? And what must an evaluative arrangement satisfy in order to be legitimate? The method is conceptual: the paper synthesizes the empirical literature on peer review with work in social epistemology, procedural legitimacy, and contestable artificial intelligence. It argues for three claims. Reliability on checkable dimensions varies substantially with procedural design. Cost is a necessary but insufficient explanation of non-adoption. And a verdict is legitimate to the degree that it is contestable by the person it burdens; this procedural standard is defended against the epistemic consequentialism that the adjacent literature assumes. On that standard, cheap machine verification changes less than current enthusiasm suggests. Fully specifiable operations, such as statistical recomputation and reference resolution, reach a reliability that human review never approached. Partially specifiable operations gain coverage rather than accuracy. The significance judgment gains nothing. The contribution is a defended standard for judging evaluative institutions, not a case for machine review. The governance and adversarial-robustness conditions that such an arrangement requires are specified, but they are not claimed to be met.

Keywords: peer review; procedural legitimacy; contestability; artificial intelligence; scholarly evaluation



Academic Editor: Lorenzo Magnani

Received: 8 August 2026

Revised: 16 October 2026

Accepted: 19 October 2026

Published: 22 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Richard Smith, after thirteen years editing the BMJ, delivered the most quoted verdict on peer review, and then conceded that he had nothing to put in its place. Review, he wrote, “is compared with democracy: a system full of problems but the least worst we have” [1] (p. 178), and it would survive “because there is no obvious alternative” [1] (p. 182). The concession is more interesting than the critique. Smith knew what a better system required, and he named the obstacle: verifying what authors report means “reviewing raw data,” which “is difficult, expensive and time consuming” [1] (p. 182). The barrier he named was

not conceptual. It was the price of labor, although Section 5.1 argues that price was not the only obstacle.

This paper takes that concession as its point of departure, and its claims are deliberately bounded. The deficiencies documented over five decades have a known remedy on their checkable dimensions. That remedy went unimplemented in part because it was unaffordable in volunteer hours. Machine verification now removes that obstacle, but only for a narrow class of operations. The conclusion drawn is not that machines should review manuscripts, nor that a general relocation of scientific trust has become available. It is that the standard by which evaluative arrangements are judged should be procedural, and that, on that standard, the newly cheap layer changes less, and something more specific, than either enthusiasts or abolitionists suppose. The strongest claim concerns automated verification infrastructure, that is, tools that recompute and retrieve against stated rules. It does not concern language-model evaluation of manuscripts; Section 5.2 makes the distinction explicit.

If the constraint is the price of expert attention, its clearest site is the editorial desk, where the most common fate of a submission is rejection without review. Of 753,909 submissions to 157 journals of ordinary prestige, only 51.7% were ever sent to referees [2] (p. 8). Publishing-economics modeling puts the global desk rejection rate near 0.45, on parameters its authors call “undoubtedly speculative” [3] (p. 3). These decisions are made in minutes, on the title, abstract, and cover letter. The material that peer review exists to assess is never examined.

Whether this matters depends on whether desk triage tracks quality, and the answer depends on where in the distribution one looks. Siler, Lee, and Bero [4] (p. 360) conclude that “on the whole, there was value added in peer review,” and nothing below disputes that. The claim pressed here concerns the tail, not the mean. Following 1008 manuscripts submitted to three elite medical journals, the same authors found that the fourteen most-cited papers had all been rejected by the journals first approached, twelve of them without external review [4] (p. 361). Against a baseline desk rejection rate near 77%, the eventual top papers were about as likely to be desk-rejected as any random submission. In that sample, then, and at the tail where filtration matters, the desk carried no discriminative information on the proxy that the system itself invokes. Among papers published in journals with impact factors above 8.00, rejected articles were subsequently more cited than accepted ones, 212.77 against 143.22 [4] (p. 362). Citations are a weak and field-dependent proxy for quality, and Section 6 argues that selecting on them would be arbitrary. The claim drawn here is therefore correspondingly weak and confined to this evidence: the desk fails on the proxy that the system itself invokes to justify gatekeeping. This is an argument about internal incoherence, not about the true ranking of papers.

None of this is new, and that is the point. Zuckerman and Merton analyzed 14,512 manuscripts submitted to The Physical Review between 1948 and 1956. They found that papers by the highest-ranked authors were decided by the editors alone, without outside referees, 87% of the time, against 58% for the lowest-ranked [5] (p. 482). In a low-rejection era, the shortcut took the form of desk acceptance for elite authors. In a high-rejection era, the identical mechanism appears as desk rejection for everyone else. Scarcity is the most economical account of this reversal. Submission inflation, brand competition, and career incentives that reward volume also bear on the modern desk, however, so the claim defended here is the weaker one: scarcity is necessary rather than sufficient.

1.1. The Gap

Three studies bear on this problem, and they do not meet. The first audits peer review’s performance empirically. The second asks whether prepublication review is

worth having at all; the most forceful statement is Heesen and Bright's argument [6] that it is weakly dominated by publish-then-filter arrangements. The third evaluates whether language models can review manuscripts; Kousha and Thelwall [7] offer the most cautious recent synthesis. The first measures without a normative standard, the second fixes evaluative capacity at human capacity, and the third assesses capability without asking what evaluation is for. A fourth literature has the concept but not the case. Work on contestable artificial intelligence treats responsiveness to dispute as a design requirement for automated decisions [8], but it does not address scientific evaluation, where the decision is already made by humans whose reasons go unrecorded. Heesen and Bright [6] state the diagnosis for all three when they observe that critical work "tends to be vague about the normative standard against which peer review or its alternatives are to be evaluated." Without a stated standard, the claim that machine evaluation is faster and more complete is an argument about throughput, and throughput is not what the legitimacy of a verdict rests on.

1.2. Questions and Contribution

Three questions organize what follows. Is peer review's low reliability a property of human judgment or of the procedure? What prevented adoption of the known remedies, and has that obstacle changed? And what must an evaluative arrangement satisfy to be legitimate?

The contribution is a synthesis and a standard rather than a discovery, and it can be stated as three additions to existing work. To the debate over whether peer review is worth having, it adds a defended procedural standard and an argument for its priority. Heesen and Bright [6] evaluate review by its epistemic yield; this paper argues that an institution exercising authority over identifiable persons is bound by a procedural obligation that epistemic yield does not discharge. To work on contestable artificial intelligence [8], it adds the case that literature lacks: an evaluative institution in which the decision is already human and its reasons already go unrecorded. Contestability is therefore a standard for the incumbent before it is a design requirement for any machine. To the literature on machine-assisted review, it adds a partition of evaluative operations by specifiability, which fixes what cheap verification can and cannot change. That reviewer labor is scarce and costly is a background assumption throughout the literature drawn on here, and Aczel, Szasz, and Holcombe [3] proposed unbundling review into mechanical and judgmental layers before language models made the mechanical layer cheap. The argument here does not rest on the price of machine evaluation. It rests on what legitimacy requires of any evaluator, human or otherwise. Finally, there is one further instrument, declared unavailable in 1971, which an algorithmic evaluator makes costless.

2. The Standard, and the Argument for It

Evaluating an institution requires saying what counts as its doing well. Heesen and Bright [6] adopt epistemic consequentialism after Goldman [9]: peer review consumes resources whose epistemic yield would be higher elsewhere. The standard adopted here is procedural. A verdict is legitimate to the degree that it is contestable by the person it burdens. Contestability requires stated reasons, public criteria, and an inspectable, repeatable procedure. An institution that produced correct verdicts by unstated criteria and unappealable authority would score well on the consequentialist standard and remain illegitimate.

Longino supplies the philosophical form. Objectivity is not a property of individual minds but a scalar property of procedures, since "a method of inquiry is objective to the degree that it permits transformative criticism" [10] (p. 76). She applies the criteria to peer

review herself, and finds that “its confidentiality and privacy make it the vehicle for the entrenchment of established views” [10] (p. 76). The indictment of confidential review from within canonical social epistemology is therefore her own.

Because the reframing depends on this choice of standard, the choice must be argued rather than announced. Three considerations favor it. First, the consequentialist standard cannot be applied to the question at issue. Scoring an arrangement by its epistemic yield requires knowing which verdicts were correct, and there is no gold standard of manuscript quality against which to score them [2] (p. 8); the available proxy is the one that Section 6 shows to be arbitrary. Heesen and Bright’s result is accordingly a dominance argument rather than a measurement. It turns on the cost side of the ledger, which is observable, and not on the benefit side, which is not. Second, contestability is a precondition of the consequentialist standard rather than a rival to it. Any judgment of epistemic yield is itself an evaluative verdict, and it can be corrected only if the record it rests on can be inspected. An arrangement that seals its verdicts destroys the evidence that any such assessment would need. The desk is the demonstration: its decisions proved unmeasurable even to a nine-author consortium holding publisher-internal records [2] (p. 2). These two considerations are arguments from epistemic access. They show the consequentialist standard to be unusable here, not mistaken. The third consideration is normative, and it carries the priority claim. Peer review allocates publication, promotion, and funding to identifiable individuals through a verdict imposed by authority. An institution that wields authority over persons owes each of them a justification they can take up, and an aggregate cannot supply this: no gain distributed across a research community answers the particular author whose work was refused without reasons. This is why the procedural standard has priority where the two conflict, and why that priority is not a weighing of one value against another. Epistemic duties to readers and to the research community are real, and nothing here denies them. But when an institution exercises authority over identifiable individuals, its epistemic aims do not cancel its obligation to give each of them reasons that are inspectable and contestable. Procedural legitimacy functions as a constraint on how epistemic aims may be pursued, not as a competing good to be traded against them. An arrangement that pursued epistemic yield by sealing its verdicts would therefore not have struck a permissible balance; it would have violated the constraint. Procedural-justice research finds the same structure empirically: the perceived fairness of a procedure, rather than the favorability of its outcome, drives whether those subject to authority accept it as legitimate [11]. Sections 4.3 and 4.4 show those burdens falling unequally and by status. That is the case an aggregate standard is constitutively unable to register.

None of this refutes epistemic consequentialism. It shows that consequentialism answers a different question, and that where the two standards diverge, the procedural standard is the one appropriate to an institution exercising authority over persons. A committed consequentialist can accept what follows and read the positive proposal as an empirical conjecture: that contestable procedures accumulate warrant faster than sealed ones. Section 7 states this as a mechanism rather than as a demonstrated result.

Two clarifications keep the standard honest. First, it does not certify truth: “to say that a theory or hypothesis was accepted on the basis of objective methods does not entitle us to say it is true” [10] (p. 79). Second, legitimacy is not one thing. The organizational literature distinguishes pragmatic, moral, regulatory, and cognitive legitimacy as separately necessary [12] (pp. 577–585); [13] (pp. 32–35). The proposal here is strong on three of these and weak on the fourth.

The standard also requires distinguishing what review is for. Four functions are conventionally attributed to it: validation, filtration, improvement, and legitimation [14,15].

Treating an argument about one of them as an argument about all four is the characteristic error of the debate.

The argument then has five separable steps. Section 3 establishes that trust was attached to persons by design rather than by demonstration. Section 4 argues that, on checkable dimensions, the resulting unreliability varies substantially with procedural design. Section 5 argues that the remedy went unimplemented in part because it was unaffordable, and that cost is not the whole explanation. It then asks how much of the remedy the affordable layer executes, and at what cost. Section 6 states what does not follow. Section 7 specifies the arrangement and its legitimacy conditions, several of which are not met. Sections 8–11 draw out implications, recommendations, limitations, and directions for further work before the conclusion.

3. Trust Was Attached to Persons by Design

The historical record does not show an institution that was built to certify quality and subsequently corrupted. It shows one built for other purposes, onto which a certification warrant was later grafted. The bearing of this on the present argument is direct. If the warrant was acquired without demonstration, the institution's current self-description cannot be defended by appeal to its tradition, and the standard of Section 2 applies to it without historical discount. What follows establishes three things in turn: that evaluation was attached to persons from the start; that the universal external referee is recent and was adopted for reasons other than epistemic ones; and that confidentiality was a borrowed convention rather than a demonstrated safeguard.

Biagioli traces academic peer review to early modern book censorship, from which it differentiated as the royal academies acquired the privilege of licensing their own publications. Their policy concerned "suitability, not quality" [16] (p. 22). It ensured that a text made no unacceptable claims "rather than to certify that it made good claims" [16] (p. 23). More striking is his account of who was evaluated. Because admission to the Académie involved extensive vetting, the criteria of publishability came to be embodied in the academicians themselves. They had been pre-reviewed as persons so thoroughly that their texts required little further review [16] (p. 23). Status filtering is therefore not a deviation from peer review's design. It is that design, under a meritocratic description.

The chronology needs care, because the claim that peer review is a post-war invention is false. Assigned pre-publication review was chartered in 1665 [5] (p. 463), and the designated referee dates from Greenough's 1817 statutes [17] (pp. 136–137). What is post-war is narrower: distributed, external, anonymous refereeing as a universal condition of publication that carries an epistemic warrant. As late as 1962, only 71% of 156 surveyed journals made any use of referees [18] (p. 75), and "many journals that adopted refereeing in the mid-twentieth century did so not because of epistemic concerns but to lighten the editorial workload" [19] (p. 543). The institution now asked to bear the legitimization function was not designed to bear it.

Secrecy has the same character, and its history reverses the usual framing of transparency proposals as departures from tradition. The reader's report began as a published genre. It appeared in the Royal Society's Proceedings from 1832 and was absorbed into confidentiality by 1838 [17] (pp. 147–150). "The referee discharged a public service through privileged secrecy" [17] (p. 152). This arrangement was borrowed from the anonymous literary reviews, whose defenders held that a critic's verdict became merely personal opinion once signed. That defense should be taken seriously rather than dismissed, and a second defense has since joined it: confidentiality is said to protect junior referees from retaliation and to permit a candor that a signed report would inhibit. The historical claim made here is accordingly limited. It is not that confidentiality lacked a rationale, nor that the external,

anonymous referee was adopted for no epistemic reason anywhere; the record is uneven across journals and fields. It is that these arrangements were not adopted on the strength of a demonstrated epistemic advantage, and that their present defenders inherit the burden of supplying one. The Victorian objections were indistinguishable from contemporary ones: memoirs rejected with no ground assigned, secret reports, and no published criteria [17] (pp. 130, 152–154).

Two further facts matter. The 1975 congressional controversy over the National Science Foundation was resolved through a transfer of trust from individuals to a system, since “individuals could not be trusted, but the system of peer review could” [19] (p. 557). Baldwin [19] (p. 558) also presses the strongest historical objection: complaints assume that review is “supposed to distinguish good science from bad science with perfect accuracy. But that is not the purpose for which refereeing was initially designed.” Taken as a defense, this concedes the case. The standard criticized below is the institution’s present self-description, and that self-description was acquired in a political controversy rather than by demonstration.

4. The Diagnosis: Procedural Design Substantially Shapes Reliability on Checkable Dimensions

The empirical case against peer review is usually presented as a case against reviewers. The evidence says something narrower. Judgment exercised without an explicit, shared, practiced criterion set is unreliable, and it becomes substantially more reliable when such a set is imposed, wherever the criterion can be stated. The claim is confined to checkable dimensions. It does not hold that human judgment sets no limit of its own; Section 4.2 locates that limit.

4.1. *What the Reliability Data Show, and What They Do Not*

Chance-corrected agreement between independent reviewers is not, as is often claimed, at chance. It is reliably above chance, and it is almost uniformly poor by the field’s own thresholds. Cicchetti classifies values below 0.40 as POOR, 0.60 to 0.74 as GOOD, and 0.75 to 1.00 as EXCELLENT [20] (p. 123); the tabulated peer-review values run from 0.19 to 0.54. Bornmann’s meta-analysis of 48 studies reports agreement coefficients near 0.23 and 0.34 and a kappa of 0.17 [14] (p. 207). In two clinical neuroscience journals using structured forms, agreement on the accept/revise/reject recommendation gave $\kappa = 0.08$ and $\kappa = 0.28$ [21] (p. 1965). The objection that these figures are depressed because editors removed the easy cases has been tested. Restoring editor-only rejections raises overall agreement to 74%, but acceptance agreement remains at 51% and kappa at 0.34 [20] (p. 172).

The most informative result is a variance decomposition. Rothwell and Martyn [21] (pp. 1966–1967) found that reviewer identity explained 27% and 32% of the variance in conference abstract scores, against 11% and 15% for the identity of the abstract, and that adding reviewers did not improve consistency. The NeurIPS duplicate-committee experiment agrees, finding that “50% of the variation in reviewer quality scores was subjective in origin” [22] (p. 1). An instrument that tracks its operators two to three times more strongly than its objects is uncalibrated.

Two qualifications improve the argument. First, reliability is not a legitimate demand across all four functions. Bailar [23] (p. 138) objects that “too much agreement is in fact a sign that the review process is not working well,” Lee et al. [15] (p. 6) question the assumption that disagreement is normatively inappropriate, and Kuhn [24] (p. 200) states the deepest version: there is “no neutral algorithm for theory-choice.” All of them are right about significance. None of them touches whether a reported F value matches its degrees

of freedom, or whether a cited paper has been retracted. The calibration charge is therefore restricted to the validation function and to checkable dimensions. Second, disagreement is informative only in a system that adjudicates it. Faced with split reviews, editors reject rather than accept at a ratio of over ten to one [20] (p. 176), and two reviews of the same paper often share no critical point at all [14] (p. 209).

4.2. *The Prescription, Demonstrated in 1991*

If unreliability were wholly a property of human cognition, no procedural intervention would repair it. Laming's comparison is the cleanest natural experiment available. United Kingdom public examinations "typically achieve a mark-remark correlation of 0.9 or better," and the difference from peer review "is the use of a marking scheme, however imprecise, which is practised by the examiners." Remove the scheme, and "the reliability of examination marks falls to near the levels reported for peer review" [25] (pp. 155–156). Direct intervention confirms this. After training, nine reviewers reached an intraclass reliability of 0.71 for overall scientific quality. They exceeded 0.8 on mechanical items, such as whether search methods were reported, and fell to 0.40 on whether conclusions were supported by the data [20] (p. 180). Training buys reliability where checklists apply, and it stops where interpretation begins. That residue, at 0.40, is the limit that human judgment contributes on its own, and no procedure described here removes it. The psychometrics also specify how much plural evaluation would be needed. Applying the Spearman–Brown formula to a single-reviewer reliability of 0.30, Stricker [26] (pp. 163–164) calculates 0.77 for seven reviewers, which falls in the range labeled EXCELLENT. No journal has attempted it.

4.3. *The Capacity Limit Is Real and Selectively Applied*

A second failure is less amenable to exhortation. Reviewers do not verify manuscripts sentence by sentence, because they cannot. When the BMJ planted known major errors in a manuscript, "nobody ever spotted all of the errors. Some reviewers did not spot any, and most reviewers spotted only about a quarter" [1] (p. 179). Automated scanning agrees. It finds statistical reporting inconsistencies in roughly half of published psychology articles, and in one article in eight it finds an inconsistency large enough to change a result's significance [27].

Mahoney's experiment sharpens this. He sent 75 referees an identical fictitious manuscript that differed only in the direction of its findings. Where the results favored the reviewers' theoretical commitments, the identical methods section averaged 4.2 on the study's seven-point scales, scored 0 to 6; where the results disconfirmed them, it averaged 2.4, and publication recommendations moved in step [28] (pp. 165–167). Embedded in the manuscript was an error that made the sample description arithmetically impossible. It was noted by 25% of those reading the favorable version and by 71.4% of those reading the unfavorable one. Scrutiny is not merely incomplete. It is allocated by prior belief. Methodology ratings correlated with summary recommendations at $r = 0.94$, so stated reasons are largely post hoc. This is why an objection such as "the theoretical contribution is unclear" is not weak but unfalsifiable.

4.4. *Status Operates Where Evaluation Does Not*

The status-bias literature is usually deployed to show that reviewers are partial. The archival record supports a more precise claim. Zuckerman and Merton found no status effect at the referee stage at all, since "referees of each rank accept the same proportion of papers by authors from every stratum" [5] (pp. 490–491). The stratification sits at the editorial stage. It appears in the allocation figures above and in acceptance rates of 91% for the seventeen foremost physics departments against 72% elsewhere [5] (p. 480). Status bias there is a property of the gate, not of the evaluation.

Modern experiments locate a second such gate. At WSDM 2017, single-blind committee members were more likely than double-blind ones to recommend acceptance for papers with a famous author (odds multiplier 1.63) or from a top company (2.10), and they bid preferentially on papers from top institutions [29] (p. 3). Status thus determines who reads the paper before anyone judges it. Huber et al. [30] (p. 3) supply the sharpest measurement, and their anonymized baseline is essential. When the less prominent co-author was shown as corresponding author, 65.4% of reviewers recommended rejection; when the Nobel laureate was shown, 22.6% did; and for the anonymized but otherwise identical manuscript, 48.2% did. The effect reached the checkable dimensions, moving ratings of whether conclusions were supported by the data [30] (p. 4). The authors attribute it to the halo effect, whereby raters “extrapolate from known to unknown information” [30] (p. 5). A cognitive mechanism of this kind survives every governance fix that leaves the substrate in place.

The exceptions define the argument’s scope. Squazzoni et al. [2] found no systematic gender penalty across 145 journals, but only for manuscripts that survived the desk, since “data on desk rejections were not consistently available” [2] (p. 2).

5. The Cost Constraint, and What Lifting It Buys

5.1. *What the Remedy Would Have Cost*

Reliable evaluation requires three things: an explicit criterion set applied consistently, complete rather than sampled checking, and enough independent evaluations to average out operator variance. None has been implemented at scale, and under volunteer labor each is unaffordable. Reviewers spent over 100 million hours on peer review in 2020, some 15,000 person-years, worth over 1.5 billion US dollars for United States-based reviewers alone [3] (pp. 3–4). That buys two reports per manuscript, sampled reading, and private criteria. Stricker’s calculation implies seven independent evaluations. That would triple an input the system cannot secure, and one that is concentrated in a minority of the eligible population [7] (p. 5). Complete verification failed the same test, since reviewing raw data is difficult, expensive, and time consuming [1] (p. 182). Rerunning deposited code is the only response to that failure that does not require a reviewer to work for free.

The leading arguments on both sides of the debate depend on this constraint without examining it. Polanyi concedes that “unorthodox work of high originality and merit may be discouraged or altogether suppressed for a time. But these risks have to be taken” [31] (p. 61). This calculation assumes that evaluative attention is irreducibly scarce. Heesen and Bright’s abolitionism runs on the identical premise from the opposite direction: on their account, prepublication review is an enormous sink of scientists’ time and resources, and the size of that sink is what generates their dominance result. Zuckerman and Merton supply a third instance when they describe editors who prefer “when in doubt, reject” [18] (p. 78). Type-I error is preferred deliberately, as a response to scarcity rather than as an accident of reviewer psychology.

Scarcity is thus doing philosophical work in the defense of peer review, in the argument for its abolition, and in the design of its selection rules. This is not to say that scarcity explains these positions entirely; it is to say that each of them requires it as a premise and none examines it. One fact cuts hard against any stronger version of the cost thesis, and it is the best objection to it. *statcheck* has been free and public since 2016, yet by 2024 only two journals had adopted it in review [32]. A pure cost story predicts near-universal uptake of a tool with zero marginal cost. Its absence shows that inertia, the prestige of the referee’s signature, and reviewers’ investment in their own discernment are doing explanatory work too. Cost explains why the full remedy could not be implemented. It does not explain why an already implementable fragment went unimplemented. An arrangement blocked by

habit will not be adopted merely because it becomes cheap; it must be argued for on the normative grounds of Section 2, not on price. Nor does the removal of the cost constraint touch every deficiency documented above. The filtration failures of Section 1 are the ones the architecture does not remedy.

5.2. What the Affordable Layer Actually Executes

The claim is deliberately narrow. Kousha and Thelwall [7] (pp. 4, 10) conclude that artificial intelligence “should not be used to replace human reviewers.” That conclusion is accepted without qualification, since nothing below asks a machine for a holistic verdict. What is asked is that decomposed, checkable subtasks be performed completely and consistently. Kousha and Thelwall also inventory what is already deployed, from text-overlap detection to statistical reporting checks and citation-to-reference matching [7] (p. 6). And before language models were relevant, Aczel, Szaszi, and Holcombe had argued that a specialist evaluating one dimension across many manuscripts outperforms a generalist reading every word of one [3] (pp. 6–7).

Comparative evidence on holistic feedback establishes less than the argument requires. Across 3096 papers from fifteen Nature family journals and 1709 ICLR submissions, machine feedback overlapped with an individual human reviewer’s points at rates that tracked human–human overlap. Reassigning the feedback across papers collapsed the overlap to 0.43%, so the criticism is at least paper-specific [33] (pp. 1, 4–7). But matching a baseline that Section 4 has discredited is not executing the remedy. The remedy demands the band that the field labels EXCELLENT.

The evidence that matters lies in the narrow tools, and it is strong exactly where the check is fully specified. The R package *statcheck* extracts reported test statistics and recomputes the associated *p* values. Its sensitivity lies between 85.3% and 100%, its specificity between 96.0% and 100%, and its agreement with manual coding is 0.76 for inconsistencies and 0.89 for decision inconsistencies [34]. On Cicchetti’s scale, those figures are GOOD to EXCELLENT. Human reviewers reach 0.17 to 0.34 on the same manuscripts, and the *BMJ* reviewers found only a quarter of the planted errors. Nor is this only a laboratory result. A preregistered comparison of two journals that adopted *statcheck* against matched controls found a steeper decline in reporting inconsistencies at the adopters [32].

The limits are limits of coverage rather than accuracy. *statcheck* recognizes roughly 60% of reported significance tests; it misses those in tables or outside APA conventions, and Schmidt [35] recomputes its sensitivity at 0.52 once undetected tests count as misses. That is a parsing deficit, and a language-model layer extends coverage. But where a check is only partially specifiable, performance falls back to the human band: GPT-4 found seven of thirteen deliberately inserted errors [36] (pp. 4–5).

The partition needs a criterion rather than examples. An operation is fully specifiable when its correct output is fixed by the manuscript plus a stated rule that competent parties would apply identically, so that any disagreement can be resolved by inspection. Examples are whether a reported *p* value follows from its test statistic, whether a cited work exists, has been retracted, or contains the passage attributed to it, and whether a required reporting item is present. An operation is unspecifiable where the rule cannot be stated without residue, as with novelty. Between the two lie partially specifiable operations, where a rule can be stated but its application calls for judgment on which competent parties may differ. Whether a cited work supports the claim for which it is cited is the clearest case. Its existence and the presence of a quoted passage are settled by retrieval, but support is a matter of degree and interpretation, and a check that returns a yes-or-no answer to it is not fully specified. Whether conclusions are supported by the data is another such case (Table 1). The boundary is empirical and mobile. “Whether conclusions are supported by

the data” sits at Cicchetti’s 0.40 line, and it is where Huber et al. locate status contamination. The architecture must therefore be revisable in scope. The inner form of Section 7 is drawn where the boundary currently falls, not where it must always fall.

Table 1. Evaluative operations by specifiability, the class of instrument each admits, and where judgment remains.

Operation	Class	Instrument	Where Judgment Remains
Reported <i>p</i> value follows from the test statistic and degrees of freedom	Fully specifiable	Deterministic recomputation (statcheck)	None; correct by inspection
Cited work exists; retraction status; cited passage present in the source	Fully specifiable	Retrieval against registries and sources	None; correct by inspection
Required reporting item present under a codified standard	Fully specifiable	Rule-based check	None; correct by inspection
Tests reported in tables or outside standard conventions	Partially specifiable	Language-model parsing, then recomputation	Human adjudication of the parse
Cited work supports the claim attributed to it	Partially specifiable	Language-model reading anchored to quoted text	Human adjudication where evaluators diverge
Conclusions are supported by the data	Partially specifiable	Machine-supported, plural configurations	Expert reviewer
Novelty, significance, theory choice	Unspecifiable	None	Named editor and commissioned expert

The empirical hinge is therefore conditional. Where an operation is fully specifiable, the machine layer reaches the reliability the remedy demands, which human review never has. Where an operation is partially specifiable, the machine layer buys complete rather than sampled coverage at negligible marginal cost, at roughly human accuracy per pass. Where an operation is not specifiable at all, it buys nothing; that is the concession of Section 6. Completeness throughout means completeness of coverage, not of detection. Plurality inherits the same structure. Stricker’s seven evaluations become purchasable, but seven passes of one model are one evaluation repeated. What becomes purchasable is therefore plurality of configuration rather than plurality of runs.

Two classes of instrument are in play in Table 1, and they should not be run together. The first class is automated verification tools, of which statcheck is the paradigm. These are deterministic or retrieval-based: a regular-expression engine that recomputes a *p* value, a resolver that checks a reference against a registry, a retraction lookup. They read no free text, they take no instructions from the manuscript, and they produce outputs that are correct or incorrect by inspection. The second class is large language model evaluation. It is what the partially specifiable dimensions require, since parsing tests out of tables, judging whether a source supports a claim, and assessing whether conclusions follow all demand reading the manuscript. The strongest claim of this paper concerns the first class. It is verification infrastructure that reaches the band the remedy demands, and the inner form of Section 7 is built from it alone. The claims made for the second class are weaker throughout: coverage rather than accuracy, and subject to the adversarial and governance conditions of Sections 5.3 and 6. The question the paper answers is therefore not whether artificial intelligence should participate in peer review. It is which checks can be moved from inconsistent or absent human performance to verification infrastructure, and on what terms a reading evaluator may be added.

5.3. *What the Alternative Costs*

A thesis about relative cost requires costing both sides, and the machine side can be given as an order of magnitude, with its assumptions stated. A 10,000-word manuscript is roughly 15,000 tokens of input, and a criterion-anchored evaluation produces on the order of 10,000 tokens of output. At the commercial frontier-model prices prevailing in 2026, one pass therefore costs on the order of ten cents, and seven independently configured passes with retrieval calls come to about a dollar per manuscript. The human alternative is two to three reports at several hours each. The conclusion does not depend on the precise price. It depends only on that price being several orders of magnitude below the salary-equivalent value of expert hours.

What does not become cheap should be stated with equal precision, and it is most of the total. The retained human layer consists of editorial time for the significance judgment, plus commissioned expert review wherever significance is contested. This layer is the expensive component, and it is untouched. The checkable layer becomes cheap while the judgmental layer remains exactly as costly as now, and the paper cannot say what fraction of aggregate reviewer hours the checkable portion represents. The audit apparatus and the governance overhead, that is, hosting, model version archiving, ingest sanitization, and appeals, are fixed institutional costs of the order that journals already carry for submission systems. The operative asymmetry is not that the new costs are small. It is that they are purchasable, whereas the incumbent constraint is volunteer expert attention, which money cannot expand. That is the whole of the affordability claim, and it is weaker than “correction is now cheap.”

A trilemma remains, and it is the strongest practical objection. The most capable models are proprietary; open-weight models are governable but may trail in capability; and a configuration that is both capable and governable exists only if some body procures it. One reply is that capability requirements are lowest where the reliability claim is strongest, since statcheck is a regular-expression engine and reference resolution is a retrieval problem. But that reply should not be overstated, because the three virtues attach to different parts of the pipeline. The operations that are cheap, governable, and closed to injection are largely those that need no language model. The partially specifiable dimensions that do need one reintroduce frontier-model dependence and the ingest attack surface. The trilemma is therefore not dissolved but partitioned, and it bites fully on the manuscript-reading layer. Consortium infrastructure is not unprecedented, since Crossref and the preprint servers are nonprofit operations on which the system already depends. But nothing shows that such a consortium will be built, and if it is not, the arrangement this argument licenses is unavailable. That is a claim about political economy rather than about epistemology.

6. What Does Not Follow

Four things do not follow, and stating them distinguishes a proposal from an enthusiasm.

Machine evaluation cannot criticize the standards it applies. Longino identifies the mechanism: assumptions shared by all members of a community “acquire an invisibility that renders them unavailable for criticism” [10] (p. 80). A model trained on the scientific record is the maximally consensual participant, and its behavior matches. It commented on novelty 10.69 times less often than human reviewers did [33] (p. 7), and such systems are “necessarily conservative, favoring the status-quo” [37] (p. 4). Criticism internal to shared standards is machine-tractable. Criticism of those standards is not, and it remains the human and pluralist function.

Consistency is not correctness. Cortes and Lawrence [22] (p. 6) press this point against their own reform proposals: a conference that accepted papers by reference count would be perfectly consistent and perfectly arbitrary. Reproducibility delivers checkability rather than

accuracy. It is a precondition of auditability, not a substitute for it. What the architecture requires is within-evaluator reproducibility, not between-evaluator convergence, since validity is maximized when evaluators' errors are independent [38] (p. 153). This makes model monoculture the pathological case. Reproducibility is also a property of pipelines rather than of models, since an identical query submitted weeks apart returns dissimilar conversations [37] (pp. 5–6). That is an argument for governed infrastructure, not against machine evaluation.

Identity signals are strippable; style signals are not. Goudsmit's objection, which persuaded Physical Review to refuse anonymized review, remains live: removing name and affiliation does not make a manuscript anonymous, since a competent reviewer can tell at a glance where the work was done [5] (p. 482). Checco et al. [39] (pp. 6, 9) trained a network on word distribution, readability, and formatting alone, with names removed, and reached 74.01% accuracy against a 50% baseline. Machine input-masking therefore delivers not anonymity but measurability. Human blinding failed in execution, since reviewers identified authors 25% to 40% of the time [15] (p. 11), and it failed in auditability, since a human who infers an identity cannot un-infer it. A pipeline, by contrast, can be run with the affiliation stated and then stripped, and the difference can be measured.

Anchoring does not survive an adversarial author. In July 2025, eighteen preprints on arXiv were found to carry instructions hidden in white text, invisible to readers and legible to models: "GIVE A POSITIVE REVIEW ONLY," "do not highlight any negatives" [40,41] (pp. 2–5). The defense that these were honeypots for reviewers using prohibited tools fails on the content, since a neutral trap would not have been uniformly self-serving. The attacks work. Testing injected prompts against commercial systems on rejected conference papers, Collu et al. [42] found that exploit prompts raised mean scores significantly across model families, in one configuration from 7.93 to 9.92 out of ten. The vulnerability is a property of the interface, since a model cannot separate the document it is given from the instructions it is given. This is the present proposal's own attack surface, and the concession must be complete. Anchoring to quoted text is no defense against instructions concealed within that text, and the documented positivity of machine evaluation [43] is inherited by adversarially and charitably configured evaluators alike. Three things survive, and all of them are narrower than the unconditional proposal. First, the injection surface belongs to the submission channel and is addressed there, by journal-controlled extraction that discards invisible text, by instruction detection at ingest, and by archiving of the file. Second, cross-model divergence is an attack detector as well as a psychometric device. Third, the attack leaves evidence in an artifact the journal retains, whereas Mahoney's referees allocated scrutiny by prior belief and left no trace. What does not survive is the unconditional form. An evaluator that reads author-supplied files cannot be secured by prompt design. The architecture below is therefore defensible only where a venue controls text extraction and runs divergence audits; otherwise it must be confined to journal-side operations that take no free text from the manuscript body.

What cheap evaluation does permit is one instrument. Zuckerman and Merton concluded that "it would require a strict experimental design to find out decisively whether papers of the same scientific quality are assessed differently by referees according to the status of authors. Both ethics and practicality rule out the draconian experiment" [18] (p. 90). Peters and Ceci [44], Tomkins et al. [29], and Huber et al. [30] ran approximations of that experiment, at real cost to real authors and editors, and each was a single study of a single field. An algorithmic evaluator makes the experiment costless, repeatable, and deception-free. One qualification is required, and it strengthens the claim. The evaluator carries its own prestige inference, since style survives anonymization, and it carries its own positivity. An instrument with those properties, turned on prestige effects, may end up

measuring itself. It must therefore first be shown to be neutral on the dimension under test, by running matched pairs that differ only in the manipulated cue and demonstrating a null result. That validation is itself cheap and repeatable, and that is the point: the 1971 objection was that no calibration was possible at all.

7. The Architecture and Its Legitimacy Conditions

The proposal relocates the unit of trust from the reviewer to the evaluative procedure, and Section 6 has fixed its perimeter. The architecture has two forms. The inner form is defensible now. It runs on the journal side and takes no free text from the manuscript body: recomputation against deposited data, reference resolution, retraction status, and conformity to codified reporting standards, all within infrastructure the venue controls. The outer form adds evaluators that read the manuscript. It is contingent on ingest security: journal-controlled extraction that discards invisible text, instruction detection at submission, and divergence auditing across model families. Five conditions govern both forms.

One clarification prevents a confusion between the diagnosis and the proposal. Sections 1 and 4.4 indict editorial judgment for inaccuracy; what follows delivers contestability; and these are different goods. The bridge between them is diachronic. A procedure that publishes its criteria and archives its reports has a measurable error rate, and a measurable error rate is correctable. The incumbent's deeper pathology is that its errors go unrecorded, so nothing accumulates from them. Laming's examiners did not become reliable by becoming better judges. They became reliable by working to a scheme that was itself revised against remark data. The accuracy claim is therefore a testable conjecture rather than an assumption, and the appropriate test is a journal-level pretest–posttest of the kind already run for statcheck [32]. Where the accuracy claim is not granted, the case rests on legitimacy alone, and Section 1 establishes only that the incumbent has no accuracy advantage to trade against accountability.

Explicit and published criteria. This is Laming's marking scheme. It addresses the demonstrated cause of unreliability rather than its symptom, and it answers Lee's objection [45] that reviewers may apply different but individually legitimate standards, since publication makes that multiplicity adjudicable rather than sealed inside verdicts. The condition is also already instantiated. Registered reports fix evaluative criteria before outcomes are known, and they have spread to hundreds of journals [46]. The condition is therefore implementable at scale, and the obstacle is neither conceptual nor technical.

Verifiability by construction. Every machine criticism must anchor to quoted text or to a recomputable claim; unanchored criticism is discarded. The scope of this condition should be stated precisely. Anchoring disciplines the evaluator's output and addresses fabrication. It does nothing about adversarial input, which is answered at ingest.

Plurality of independently configured evaluators. There must be at least two, and divergence between them is treated as a signal that requires adjudication, not as noise to be averaged. In the outer form, one evaluator is prompted adversarially and one charitably. The divergence record then doubles as the injection audit of Section 6, and it satisfies Kraemer's independence requirement.

Published procedure, reports, and audits. Confidentiality attaches to unpublished science, not to the evaluation of it. As Section 3 showed, this condition restores the reader's report to the genre in which it began. Bravo et al. [47] find, across five journals, that publishing reports did not degrade referee behavior.

Named human responsibility, with the significance judgment reserved. No manuscript is accepted or rejected by a machine. A named editor signs every decision, and desk rejections must state reasons that engage the manuscript's content. That editor is not the editor of Section 1, who decided in minutes on surface signals and left no record. It is an

editor who decides against published criteria, on an anchored report, stating reasons, and open to appeal. What improves is not the mind but the bindings on it. Merton anticipated this when he observed that “since accountability is itself so much a component of the ethos of science, it may be that the practice of maintaining anonymity of referees will increasingly go by the board” [5] (p. 492).

Three objections deserve answers.

Only a peer can review. Hosseini and Horbach [37] (p. 6) press the strongest version of this objection: peerhood denotes not only expertise but social belonging to an academic community, and it is unclear that machine evaluators satisfy either condition. The concession should be full, since legitimation is tied to the evaluator’s identity in a way that validation is not. But the social component attaches to the accountable editor and to the expert commissioned where significance is contested, not to the checker of citations and degrees of freedom. A procedure can be trusted; only a person can be accountable.

Transparency reforms have been tried and failed. They have: randomized trials of open review found no improvement in quality [15] (pp. 12–13). But each of those reforms failed through one mechanism, namely raising the cost of the scarce input, and the architecture proposed here is the first that lowers it. The BMJ adopted open review despite the null result, on the ground that consequential judgments should be accountable [1] (p. 181). Transparency was tried as a quality intervention and failed as one. It is proposed here as an accountability intervention.

This is solutionism, and it will be mythologized in turn. Csiszar [17] (p. 284) presses the political economy objection against this class of proposal, and his own thesis supplies half of the reply: science’s judgment infrastructure has always been built by adapting commercial forms, so governance is what matters. The other half of the reply is the concession of Section 5.3, that consortium governance is a condition on the proposal rather than a prediction about the industry. Biagioli [16] (pp. 34–35) presses the deeper version of the objection. Review is revered as an impersonal, self-regulating mechanism, in the way markets are, and its acknowledged failures are absorbed by the myth rather than counted against it. A governed pipeline is an obvious candidate for the same treatment, and “the audited machine cannot be biased” is the successor myth. Auditability does not immunize an institution against mythologization; it only makes the myth checkable. The safeguard is that audits be run and published by parties who do not operate the pipeline, and this displaces the problem rather than dissolving it. The same capability could equally industrialize the injustice it claims to remedy, and the difference is one of governance rather than technology. A filtration-first design inherits the function of the two-minute scan. A feedback-first design returns the evaluation to the author and leaves the decision with a named editor.

8. Implications

Three implications follow. For the debate over whether prepublication review is worth having, Heesen and Bright’s dominance result and Polanyi’s defense both treat evaluative attention as a fixed quantity. Once the checkable layer is separable and cheap, neither position can be stated in its original form without saying which layer it concerns. For journals, the desk rather than the referee is where legitimacy is most cheaply gained. Stating reasons against published criteria costs an editor little, and it is the one reform that no reviewer shortage blocks. For the study of automated evaluation, the gain worth claiming is measurability rather than judgment. An evaluator rerun with a cue stated and then stripped turns the experiment that Zuckerman and Merton declared impossible into a routine calibration.

9. Recommendations

The five conditions of Section 7 are the paper's recommendations, and they can be adopted in an order that tracks their risk. Published criteria and reasoned desk decisions require no machine, and they are available to any journal now. The inner form consists of recomputation against deposited data, reference and retraction resolution, and conformity checks against codified reporting standards. It runs on infrastructure the venue controls and takes no free text from the manuscript. It can be adopted wherever a journal controls its own submission pipeline, and its effect can be measured by the pretest–posttest design already run for statcheck [32]. The outer form, evaluators that read the manuscript, should be adopted only where ingest security and divergence auditing are in place, and its reports should be returned to the author as feedback rather than used as inputs to filtration. In every form, audits should be run and published by parties who do not operate the pipeline.

10. Limitations

Two conditions on the arrangement are not currently satisfied. First, an evaluator that reads author-supplied files can be manipulated by hidden instructions, which confines the defensible version to venues that control their own text extraction. Second, the governed, plural infrastructure that the standard requires does not exist and may not be built. Both are conditions on legitimacy rather than on feasibility, so failing them defeats the arrangement, not the standard. Two limits attach to the argument itself. The paper cannot say what fraction of aggregate reviewer hours the checkable portion represents, so it establishes that a saving exists but not its size (Section 5.3). And the coverage of the demonstrated instruments is partial, since statcheck recognizes roughly 60% of reported tests (Section 5.2), so the reliability figures cited hold within the class of operations the tools can parse. Finally, the claim that contestable procedures accumulate warrant faster than sealed ones is a conjecture about mechanism. It is stated in Section 7 as testable, and it is not claimed here to have been shown.

11. Future Research Directions

Three questions are left in testable form. First, whether a procedure that records its criteria, reports, and errors accumulates warrant that a sealed one cannot. The design is a journal-level pretest–posttest with matched controls, of the kind already run for statcheck [32]. Second, whether the specifiability boundary moves. The operations classed as partially specifiable in Table 1 should be re-tested against Cicchetti's bands as parsing and retrieval improve. Third, the status experiment. An evaluator shown to be neutral on the manipulated cue by matched pairs (Section 6) can be run across fields and years at negligible cost, turning a literature of single studies into a monitored quantity. None of these requires the outer form; all can begin with the inner one.

12. Conclusions

The three questions posed at the outset admit direct answers.

Low reliability on checkable dimensions varies substantially with procedural design, and it is not fixed by the human cognitive substrate. Judgment of the same kind, applied to written work whose assessment criteria can be specified in advance, achieves mark–remark correlations above 0.9 under a practiced criterion set, and it falls to peer review's levels when that set is withdrawn [25] (pp. 155–156). Trained reviewers exceed 0.8 on mechanical items while remaining at 0.40 on interpretive ones [20] (p. 180). The restriction to specifiable criteria was made in the body of the paper, and it is not relaxed here. The 0.40 residue on interpretive items marks the limit that human judgment contributes on its own.

What prevented adoption of the known remedies was cost, but only in part. The scarcity assumption is load-bearing on every side of the debate, and capability has altered it unevenly. On fully specifiable checks, a narrow class for which statistical recomputation and retrieval-based verification are the demonstrated cases, the affordable layer reaches a reliability the volunteer system never approached. On partially specifiable checks, it buys coverage rather than accuracy. On the significance judgment, it buys nothing, because there cost was never the obstacle. The near-absence of statcheck adoption shows that price was not the whole obstacle even where it bound. This is why the case must be normative rather than economic.

What legitimacy requires is that verdicts be contestable. This entails published criteria, anchored and checkable criticism, plural and independently configured evaluation, published procedure and reports, and named human responsibility for decisions. Measured against the four-part taxonomy, such an arrangement can claim pragmatic, regulatory, and moral legitimacy: it works on the checkable dimensions, it is auditable by construction, and it evaluates texts where the present system filters persons. It is weakest on cognitive legitimacy, which no argument can supply.

Stated once without qualification: an institution that refuses work without stating reasons, against criteria it does not publish, by a procedure no one can rerun, is illegitimate. It is now illegitimate without excuse, since the cost that once explained the refusal has lapsed where reasons are cheapest to give. Smith was right in 2006 that peer review was the least worst arrangement available, and he was right about why no alternative was in view. The constraint he named has lifted on one narrow layer. What remains is to specify an arrangement that deserves the trust it will be asked to carry.

Author Contributions: Conceptualization, M.A.; methodology, M.A.; formal analysis, M.A.; writing—original draft preparation, M.A. and A.A.; writing—review and editing, M.A. and A.A.; project administration, M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Smith, R. Peer review: A flawed process at the heart of science and journals. *J. R. Soc. Med.* **2006**, *99*, 178–182. [\[CrossRef\]](#)
2. Squazzoni, F.; Bravo, G.; Farjam, M.; Marusic, A.; Mehmani, B.; Willis, M.; Birukou, A.; Dondio, P.; Grimaldo, F. Peer review and gender bias: A study on 145 scholarly journals. *Sci. Adv.* **2021**, *7*, eabd0299. [\[CrossRef\]](#)
3. Aczel, B.; Szaszi, B.; Holcombe, A.O. A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. *Res. Integr. Peer Rev.* **2021**, *6*, 14. [\[CrossRef\]](#)
4. Siler, K.; Lee, K.; Bero, L. Measuring the effectiveness of scientific gatekeeping. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 360–365. [\[CrossRef\]](#)
5. Merton, R.K. *The Sociology of Science: Theoretical and Empirical Investigations*; Storer, N.W., Ed.; University of Chicago Press: Chicago, IL, USA, 1973.
6. Heesen, R.; Bright, L.K. Is peer review a good idea? *Br. J. Philos. Sci.* **2021**, *72*, 635–663. [\[CrossRef\]](#)
7. Kousha, K.; Thelwall, M. Artificial intelligence to support publishing and peer review: A summary and review. *Learn. Publ.* **2024**, *37*, 4–12. [\[CrossRef\]](#)
8. Alfrink, K.; Keller, I.; Kortuem, G.; Doorn, N. Contestable AI by design: Towards a framework. *Minds Mach.* **2023**, *33*, 613–639. [\[CrossRef\]](#)
9. Goldman, A.I. *Knowledge in a Social World*; Oxford University Press: Oxford, UK, 1999.

10. Longino, H.E. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*; Princeton University Press: Princeton, NJ, USA, 1990.
11. Tyler, T.R. *Why People Obey the Law*; Princeton University Press: Princeton, NJ, USA, 2006.
12. Suchman, M.C. Managing legitimacy: Strategic and institutional approaches. *Acad. Manag. Rev.* **1995**, *20*, 571–610. [[CrossRef](#)]
13. Deephouse, D.L.; Bundy, J.; Tost, L.P.; Suchman, M.C. Organizational legitimacy: Six key questions. In *The SAGE Handbook of Organizational Institutionalism*, 2nd ed.; Greenwood, R., Oliver, C., Lawrence, T.B., Meyer, R.E., Eds.; SAGE: London, UK, 2017; pp. 27–54.
14. Bornmann, L. Scientific peer review. *Annu. Rev. Inf. Sci. Technol.* **2011**, *45*, 199–245. [[CrossRef](#)]
15. Lee, C.J.; Sugimoto, C.R.; Zhang, G.; Cronin, B. Bias in peer review. *J. Am. Soc. Inf. Sci. Technol.* **2013**, *64*, 2–17. [[CrossRef](#)]
16. Biagioli, M. From book censorship to academic peer review. *Emergences* **2002**, *12*, 11–45. [[CrossRef](#)]
17. Csiszar, A. *The Scientific Journal: Authorship and the Politics of Knowledge in the Nineteenth Century*; University of Chicago Press: Chicago, IL, USA, 2018.
18. Zuckerman, H.; Merton, R.K. Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva* **1971**, *9*, 66–100. [[CrossRef](#)]
19. Baldwin, M. Scientific autonomy, public accountability, and the rise of “peer review” in the Cold War United States. *Isis* **2018**, *109*, 538–558. [[CrossRef](#)]
20. Cicchetti, D.V. The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behav. Brain Sci.* **1991**, *14*, 119–186. [[CrossRef](#)]
21. Rothwell, P.M.; Martyn, C.N. Reproducibility of peer review in clinical neuroscience: Is agreement between reviewers any greater than would be expected by chance alone? *Brain* **2000**, *123*, 1964–1969. [[CrossRef](#)]
22. Cortes, C.; Lawrence, N.D. Inconsistency in conference peer review: Revisiting the 2014 NeurIPS experiment. *arXiv* **2021**, arXiv:2109.09774.
23. Bailar, J.C. Reliability, fairness, objectivity and other inappropriate goals in peer review. *Behav. Brain Sci.* **1991**, *14*, 137–138. [[CrossRef](#)]
24. Kuhn, T.S. *The Structure of Scientific Revolutions*, 3rd ed.; University of Chicago Press: Chicago, IL, USA, 1996.
25. Laming, D. Why is the reliability of peer review so low? *Behav. Brain Sci.* **1991**, *14*, 154–156. [[CrossRef](#)]
26. Stricker, L.J. Disagreement among journal reviewers: No cause for undue alarm. *Behav. Brain Sci.* **1991**, *14*, 163–164. [[CrossRef](#)]
27. Nuijten, M.B.; Hartgerink, C.H.J.; van Assen, M.A.L.M.; Epskamp, S.; Wicherts, J.M. The prevalence of statistical reporting errors in psychology (1985–2013). *Behav. Res. Methods* **2016**, *48*, 1205–1226. [[CrossRef](#)]
28. Mahoney, M.J. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cogn. Ther. Res.* **1977**, *1*, 161–175. [[CrossRef](#)]
29. Tomkins, A.; Zhang, M.; Heavlin, W.D. Reviewer bias in single- versus double-blind peer review. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 12708–12713. [[CrossRef](#)]
30. Huber, J.; Inoua, S.; Kerschbamer, R.; König-Kersting, C.; Palan, S.; Smith, V.L. Nobel and novice: Author prominence affects peer review. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2205779119. [[CrossRef](#)]
31. Polanyi, M. The republic of science: Its political and economic theory. *Minerva* **1962**, *1*, 54–73.
32. Nuijten, M.B.; Wicherts, J.M. Implementing statcheck during peer review is related to a steep decline in statistical-reporting inconsistencies. *Adv. Methods Pract. Psychol. Sci.* **2024**, *7*, 25152459241258945. [[CrossRef](#)]
33. Liang, W.; Zhang, Y.; Cao, H.; Wang, B.; Ding, D.Y.; Yang, X.; Vodrahalli, K.; He, S.; Smith, D.S.; Yin, Y.; et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI* **2024**, *1*, AIoa2400196. [[CrossRef](#)]
34. Nuijten, M.B.; van Assen, M.A.L.M.; Hartgerink, C.H.J.; Epskamp, S.; Wicherts, J.M. The validity of the tool “statcheck” in discovering statistical reporting inconsistencies. *PsyArXiv* **2017**, 1–32. [[CrossRef](#)]
35. Schmidt, T. Statcheck does not work: All the numbers. Reply to Nuijten et al. (2017). *PsyArXiv* **2017**, 1–5. [[CrossRef](#)]
36. Liu, R.; Shah, N.B. ReviewerGPT? An exploratory study on using large language models for paper reviewing. *arXiv* **2023**, arXiv:2306.00622.
37. Hosseini, M.; Horbach, S.P.J.M. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Res. Integr. Peer Rev.* **2023**, *8*, 4. [[CrossRef](#)]
38. Kraemer, H.C. Do we really want more “reliable” reviewers? *Behav. Brain Sci.* **1991**, *14*, 152–154. [[CrossRef](#)]
39. Checco, A.; Bracciale, L.; Loreti, P.; Pinfield, S.; Bianchi, G. AI-assisted peer review. *Humanit. Soc. Sci. Commun.* **2021**, *8*, 25. [[CrossRef](#)]
40. Sugiyama, S.; Eguchi, R. “Positive review only”: Researchers hide AI prompts in papers. *Nikkei Asia*, 1 July 2025. [[CrossRef](#)]
41. Lin, Z. Hidden prompts in manuscripts exploit AI-assisted peer review. *arXiv* **2025**, arXiv:2507.06185.
42. Collu, M.G.; Janssen-Groesbeek, R.; Koffas, S.; Conti, M.; Picek, S. Publish to perish: Prompt injection attacks on LLM-assisted peer review. *arXiv* **2025**, arXiv:2508.20863.

43. Zhou, R.; Chen, L.; Yu, K. Is LLM a reliable reviewer? A comprehensive evaluation of LLM on automatic paper reviewing tasks. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, 20–25 May 2024; pp. 9340–9351.
44. Peters, D.P.; Ceci, S.J. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behav. Brain Sci.* **1982**, *5*, 187–195. [[CrossRef](#)]
45. Lee, C.J. A Kuhnian critique of psychometric research on peer review. *Philos. Sci.* **2012**, *79*, 859–870. [[CrossRef](#)]
46. Chambers, C.D.; Tzavella, L. The past, present and future of Registered Reports. *Nat. Hum. Behav.* **2022**, *6*, 29–42. [[CrossRef](#)]
47. Bravo, G.; Grimaldo, F.; López-Iñesta, E.; Mehmani, B.; Squazzoni, F. The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nat. Commun.* **2019**, *10*, 322. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.