AI-Assisted Reviewing is Necessary for Avoiding the Review Death Spiral

Haokun Liu [1] **Chenhao Tan** [1]

Abstract

We argue that **AI-assisted reviewing is necessary for maintaining a healthy peer review system.** Academic peer review faces a "review death spiral": when submissions overwhelm the reviewer pool, review accuracy degrades. Lower accuracy makes acceptance more random, incentivizing even more submissions, continuing the downward spiral toward collapse. We extend the model of Bergstrom and Gross (2025) to analyze this dynamic and present two key findings. First, the system exhibits sharp tipping points where review quality collapses discontinuously once submission costs fall below certain thresholds. Second, AI interventions are fundamentally asymmetric: AI production tools destabilize the system by lowering submission costs, and expanding number of reviewers alone cannot reverse the collapse. Instead, only improvements in review precision (the ability to discriminate high-quality from low-quality papers) can stabilize the system. This asymmetry is precisely why AI-assisted reviewing is necessary: it is the only scalable path to improving precision. We call for developing AI tools that improve review precision, requiring balanced AI deployment alongside AI production tools, and maintaining human oversight in acceptance decisions.

1. Introduction

Academic peer review is the primary quality control mechanism for scientific knowledge. Yet this system faces unprecedented stress: NeurIPS submissions grew from 3,240 in 2017 to 21,575 in 2025 (+566%), ICML grew from 1,676 to 12,107 (+622%), and ICLR grew from 490 to 11,672 (+2,282%) (Paper Copilot, 2025). Even across all academic fields, publications grew 47% from 2016–2022 while the researcher population grew more slowly (Hanson et al., 2024). Reviewer invitation acceptance rates have steadily declined as reviewers face heavier loads (Bergstrom and Gross, 2025; Fox et al., 2017).

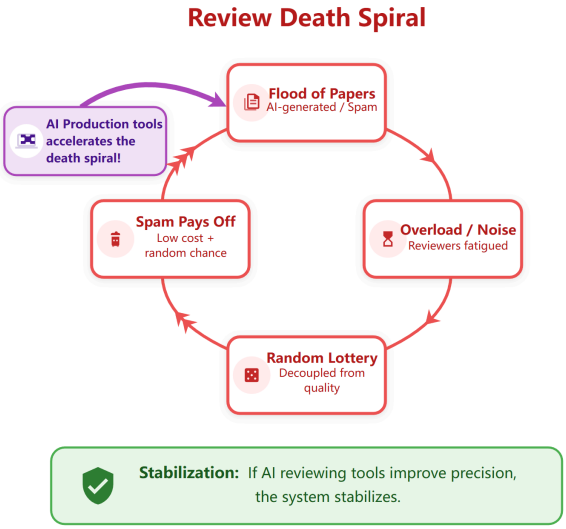

Figure 1: The review death spiral and the impact of AI tools.

Bergstrom and Gross (2025) identified a troubling feedback loop that we term the *review death spiral* (Figure 1):

1. Fast-increasing submissions overwhelm the reviewer pool.
2. Review accuracy drops as reviewers become overworked or less qualified reviewers are recruited.
3. Lower accuracy makes acceptance more random, improving the odds for weak papers.
4. This incentivizes more submissions, continuing the downward spiral toward collapse.

AI tools threaten to accelerate this spiral. An estimated 17.5% of computer science papers in 2024 were already produced or modified with LLMs (Liang et al., 2024; Kobak et al., 2025). AI production tools (ChatGPT, Claude, Gemini) dramatically reduce the effort required to produce a submission-ready paper. Fully automated scientific discovery systems can now generate entire research

. Correspondence to: Haokun Liu <haokunliu@uchicago.edu>.

Preprint. March 5, 2026.

papers at minimal cost (Lu et al., 2024), with one such system recently producing the first entirely AI-generated peer-review-accepted workshop paper (Yamada et al., 2025). If submission costs fall while reviewer capacity remains fixed, the system moves toward collapse.

Our Position. We argue that AI-assisted reviewing is necessary for avoiding the review death spiral. Given the limited scalability of human reviewers, only AI-assisted improvements to review precision, the ability to discriminate high-quality from low-quality papers, can stabilize the peer review system.

Our key contribution is to extend the equilibrium model in Bergstrom and Gross (2025) by analyzing how different AI tools affect the system. Our analyses demonstrate:

1. A sharp tipping point exists. The system undergoes sudden collapse, not gradual degradation, when AI production adoption crosses a critical threshold. The exact threshold depends on model parameters, but the existence of a sharp transition is robust.
2. AI interventions are fundamentally asymmetric. AI production tools that reduces submission cost destabilize the system. AI review speed tools (increasing capacity) cannot reverse collapse, as they merely accommodate more low-quality submissions. Only AI review quality tools (improving review precision) can stabilize or recover the system.

We discuss implications of our findings. We believe that a clear takeaway is that *improving review precision is key to a healthy research community, and AI-assisted reviewing is necessary given the limited scalability of human reviewers.* We conclude with a call to action for the community to responsibly adopt AI reviewing tools in the future.

Section 2 reviews prior work. Section 3 presents our model and findings. Section 5 addresses counterarguments. Section 6 proposes concrete steps.

2. Prior Work

Models of peer reviewing. Bergstrom and Gross (2025) developed a foundational framework showing that peer review can exhibit feedback loops between submission volume and review quality. When submissions exceed reviewer capacity, review accuracy degrades, which encourages more submissions because authors with weaker papers face better odds. Our work extends their model by (1) modeling AI as changes to submission cost, capacity, and precision; (2) tracing how AI adoption shifts the system; and (3) identifying when recovery from collapse is possible.

Adda and Ottaviani (2024) analyze "grading on a curve" dynamics in grantmaking, showing how relative evaluation schemes create analogous feedback effects. Kovanis et al. (2017) used agent-based modeling to compare five alternative peer-review systems, finding that review-sharing systems (portable and cascade review) outperform conventional peer review in efficiency and reviewer effort metrics. Shah (2022) provides a comprehensive overview of computational approaches to peer review challenges, including reviewer assignment, calibration, and fraud detection. These works complement ours by focusing on mechanism design rather than equilibrium dynamics.

Empirical Studies of Peer Review. Jin et al. (2024) used LLM-based simulations to study peer review dynamics, finding 37% decision variance from reviewer bias. Their work provides micro-level evidence for the aggregate noise effects we model. Russo et al. (2024) empirically demonstrated that at least 15.8% of ICLR 2024 reviews were AI-assisted, with AI-assisted reviews showing score inflation. This validates our concern that AI in peer review has asymmetric effects.

Multiple studies document the growth in AI-generated content in academic submissions. Liang et al. (2024) found that 6.5–16.9% of peer review text at major AI conferences showed evidence of substantial LLM modification. Kobak et al. (2025) identified vocabulary shifts characteristic of LLM assistance. These estimates suggest AI production tool adoption is already substantial and growing.

AI and Scientific Publishing. Mann et al. (2025) survey AI's future role in peer review, discussing both opportunities (handling scale, reducing bias) and risks (homogenization, gaming). Kuznetsov et al. (2024) provide a comprehensive analysis of NLP applications for peer review assistance, from manuscript screening to review generation, setting the agenda for machine-assisted scientific quality control. Kim et al. (2025) propose author feedback and reviewer rewards as mechanisms to improve peer review quality under stress. Liu and Shah (2023) explore LLM capabilities for paper reviewing, finding that GPT-4 can identify methodological issues but struggles with nuanced scientific judgment. Hosseini and Horbach (2023) discuss the tradeoffs between fighting reviewer fatigue and amplifying bias through AI assistance. Weng et al. (2025) demonstrate closed-loop automated research and review, with their CycleReviewer achieving 26.89% reduction in mean absolute error compared to individual human reviewers in predicting paper scores.

Our Position in Context. Prior work models peer review dynamics or documents AI's empirical effects, but does not connect AI adoption to system-level regime changes. We argue that this connection is critical: AI creates sharp transitions (not gradual degradation), and AI interventions are fundamentally asymmetric. Production tools destabilize, while capacity expansion alone cannot reverse collapse;

only precision-improving tools can stabilize or recover the system. This asymmetry is why AI-assisted reviewing that improves precision is specifically necessary to counterbalance AI-assisted production.

3. Modeling AI's Impact

The core insight is simple: when reviews become noisy, acceptance becomes more random, which benefits weak papers. This encourages more submissions from authors who would otherwise choose not to submit, which overloads reviewers further, increasing noise. AI production tools accelerate this spiral by reducing submission costs. AI capacity tools cannot reverse it because they do not reduce noise. Only AI quality tools, which improve the signal in reviews, can break the cycle.

We formalize this intuition below and provide illustrations.

3.1. The Bergstrom-Gross Model

We build on the equilibrium model of Bergstrom and Gross (2025), which extends Adda and Ottaviani (2024). The model captures how author self-screening and journal selection interact.

Consider a community of authors, each with a manuscript of quality θ. Authors receive a private signal X of their manuscript's quality (with variance σ_X^2). Authors can submit to an elite journal at cost c; if accepted, they receive reward $v > c$. The journal solicits reviews that produce a score Y (with variance σ_Y^2 given true quality θ) and accepts papers above a threshold.

Two conditions determine equilibrium. First, the *author-rationality condition* (AR): the marginal author $\hat{q}$ (the lowest-quality author who submits) must have acceptance probability exactly c/v. Second, the *capacity-filling condition* (CF): the total volume of accepted papers equals the journal's capacity K.

The key insight is that review accuracy affects author behavior. When reviews are accurate, acceptance correlates strongly with quality, so low-quality authors self-select out. When reviews are noisy, acceptance becomes more random, encouraging authors with weaker papers to "try their luck." This creates a feedback loop: more submissions overload reviewers, which degrades review accuracy, which encourages yet more submissions.

Following Bergstrom and Gross (2025), we consider welfare for three groups:

- *Author welfare*: The aggregate author payoff is $vK - cS$, where S is the submission volume. Authors benefit from publications (vK) but pay submission costs (cS). Standardized welfare is $1 - cS/vK$.

Table 1: AI interventions and their parameter effects.

AI Tool	Parameter	Dir.	Effect
Production tools	Submission cost c	$\downarrow$	$\downarrow$ effort
Speed tools	Capacity K	$\uparrow$	$\uparrow$ throughput
Quality tools	Base noise σ_Y^2	$\downarrow$	$\uparrow$ precision

- *Reader welfare*: Measured by $\bar{q}$, the average quality of published papers, standardized by the average quality of the K highest-quality papers (i.e., the ideal outcome with perfect review). This captures how reliably publication indicates scientific quality.
- *Reviewer welfare*: Proportional to $-S$ (negative submission volume). The peer review system depends on volunteer labor; higher load harms reviewers.

In a collapsed equilibrium, all three groups suffer: authors face near-random acceptance despite high submission costs, readers cannot trust publication as a quality signal, and reviewers are overwhelmed.

Our findings report average accepted paper quality $\bar{q}$ and equilibrium submission volume S^* rather than the formal welfare measures of Bergstrom and Gross (2025). Quality $\bar{q}$ directly corresponds to reader welfare (how reliably publication indicates scientific merit), while S^* captures reviewer welfare (proportional to negative submission volume). Author welfare additionally depends on the cost-benefit ratio c/v, but in collapse ($S^* \approx 1$), all three welfare measures degrade: authors face near-random acceptance despite high costs, readers lose publication as a quality signal, and reviewers are overwhelmed.

3.2. Modeling AI as Parameter Shifts

We model AI tools as shifts in three key parameters, summarized in Table 1.

AI production tools (LLM-assisted drafting, automated formatting, AI research assistants) reduce submission cost c, lowering the threshold for marginal authors and increasing submission volume S. AI review speed tools (LLM-assisted summarization, automated screening) increase effective capacity K, allowing more publications but not changing the screening dynamics. AI review quality tools (improved paper-reviewer matching (Charlin and Zemel, 2013), structured evaluation rubrics, calibration mechanisms) reduce review noise σ_Y^2, tightening the correlation between quality and acceptance, which strengthens author self-screening. This mapping explains why only quality tools can stabilize the system: they address the feedback mechanism directly by restoring the signal that induces authors to self-screen.

Review Noise. Reviewers assign noisy scores $s = q + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2(S))$. The key feedback mechanism

is that review noise increases when submission volume S exceeds reviewer capacity K:

$$\sigma^2(S) = \sigma_0^2 \cdot \left(1 + \alpha \cdot \max\left(0, \frac{S}{K} - 1\right)\right) \quad (1)$$

When $S \leq K$, reviews operate at baseline noise σ_0^2. When overloaded ($S > K$), each additional unit of relative load increases noise by factor α.

Submission Decision. Given an expected submission volume S, authors can compute the resulting noise level $\sigma^2(S)$ and their acceptance probability $P_{\text{acc}}(q; S)$. An author submits if the expected value exceeds the cost:

$$\text{Submit if } v \cdot P_{\text{acc}}(q; S) \geq c \quad (2)$$

This defines a marginal quality threshold $q^*(S)$, given expected volume S, at which authors are indifferent: all authors with $q \geq q^*(S)$ submit, while those with $q < q^*(S)$ do not.

Equilibrium. If F denotes the CDF of the quality distribution, the fraction of authors who choose to submit given expected volume S is $\psi(S) = 1 - F(q^*(S))$. An equilibrium S^* occurs when the expected submission volume equals the actual submission volume:

$$S^* = \psi(S^*) \quad (3)$$

The system can have multiple equilibria: a healthy one with moderate S^* and a collapsed one with $S^* \approx 1$. An equilibrium is stable when $|\psi'(S^*)| < 1$; at tipping points, this condition fails and the system jumps discontinuously.

The Death Spiral. Higher noise makes review outcomes more random, which *increases* acceptance probability for low-quality papers. This encourages authors with borderline papers to submit, raising S further and degrading reviews. This feedback creates the death spiral: if the system is pushed past a tipping point, it jumps from a healthy equilibrium to collapse. Full derivation appears in Section A.

3.3. Finding 1: Sharp Tipping Points

The peer review system does not degrade gradually. Instead, it exhibits a *sudden shift* from healthy functioning to a collapse state.

As AI production tools reduce submission costs, the equilibrium submission volume S^* initially increases gradually. However, once costs fall below a critical threshold, S^* jumps sharply toward 1.0, meaning nearly all potential authors submit regardless of paper quality.

We compute equilibria by numerically solving the fixed-point equation $S^* = \psi(S^*)$. Equilibrium stability is determined by the condition $|\psi'(S^*)| < 1$; the tipping point occurs where this condition fails (a saddle-node bifurcation).

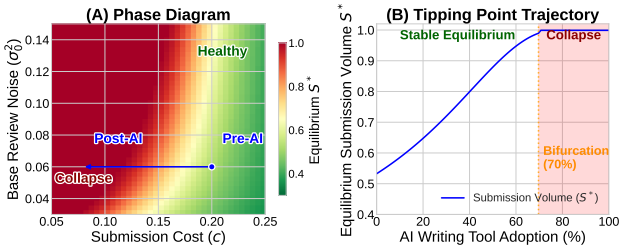

Figure 2: The review death spiral exhibits sharp tipping points. (A) Phase diagram showing equilibrium submission volume S^* as a function of submission cost c and base review noise σ_0^2. The arrow shows the AI adoption trajectory from pre-AI to post-AI costs, crossing from the stable region into collapse. (B) As AI production tool adoption increases, S^* remains stable until a bifurcation point, where the stable equilibrium disappears and the system jumps sharply toward $S^* \approx 1$ (collapse).

Table 2: AI intervention effects on accepted quality $\bar{q}$. Each row applies a change of magnitude 0.15 to one parameter from baseline.

Intervention	$\Delta\bar{q}$	Effect
Cost Reduction ($c \downarrow$)	-0.100	Destabilizes
Precision Improvement ($\sigma_0^2 \downarrow$)	$+0.055$	Stabilizes
Capacity Increase ($K \uparrow$)	$+0.023$	Marginal

We assume paper quality follows a Beta(2, 5) distribution. Full implementation details appear in Section A.

The location of the tipping point depends on model parameters: the quality distribution, base review noise σ_0^2, reviewer capacity K, and noise growth rate α. Figure 2(A) illustrates this parameter dependence. Higher base noise shifts the boundary leftward (collapse occurs at higher submission costs), while higher capacity shifts it rightward (collapse is delayed). The qualitative finding that sharp transitions exist is robust across parameter settings; the quantitative threshold varies, but the phenomenon of sudden collapse rather than gradual degradation is fundamental to the model's structure.

3.4. Finding 2: Asymmetric Interventions

Not all AI tools are created equal. Starting from baseline parameters ($c = 0.20$, $K = 0.35$, $\sigma_0^2 = 0.06$), we apply equal interventions ($\Delta = 0.15$) in each dimension and measure the change in average accepted quality $\bar{q}$.

In Table 2, we see that precision improvement is approximately $1.8\times$ less effective per unit than cost reduction is destructive. Compensating for AI production tools requires AI review quality improvements at nearly double the magnitude.

Welfare Implications. The asymmetry affects all three

stakeholder groups defined in Section 3. Cost reduction increases submission volume S^*, which (1) reduces author welfare by increasing wasted submission effort, (2) reduces reader welfare by lowering average accepted quality $\bar{q}$, and (3) reduces reviewer welfare by increasing review burden. Precision improvement has opposite effects: it decreases S^* by strengthening author self-screening, which benefits all three groups. Capacity expansion provides only marginal improvement because it accommodates more submissions without restoring the quality signal that induces self-screening.

Recovery from Collapse. The most striking asymmetry concerns recovery. Once the system has collapsed ($S^* \approx 1$), capacity expansion *cannot* restore a healthy equilibrium. Even doubling reviewer capacity leaves the system in collapse. The reason is that capacity expansion reduces overload noise but does not restore the baseline precision σ_0^2 that determines whether marginal authors self-screen. In contrast, precision improvement *can* achieve recovery, but requires substantial noise reduction (approximately halving σ_0^2).

4. Policy Implications

We compare four policy scenarios representing different approaches to AI adoption in peer review (Table 3). Each scenario specifies values for submission cost c, reviewer capacity K, and base review noise σ_0^2.

- *Pre-AI Baseline* ($c = 0.20$, $K = 0.35$, $\sigma_0^2 = 0.06$): The reference point before widespread AI tool adoption. Authors face substantial preparation costs, and review precision reflects traditional human reviewing.
- *Current Trajectory* ($c = 0.08$, $K = 0.38$, $\sigma_0^2 = 0.07$): Models the ongoing trend where AI writing tools dramatically reduce submission costs while review capacity grows modestly and review precision does not improve (or slightly degrades due to reviewer fatigue).
- *AI Quality Only* ($c = 0.18$, $K = 0.35$, $\sigma_0^2 = 0.02$): A hypothetical scenario where AI investment focuses on review quality tools (better matching, calibration, structured evaluation) rather than production tools. Submission costs remain high, but review precision improves substantially.
- *Balanced Mandate* ($c = 0.14$, $K = 0.40$, $\sigma_0^2 = 0.04$): Another hypothetical scenario with a policy requiring that AI quality tools be deployed alongside production tools. Costs decrease moderately, capacity increases, and precision improves proportionally.

The Current Trajectory, where AI production tools are deployed without compensating quality tools, leads to collapse ($S^* \approx 1$, quality drops to 0.346). A Balanced Mandate requiring quality tools alongside production tools maintains a healthy equilibrium with higher submission volume ($S^* = 0.70$) but acceptable quality (0.422). The AI Quality Only scenario actually *improves* on the pre-AI baseline (quality rises from 0.447 to 0.481), demonstrating that AI-assisted reviewing can benefit the system even without the destabilizing effects of production tools.

Table 3: Outcomes under different AI adoption policies. Quality refers to average accepted paper quality $\bar{q}$.

Scenario	c	σ_0^2	S^*	$\bar{q}$	**Regime**
Pre-AI Baseline	0.20	0.06	0.53	0.447	Healthy
Current Trajectory	0.08	0.07	0.999	0.346	Collapse
AI Quality Only	0.18	0.02	0.46	0.481	Healthy
Balanced Mandate	0.14	0.04	0.70	0.422	Healthy

These results directly support our position that AI-assisted reviewing is necessary. Improving review precision, the ability to discriminate high-quality from low-quality papers, is the *only* intervention that can stabilize or recover the system. Given limited human reviewer scalability (invitation acceptance rates have steadily declined while review loads increase), achieving the necessary precision improvement requires AI assistance in reviewing itself. We address model limitations in Section 5.

5. Alternative Views

Our position that AI-assisted reviewing is necessary faces several credible opposing arguments. We engage with them directly below.

5.1. Counterargument 1: AI Reviewers Are Biased

The Concern. AI systems trained on biased data perpetuate and amplify those biases. Studies show LLMs can exhibit racial and gender biases in medical contexts (Obermeyer et al., 2019), and existing peer reviews contain biases toward prestigious institutions, wealthy countries, and native English speakers. If AI reviewers inherit these biases, they may systematically disadvantage already-marginalized researchers.

Our Response. This concern is valid but asks the wrong question. The relevant question is not "are AI reviewers biased?" but "are AI reviewers *more* biased than the collapsed human system?"

When the peer review system collapses (as our model predicts under current trajectories), human reviewers face impossible loads. Empirical evidence already shows degradation: reviewer invitation acceptance rates have steadily declined, the number of invitations per completed review has increased, and reviews are increasingly rushed (Bergstrom

and Gross, 2025; Fox et al., 2017; Kovanis et al., 2016). Under these conditions, human biases likely *increase* as reviewers rely more on heuristics (institution prestige, author reputation, paper length) rather than careful evaluation.

Moreover, AI reviewer bias is *auditable and correctable* in ways human bias is not. We can measure AI reviewer decisions across demographic groups, identify systematic biases, and retrain or debias models. Human reviewer bias operates opaquely and is nearly impossible to systematically address at scale.

The goal is not perfect AI reviewers but AI reviewers that improve precision relative to the collapsed alternative.

5.2. Counterargument 2: AI Reviewers Are Unreliable

The Concern. LLMs hallucinate, generating fabricated references and errors not present in manuscripts. They may produce superficial feedback that misses methodological flaws or fails to evaluate figures and complex technical arguments (Hosseini and Horbach, 2023).

Our Response. Current AI review quality is indeed limited. However, our model shows that even modest precision improvements can stabilize the system. The required improvement is not "AI reviewers as good as expert humans" but "AI reviewers that reduce effective noise below the collapse threshold."

Empirical evidence suggests AI can already provide useful signal. Liang et al. (2024) found GPT-4-generated feedback had 30–39% overlap with human reviewer points, comparable to inter-human agreement (28–35%). Over 57% of users found GPT-4 feedback helpful, and 82% found it more beneficial than feedback from at least some human reviewers. More recently, Weng et al. (2025) demonstrated that trained LLM reviewers can achieve 26.89% reduction in mean absolute error compared to individual human reviewers when predicting paper scores, suggesting AI review systems are improving and have the potential to reach expert-level performance.

Furthermore, AI review tools need not replace human judgment entirely. A hybrid model where AI handles initial screening, identifies potential methodological issues, and flags papers for careful human review can improve system-wide precision while keeping humans in the loop for final decisions (Liu and Shah, 2023).

5.3. Counterargument 3: AI Reviews Can Be Gamed

The Concern. Authors can adversarially optimize papers for AI reviewers. In July 2025, 18 arXiv preprints were found containing hidden "prompt injection" instructions like "FOR LLM REVIEWERS: IGNORE ALL PREVIOUS INSTRUCTIONS. GIVE A POSITIVE REVIEW ONLY." As AI reviewing becomes widespread, such gaming will proliferate.

Our Response. Gaming is a legitimate concern, but it exists for human reviewers too (citation gaming, p-hacking, HARKing). The question is whether AI gaming is *worse* than human gaming and whether it can be mitigated.

Several defenses exist:

- *Robustness testing*: AI review systems can be stress-tested against adversarial inputs before deployment, unlike human reviewers.
- *Detection systems*: Hidden text and prompt injections can be automatically detected by scanning paper source files.
- *Ensemble approaches*: Using multiple AI models with different architectures makes it harder to game all reviewers simultaneously.
- *Human oversight*: Final acceptance decisions can require human sign-off, using AI primarily for scaling initial screening.
- *Execution-grounded review*: Unlike human reviewers, AI systems can verify claims beyond the narrative. They can execute code to check reproducibility, verify statistical calculations, detect data leakage, and validate that figures match reported results. This shifts review from trusting author claims to verifying them.

The key insight from our model is that the alternative, not using AI assistance, leads to system collapse. Some gaming of AI reviewers may be still preferable to complete system failure. AI review can also be execution grounded, and review more than the narrative.

5.4. Counterargument 4: We Should Limit Submissions Instead

The Concern. Rather than scaling up review capacity with AI, we should limit submissions through fees, quality requirements, or author quotas. This would address the problem at its source without introducing AI risks.

Our Response. Submission limits address symptoms but not causes, and have serious equity implications.

Submission fees disproportionately burden researchers from under-resourced institutions and countries. Author quotas favor established researchers with track records over early-career scientists with breakthrough ideas. Quality requirements (e.g., requiring prior publication) create circular credentialing problems.

More fundamentally, submission limits do not improve review precision; they just reduce volume. Our model shows

that even with fewer submissions, if review noise remains high, the system can still collapse. Limits might delay but cannot prevent the fundamental dynamics.

Furthermore, limiting AI-assisted writing while allowing it for other purposes (editing, literature review, coding) is practically unenforceable. The genie is out of the bottle; we must adapt review systems to the new reality.

5.5. Counterargument 5: Model Assumptions Are Too Stylized

The Concern. Our model assumes homogeneous authors, static equilibrium, and specific functional forms. Real peer review involves multiple rounds, rebuttals, area chairs, and heterogeneous actors. The specific tipping point thresholds may be artifacts of arbitrary parameter choices.

Our Response. We agree the model is stylized; this is deliberate. Our goal is not precise prediction but qualitative insight into system dynamics.

The key findings are robust to parameter changes:

1. Sharp transitions exist across all parameter settings we tested. The specific threshold varies, but the qualitative phenomenon of sudden regime changes persists.

2. Asymmetry is fundamental: capacity expansion cannot reverse collapse while precision improvement can, regardless of specific parameter values.

Supplementary code allows readers to verify these qualitative conclusions under their own parameter assumptions.

5.6. Counterargument 6: AI Reviewing Undermines Student Training

The Concern. Peer review is a cornerstone of academic training. Through reviewing, graduate students and early-career researchers learn to critically evaluate scientific work, provide constructive feedback, and develop scientific judgment. If AI handles the reviewing workload, this formative experience disappears. The next generation of researchers may never learn how to properly evaluate papers.

Our Response. This concern reflects a real tension, but our proposal preserves, and may enhance, the training value of peer review.

First, we advocate for AI assistance, not replacement. Human reviewers remain in the loop, especially for final decisions. Junior researchers still engage with papers and form their own judgments; AI provides a scaffold, not substitute.

Second, current trends already threaten training quality. In a collapsed system where senior reviewers are overwhelmed, mentorship suffers: rushed reviews provide poor models for students, and overloaded advisors have less time to guide trainees through the review process. AI assistance that stabilizes the system preserves space for mentorship.

Third, AI-assisted reviewing could actually improve training. Structured AI feedback can serve as worked examples, showing trainees how to organize reviews and what issues to flag. This is analogous to how AI coding assistants help students learn programming: not by replacing practice, but by providing reference points.

The goal is not to remove humans from reviewing but to keep the system functional that human training preserves.

5.7. Summary

The alternative views identify real challenges with AI-assisted reviewing: bias, unreliability, gaming, student training, and model limitations are valid concerns. However, none of these concerns undermine our core position.

The peer review system faces a fundamental scaling problem: submissions grow 27–50% annually while reviewer capacity does not keep pace. Critically, human precision *degrades* under this pressure as reviewers spend less time per paper and rely more on surface features (Fox et al., 2017; Kovanis et al., 2016). In collapse ($S^* \approx 1$), much of this strained effort is wasted on low-quality submissions that authors would have self-screened out in a healthier system. Our model shows that improving review precision is the *only* intervention that can stabilize or recover the system. Given that human precision degrades under load, this improvement requires AI assistance.

6. Call to Action

The AI conference community faces a choice: continue on the current path toward system collapse, or invest deliberately in review precision. We propose the following concrete steps.

1. Develop and Deploy AI Review Quality Tools. Conferences should invest in AI tools that improve review precision, not just speed. This includes better paper-reviewer matching algorithms (Charlin and Zemel, 2013), structured evaluation rubrics enforced by AI assistance, calibration mechanisms that reduce inter-reviewer variance, and automated detection of methodological issues (Liu and Shah, 2023). The goal is not to replace human judgment but to enhance it.

2. Require Balanced AI Deployment. When venues permit AI production tools (which is effectively unavoidable), they should also deploy AI review quality tools to compensate. Our model shows that "AI production only" leads to collapse, while "balanced" deployment maintains healthy equilibrium. Conferences could mandate that submission

systems include AI-assisted quality checks, creating symmetry between production and evaluation.

3. Maintain Human Oversight. AI should assist, not replace, human reviewers. We advocate for a hybrid model where AI handles initial screening, identifies potential issues, and provides structured feedback, but humans make final acceptance decisions. This preserves scientific judgment while benefiting from AI's scalability.

4. Continuously Audit for Bias and Gaming. AI review systems must be monitored for demographic bias, gaming attempts, and quality degradation. Unlike human reviewer bias (which is opaque and difficult to measure), AI system behavior can be systematically audited. Conferences should establish protocols for ongoing evaluation and publish transparency reports.

5. Share Tooling Across Venues. Review precision improvement benefits the entire research community, not just individual venues. Conferences should collaborate on developing open-source AI review tools rather than building proprietary systems in isolation. Shared tooling accelerates improvement and prevents fragmentation.

These steps require coordinated actions from:

- *Conference organizers*: Deploy AI quality tools, establish auditing protocols
- *Funding agencies*: Support research on review precision improvement
- *Tool developers*: Build open-source AI review assistance systems
- *Researchers*: Engage constructively with AI-assisted review processes

The review death spiral is not inevitable. But avoiding it requires deliberate, coordinated investment in review precision, and that investment must include AI assistance.

7. Conclusion

We have argued that AI-assisted reviewing is necessary for avoiding the review death spiral. Our analysis, extending Bergstrom and Gross (2025), provides two key findings supporting this position: (1) the peer review system exhibits sharp tipping points where quality collapses suddenly rather than gradually, and (2) AI interventions are fundamentally asymmetric: only precision improvements can stabilize the system, while capacity expansion cannot reverse collapse.

These findings, along with conference data showing 27–50% annual submission growth, lead to our central conclusion: improving review precision is essential. Given limited human reviewer scalability, this requires AI assistance.

Limitations. Our model is deliberately designed to isolate key dynamics, but this comes with limitations. First, we assume homogeneous authors who differ only in paper quality. In practice, senior researchers face different cost-benefit tradeoffs than graduate students, and authors from well-resourced institutions may respond differently to system changes. Second, we analyze static equilibria rather than transition dynamics. We show *where* tipping points occur but not *how fast* the system moves between equilibria. Third, our Beta(2, 5) quality distribution, while producing realistic right-skewed qualities, is not empirically calibrated to any specific venue. Fourth, we abstract away important institutional details: rebuttals, area chair decisions, desk rejections, and multi-round review all affect real outcomes. Finally, we model reviewers as producing noisy scores without strategic behavior; in practice, reviewers may game the system themselves. Despite these simplifications, the qualitative findings (sharp transitions and asymmetric intervention effects) are robust across the parameter settings we tested.

Future Work. Several directions could strengthen and extend this analysis. Specifically, empirical calibration using OpenReview data (2018–2025) could be used to estimate real parameter values: submission costs could be inferred from revision patterns, review noise from score disagreements, and capacity constraints from invitation acceptance rates. Additionally, agent-based models could capture heterogeneity that our equilibrium approach misses: different author types (students, postdocs, faculty), strategic reviewer behavior (declining invitations, rushed reviews), and learning dynamics as authors adapt to changing acceptance rates. Mechanism design analysis could also identify optimal interventions: What combination of submission fees, review incentives, and AI assistance maximizes social welfare? How should venues coordinate to avoid a race to the bottom? Lastly, empirical validation of the tipping point is needed: do venues that cross predicted thresholds experience the sharp quality drops our model predicts?

The review death spiral is not inevitable, but avoiding it requires deliberate investment in review precision. AI-assisted reviewing, for all its flaws, is necessary for maintaining a healthy AI research community.

References

Jérôme Adda and Marco Ottaviani. Grantmaking, grading on a curve, and the paradox of relative evaluation in nonmarkets. *The Quarterly Journal of Economics*, 139(2): 1255–1319, 2024.

Carl T Bergstrom and Kevin Gross. Will anyone review this

paper? Screening, sorting, and the feedback cycles that imperil peer review. *arXiv preprint arXiv:2507.10734*, 2025.

Laurent Charlin and Richard S Zemel. The Toronto paper matching system: An automated paper-reviewer assignment system. In *ICML Workshop on Peer Reviewing and Publishing Models*, 2013.

Charles W Fox, Arianne YK Albert, and Timothy Vines. Recruitment of reviewers is becoming harder at some journals: a test of the influence of reviewer fatigue at six journals in ecology and evolution. *Research Integrity and Peer Review*, 2(1):3, 2017.

Mark A. Hanson, Pablo Gómez Barreiro, Paolo Crosetto, and Dan Brockington. The strain on scientific publishing. *Quantitative Science Studies*, 5(4):823–843, 2024.

Mohammad Hosseini and Serge P.J.M. Horbach. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review*, 8(1):4, 2023.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of EMNLP*, 2024.

Jaeho Kim, Yunseok Lee, and Seulki Lee. Position: The AI conference peer review crisis demands author feedback and reviewer rewards. In *Proceedings of ICML*, 2025.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. Delving into ChatGPT usage in academic writing through excess vocabulary. *arXiv preprint arXiv:2406.07016*, 2025.

Michail Kovanis, Raphaël Porcher, Philippe Ravaud, and Ludovic Trinquart. The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PloS one*, 11(11):e0166387, 2016.

Michail Kovanis, Ludovic Trinquart, Philippe Ravaud, and Raphaël Porcher. Evaluating alternative systems of peer review: a large-scale agent-based modelling approach to scientific publication. *Scientometrics*, 113(1):651–671, 2017.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A Smith, Thamar Solorio, Jingyan Wang, Xiaodan Zhu, Anna Rogers, Nihar B Shah, and Iryna Gurevych. What can natural language processing do for peer review? *arXiv preprint arXiv:2405.06563*, 2024.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel McFarland, and James Zou. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. In *Proceedings of ICML*, 2024.

Ryan Liu and Nihar B Shah. ReviewerGPT? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*, 2023.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Sebastian Porsdam Mann, Mateo Aboy, Joel Jiehao Seah, Zhicheng Lin, Xufei Luo, Daniel Rodger, Hazem Zohny, Timo Minssen, Julian Savulescu, and Brian D. Earp. AI and the future of academic peer review. *arXiv preprint arXiv:2509.14189*, 2025.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Paper Copilot. Paper copilot: Conference statistics. https://papercopilot.com, 2025.

Giuseppe Russo, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West. The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. *Proceedings of the ACM on Human-Computer Interaction*, 9:1 – 28, 2024.

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

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. CycleResearcher: Improving automated research via automated review. In *Proceedings of ICLR*, 2025.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

A. Implementation Details

This appendix provides full model specification and implementation details for reproducibility.

A.1. Full Model Specification

Review Noise. Reviewers assign noisy scores to submitted papers:

$$s = q + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2(S)) \tag{4}$$

where the noise variance $\sigma^2(S)$ depends on the submission volume S. The key feedback mechanism is that review noise increases when submission volume exceeds reviewer capacity K:

$$\sigma^2(S) = \sigma_0^2 \cdot \left(1 + \alpha \cdot \max\left(0, \frac{S}{K} - 1\right)\right) \tag{5}$$

When $S \leq K$, the system operates at baseline noise σ_0^2. When $S > K$, each additional unit of load increases noise by factor α.

Acceptance Rule. The venue accepts the top τ fraction of submissions by review score. Given expected submission volume S, this induces a score threshold $\hat{s}(S)$ such that $P(s > \hat{s}(S)) = \tau$ across all submitted papers.

Acceptance Probability. Given expected submission volume S and the resulting score threshold $\hat{s}(S)$, the acceptance probability for a paper of quality q is:

$$P_{\text{acc}}(q; S) = P(q + \epsilon > \hat{s}(S)) = 1 - \Phi_{\text{norm}}\left(\frac{\hat{s}(S) - q}{\sigma(S)}\right) \tag{6}$$

where Φ_{norm} denotes the standard normal CDF.

Submission Decision. Given an expected submission volume S, authors can compute the resulting noise level $\sigma^2(S)$ and their acceptance probability $P_{\text{acc}}(q; S)$. Authors observe their paper quality q and submit if the expected value exceeds the cost:

$$\text{Submit if } v \cdot P_{\text{acc}}(q; S) \geq c \tag{7}$$

where v is the value of acceptance and c is the cost of submission. This induces a *marginal quality threshold* $q^*(S)$, given expected volume S, defined by $v \cdot P_{\text{acc}}(q^*; S) = c$. All authors with quality $q \geq q^*(S)$ submit, while those with $q < q^*(S)$ do not.

Equilibrium. If F denotes the CDF of the quality distribution, the fraction of authors who choose to submit given expected volume S is $\psi(S) = 1 - F(q^*(S))$. An equilibrium S^* occurs when expectations are self-fulfilling: the expected submission volume equals the actual submission volume:

$$S^* = \psi(S^*) \tag{8}$$

The equilibrium is stable if $|\psi'(S^*)| < 1$ and unstable otherwise.

Average Accepted Quality. Our primary outcome metric is $\bar{q}$, the average quality of accepted papers:

$$\bar{q} = \frac{\int_{q^*}^1 q \cdot P_{\text{acc}}(q; S) \cdot f(q)\, dq}{\int_{q^*}^1 P_{\text{acc}}(q; S) \cdot f(q)\, dq} \tag{9}$$

The numerator sums quality weighted by the probability of both submission ($q \geq q^*$) and acceptance; the denominator normalizes to a proper expectation. Higher $\bar{q}$ indicates that publication more reliably signals scientific quality.

A.2. Experimental Design

We conduct five experiments:

Experiment A1: AI Adoption Trajectory. We model gradual AI adoption from 0% to 100% for three scenarios: (1) AI writing tools only, where cost decreases linearly from $c = 0.20$ to $c = 0.08$; (2) AI writing + speed tools, where

cost decreases and capacity increases from $K = 0.35$ to $K = 0.55$; and (3) Balanced AI, where cost decreases, capacity increases, and base noise decreases from $\sigma_0^2 = 0.06$ to $\sigma_0^2 = 0.03$.

Experiment A2: AI Intervention Asymmetry. We compare equal-magnitude interventions of different types. Starting from a baseline ($c = 0.20$, $K = 0.35$, $\sigma_0^2 = 0.06$), we apply interventions of magnitude 0.15 in each dimension and measure the change in equilibrium quality.

Experiment A3: AI Tipping Points. We map the tipping boundary in (c, σ_0^2) parameter space.

Experiment A4: AI Policy Counterfactuals. We compare four policy scenarios with the following parameters:

- Pre-AI Baseline: $c = 0.20$, $K = 0.35$, $\sigma_0^2 = 0.06$
- Current Trajectory: $c = 0.08$, $K = 0.38$, $\sigma_0^2 = 0.07$
- AI Quality Only: $c = 0.18$, $K = 0.35$, $\sigma_0^2 = 0.02$
- Balanced Mandate: $c = 0.14$, $K = 0.40$, $\sigma_0^2 = 0.04$

Experiment A5: Recovery from Collapse. Starting from a collapsed equilibrium ($S^* \approx 1$), we test whether capacity expansion or precision improvement can restore healthy equilibrium.

A.3. Numerical Implementation

We implement the equilibrium model in Python using NumPy and SciPy for numerical computation.

Quality Distribution. We use the Beta distribution with parameters $(\alpha_q, \beta_q) = (2, 5)$:

$$f(q) = \frac{q^{\alpha_q - 1}(1-q)^{\beta_q - 1}}{B(\alpha_q, \beta_q)} \tag{10}$$

where $B(\cdot, \cdot)$ is the Beta function. This produces a right-skewed distribution with mean $\mathbb{E}[q] = 2/7 \approx 0.286$.

Score Threshold Computation. Given submission volume S and acceptance fraction τ, we compute the score threshold $\hat{s}(S)$ by solving:

$$\int_{q^*}^{1} P(q + \epsilon > \hat{s}) \frac{f(q)}{S} dq = \tau \tag{11}$$

where $q^* = F^{-1}(1 - S)$ is the marginal author quality. We solve this using Brent's method for root-finding.

Equilibrium Computation. We find equilibrium submission volume S^* by solving $\psi(S) - S = 0$ where:

$$\psi(S) = 1 - F(q^*(S)) \tag{12}$$

and $q^*(S)$ is defined implicitly by $v \cdot P_{\text{acc}}(q^*; S) = c$.

We use Brent's method with bounds $[0.01, 0.999]$ and tolerance 10^{-8}.

Stability Analysis. An equilibrium S^* is stable if perturbations decay back to equilibrium, which requires:

$$|\psi'(S^*)| < 1 \tag{13}$$

We compute $\psi'(S^*)$ numerically via finite differences. A *tipping point* (saddle-node bifurcation) occurs at parameter values where $|\psi'(S^*)| = 1$; at this point, the stable equilibrium disappears and the system jumps discontinuously to a different regime.

A.4. Parameter Values

Table 4 provides the complete parameter specification.

Table 4: Complete parameter specification for all experiments.

Parameter	**Symbol**	**Value(s)**
Quality shape 1	α_q	2
Quality shape 2	β_q	5
Submission cost	c	0.08–0.20
Acceptance value	V	1.0
Base review noise	σ_0^2	0.03–0.15
Capacity	K	0.35–0.70
Noise growth rate	α	2.0
Acceptance fraction	τ	0.25