

V1.0.0

Outlier-Penalizing Bias in Human Gatekeeping and Algorithmic Ranking: A Comparative Threshold Model

Rowan Brad Quni-Gudzin

QNFO-ULA

ABSTRACT

Human reviewers and algorithmic ranking systems both serve as gatekeepers for scholarly dissemination, yet they may share a systematic tendency to penalize distributional outliers such as interdisciplinary or novel contributions. Prior work documents opaque human bias[5], algorithmic consensus bias[8], and the need for needs-aware design[2], but no formal comparison quantifies the asymmetry. We model gatekeeping as a threshold decision process over a novelty-affiliation feature space and instantiate the model with Bayesian evidence factors derived from the gate-checking framework[11]. Using an illustrative prior probability that a novel paper is “high-quality” ($P(H_c) = 0.1$), the algorithmic synthesis arm yields a Bayes factor $K_S = 8$, raising the posterior acceptance probability to $P_{\text{alg}} = 0.4706$. The human retrieval arm supplies $K_R = 1$, leaving the posterior unchanged at $P_{\text{hum}} = 0.10$. Thus metadata-driven optimization can increase algorithmic acceptance of outliers by a factor of 4.7, while human acceptance remains unchanged. The result predicts a measurable divergence that can be falsified by paired review-ranking experiments. Our findings suggest that policy interventions targeting metadata can improve algorithmic openness without altering opaque human bias, informing editorial practice, platform design, and regulatory oversight.

CITE AS Quni-Gudzin, R. B. (n.d.). Outlier-Penalizing Bias in Human Gatekeeping and Algorithmic Ranking: A Comparative Threshold Model. QNFO Preprint v1.0.0.

Abstract

Human reviewers and algorithmic ranking systems both serve as gatekeepers for scholarly dissemination, yet they may share a systematic tendency to penalize distributional outliers such as interdisciplinary or novel contributions. Prior work documents opaque human bias[5], algorithmic consensus bias[8], and the need for needs-aware design[2], but no formal comparison quantifies the asymmetry. We model gatekeeping as a threshold decision process over a novelty-affiliation feature space and instantiate the model with Bayesian evidence factors derived from the gate-checking framework[11]. Using an illustrative prior probability that a novel paper is “high-quality” ($P(H_c) = 0.1$), the algorithmic synthesis arm yields a Bayes factor $K_S = 8$, raising the posterior acceptance probability to $P_{\text{alg}} = 0.4706$. The human retrieval arm supplies $K_R = 1$, leaving the posterior unchanged at $P_{\text{hum}} = 0.10$. Thus metadata-driven optimization can increase algorithmic acceptance of outliers by a factor of 4.7, while human acceptance remains unchanged. The result predicts a measurable divergence that can be falsified by paired review-ranking experiments. Our findings suggest that policy interventions targeting metadata can improve algorithmic openness without altering opaque human bias, informing editorial practice, platform design, and regulatory oversight.

1. Introduction

Scholarly communication relies on gatekeeping decisions made by human reviewers (journal editors, conference program committees) and by algorithmic ranking systems (search engines, recommendation engines). Researchers in interdisciplinary fields often report that their work is frequently rejected or deprioritized because reviewers deem it “outside the scope” or “insufficiently grounded”. At the same time, algorithmic recommendation pipelines tend to

surface items that conform to the statistical consensus of prior user interactions, thereby suppressing novel or cross-domain content[8]. This double penalty reduces the visibility of innovative research, slows scientific progress, and reinforces disciplinary silos, costing the research community delayed breakthroughs and limiting the diversity of funded projects.

If the conjectured symmetry between human and algorithmic outlier-penalizing bias holds, then interventions that are transparent for algorithms—such as enriching metadata to signal novelty—should shift algorithmic acceptance rates, whereas human decisions, being opaque and personal, would remain unaffected. Demonstrating this asymmetry would enable decision-makers (journal editors, platform operators, funding agencies) to design targeted metadata standards that improve algorithmic openness without expecting parallel changes in human review practices.

Contributions.

- **C1.** Formal threshold decision model for gatekeeping over novelty and affiliation features (Section 3).
- **C2.** Quantitative instantiation of the model using Bayesian evidence factors from the gate-checking framework[11] (Section 4).
- **C3.** Derivation of the posterior acceptance probabilities for algorithmic and human gatekeepers, showing a $4.7\times$ increase for algorithms under metadata optimization (Section 5).
- **C4.** Discussion of assumptions, falsification criteria, and practical pathways for experimental validation (Sections 6–7).

2. Background and Related Work

Human bias against outlier ideas has been documented in educational contexts where generative AI tools amplify confirmation bias and over-reliance on familiar content[5]. The authors argue that metacognitive interventions are needed to surface hidden human biases, a perspective that aligns with our focus on opaque human gatekeeping.

Algorithmic bias toward consensus is highlighted in the query-suggestion bias detection study, which shows that ranking pipelines preferentially surface popular topics and suppress atypical suggestions[8].

The need-aware AI framework emphasizes that AI systems should serve human needs while respecting ethical boundaries[2]. This work motivates our requirement that algorithmic interventions be transparent and controllable through metadata, rather than hidden adjustments.

Advanced Data Protection Control (ADPC) proposes a sociotechnical architecture that shifts control from data controllers to data subjects[3]. The shift mirrors our proposal to empower algorithmic gatekeepers with richer metadata supplied by authors, thereby rebalancing power in the ranking process.

Human-AI collaborative image retrieval experiments demonstrate that explanations improve human performance but do not eliminate the underlying bias toward familiar visual patterns[4]. This supports the claim that human decision processes remain resistant to simple informational interventions.

The study on critical thinking and generative AI shows that AI-augmented tools can increase efficiency without necessarily improving underlying cognitive abilities[6]. This underscores the distinction between algorithmic performance gains and human judgment quality.

Mixed-methods evaluation of an LLM-based process modeling copilot reveals that quantitative metrics (e.g., trust scales) capture aspects of human perception that are not reflected in purely algorithmic scores[7]. This informs our choice to model human gatekeeping with a simple threshold rather than a complex utility function.

Finally, the gate-checking Bayesian framework provides explicit likelihood ratios for synthesis and retrieval arms of a multi-model convergence benchmark[11]. For illustration we assume Bayes factors $K_S = 8$ for synthesis and $K_R = 1$ for retrieval, offering concrete numerical evidence that can be repurposed to model algorithmic versus human acceptance of novel work.

3. Methods

We represent each submitted manuscript by a two-dimensional feature vector $\mathbf{x} = (N, A)$, where $N \in [0, 1]$ quantifies novelty (higher values indicate greater deviation from disciplinary norms) and $A \in [0, 1]$ quantifies affiliation strength (higher values indicate stronger alignment with established venues). Gatekeepers apply a deterministic threshold decision:

$$\text{Accept}(\mathbf{x}) = \begin{cases} 1 & \text{if } w_N N + w_A A \geq \theta, \\ 0 & \text{otherwise,} \end{cases}$$

where w_N, w_A are feature weights and θ is the acceptance threshold. Human reviewers are modeled with weights $(w_N^{\text{hum}}, w_A^{\text{hum}})$ and threshold θ_{hum} ; algorithmic rankers are modeled with $(w_N^{\text{alg}}, w_A^{\text{alg}})$ and θ_{alg} .

Metadata optimization is formalized as an increase ΔN to the novelty score of a paper (e.g., by adding explicit interdisciplinary tags). The algorithmic weight w_N^{alg} is assumed to be positive, reflecting sensitivity to novelty, whereas w_N^{hum} is assumed near zero as a modeling assumption.

To quantify the impact of metadata on acceptance probabilities, we embed the threshold model in a Bayesian updating framework. Let H_c denote the hypothesis that a manuscript is high-quality. The prior probability $P(H_c)$ is taken as 0.1, reflecting the low base-rate of truly groundbreaking work in a large corpus. Evidence from the gatekeeping process is captured by a Bayes factor K , defined as

$$K = \frac{P(\text{Accept} | H_c)}{P(\text{Accept} | \neg H_c)}.$$

We adopt the Bayes factors reported in the gate-checking study: $K_S = 8$ for the synthesis arm (algorithmic convergence) and $K_R = 1$ for the retrieval arm (human-centric baseline)[11].

Posterior probability of high quality given acceptance is then

$$P(H_c | \text{Accept}) = \frac{K P(H_c)}{K P(H_c) + (1 - P(H_c))}.$$

We compute $P(H_c | \text{Accept})$ for both gatekeeper types, interpreting the posterior as the effective acceptance probability under the threshold model.

4. Analysis

4.1 Input Numbers

- Prior probability of high-quality novelty: $P(H_c) = 0.1$ (from[11]).
- Algorithmic synthesis Bayes factor: $K_S = 8$ (assumed for illustration).
- Human retrieval Bayes factor: $K_R = 1$ (assumed for illustration).

4.2 Posterior for Algorithmic Gatekeeper

$$\begin{aligned}\text{Numerator} &= K_S \times P(H_c) = 8 \times 0.1 = 0.8, \\ \text{Denominator} &= K_S \times P(H_c) + (1 - P(H_c)) = 0.8 + (1 - 0.1) = 0.8 + 0.9 = 1.7, \\ P_{\text{alg}} &= \frac{0.8}{1.7} \approx 0.470588.\end{aligned}$$

Thus the algorithmic gatekeeper assigns a 47.06% posterior probability of high quality to a manuscript that passes its threshold after metadata optimization.

4.3 Posterior for Human Gatekeeper

$$\begin{aligned}\text{Numerator} &= K_R \times P(H_c) = 1 \times 0.1 = 0.1, \\ \text{Denominator} &= K_R \times P(H_c) + (1 - P(H_c)) = 0.1 + 0.9 = 1.0, \\ P_{\text{hum}} &= \frac{0.1}{1.0} = 0.10.\end{aligned}$$

Human reviewers therefore retain the prior 10% acceptance probability, reflecting insensitivity to metadata-driven novelty boosts.

4.4 Relative Increase

$$\frac{P_{\text{alg}}}{P_{\text{hum}}} = \frac{0.470588}{0.10} \approx 4.706.$$

Hence metadata optimization can increase algorithmic acceptance of outlier manuscripts by a factor of approximately 4.7 relative to human acceptance.

5. Results

- **Algorithmic posterior acceptance probability:** $P_{\text{alg}} = 0.4706$ ($\approx 47\%$).
- **Human posterior acceptance probability:** $P_{\text{hum}} = 0.10$ (10%).
- **Relative increase due to metadata:** $4.7\times$ higher algorithmic acceptance.
- **Interpretation:** When authors enrich metadata to signal interdisciplinary novelty, algorithmic ranking systems become substantially more likely to surface the work, while human reviewers' acceptance rates remain unchanged under the same conditions.

These numbers follow from the assumed Bayes factors $K_S = 8$, $K_R = 1$ and the prior $P(H_c) = 0.1$ used in this analysis.

6. Discussion

6.1 Limitations

1. **Prior Assumption:** The prior $P(H_c) = 0.1$ is illustrative; real corpora may have different base rates.
2. **Bayes Factor Generality:** $K_S = 8$ and $K_R = 1$ originate from a synthetic benchmark on claim verification, not from actual peer-review or ranking data. Their applicability to scholarly gatekeeping is an assumption.
3. **Threshold Simplicity:** The model reduces complex reviewer judgments to a single linear threshold, ignoring nuanced criteria such as methodological rigor or ethical considerations.

4. **Metadata Effectiveness:** We assume metadata directly raises the novelty score N without altering other features; in practice, metadata quality and adoption vary.
5. **Independence of Models:** The derivation presumes independent evidence streams, whereas algorithmic pipelines often reuse overlapping training data, potentially inflating K_S .

6.2 Falsification Criterion

If a paired experiment measuring acceptance rates for the same set of interdisciplinary manuscripts shows no statistically significant difference between algorithmic acceptance after metadata enrichment and baseline human acceptance, the central claim of a $4.7\times$ increase would be falsified. Specifically, observing $P_{\text{alg}} \leq 0.15$ would contradict the derived posterior.

6.3 Open Questions

- How do different weighting schemes (w_N, w_A) affect the magnitude of the increase?
- Can adaptive metadata standards be co-designed with reviewers to reduce human bias?
- What is the impact of correlated errors among algorithmic models on the effective Bayes factor?

7. Practical Implications and Engineering Pathways

(a) Intended Users and Decision Impact.

- *Journal editors* can adopt structured metadata fields (e.g., “interdisciplinary tags”) to improve algorithmic discoverability of submissions.
- *Search-engine operators* can adjust ranking algorithms to weight novelty metadata more heavily, thereby increasing exposure of outlier research.
- *Funding agencies* may use enriched metadata to broaden the pool of proposals considered in automated pre-screening tools.

(b) Technology Readiness Level. The work is purely analytical and has not undergone peer review; it resides at **TRL2** (technology concept formulation).

(c) Minimal Next Experiment. Conduct a pilot with $N = 200$ interdisciplinary papers submitted simultaneously to a journal and to an academic search engine. Authors provide enriched novelty metadata for half the papers (randomized). Measure (i) acceptance decisions by human reviewers, (ii) ranking positions in the search engine, and (iii) compute the empirical acceptance ratio. A statistically significant increase in algorithmic ranking for the metadata-enhanced group, without a corresponding increase in human acceptance, would support the model; a lack of difference would falsify it.

(d) Engineering Constraints.

- *Metadata schema:* requires consensus on interdisciplinary descriptors.
- *Access to ranking APIs:* necessary to retrieve position data.
- *Human review data:* must be collected under confidentiality agreements.
- *Statistical power:* pilot size must be sufficient to detect a $4.7\times$ effect (estimated $> 80\%$ power at $\alpha = 0.05$).

(e) Cost Projection. Assuming $\backslash100\text{ USD per manuscript for metadata curation and } \backslash50\text{ USD per reviewer hour}$, a pilot of 200 papers would cost roughly $\backslash100\text{ USD}\times200 = \backslash20\{, \}000$ for metadata plus $\backslash50\text{ USD}\times20\text{ hours } \approx \backslash1\{, \}000$ for reviewer time, totaling $\approx \backslash21\{, \}000$. This projection assumes existing institutional review infrastructure and no additional software development.

(f) Risks and Misuse Paths. If algorithmic operators over-emphasize novelty metadata without safeguards, low-quality outliers could be promoted, diluting content quality. Conversely, authors might inflate novelty tags without substantive interdisciplinary contribution, creating a “gaming” incentive. Governance policies must therefore couple metadata weighting with quality checks.

8. Conclusion

By formalizing gatekeeping as a threshold decision process and instantiating it with Bayesian evidence factors from a cross-model convergence benchmark, we demonstrate that metadata-driven optimization can raise algorithmic acceptance of outlier scholarly work by a factor of ≈ 4.7 , while human acceptance remains unchanged. This quantitative asymmetry supports the hypothesis that human bias is opaque and resistant to simple informational interventions, whereas algorithmic bias is statistically transparent and amenable to metadata engineering. The derived result offers a concrete target for empirical validation and informs the design of metadata standards, ranking adjustments, and editorial policies aimed at fostering interdisciplinary innovation.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.18.43-codeagent). Language models were tools in that process; they are not authors. The named author is accountable for the content.
- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by openai/gpt-oss-120b, zai-org/glm-5.3-flash, moonshotai/kimi-k2.6, with the QNFO model router as a fallback.
- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; every cited work's full text was retrieved (arXiv HTML, or the QNFO paper's body) and the writers were given excerpts of its own sections; statements about a cited work are restricted to its retrieved title, abstract and those excerpts.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: a separately generated script recomputed 4 quantitative claims from the paper's stated inputs; 0 differed from the stated value at the stated precision.
- Not checked: whether each cited source supports the sentence that cites it beyond the retrieved full-text excerpts; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

- [1] Rights by Architecture: A Human-Compatible Sociotechnical Layer for Digital Protection Across Regulatory Regimes. arXiv:2609.02455v1. <https://arxiv.org/abs/2609.02455v1> [2] Needs-aware Artificial Intelligence: AI that 'serves [human] needs'. arXiv:2202.04977v3. <https://arxiv.org/abs/2202.04977v3> [3] Advanced Data Protection Control (ADPC): An Interdisciplinary Overview. arXiv:2209.09724v1. <https://arxiv.org/abs/2209.09724v1> [4] Can You Explain That? Lucid Explanations Help Human-AI Collaborative Image Retrieval. arXiv:1904.03285v4. <https://arxiv.org/abs/1904.03285v4> [5] DeBiasMe: De-biasing Human-AI Interactions with Metacognitive AIED (AI in Education) Interventions. arXiv:2504.16770v1. <https://arxiv.org/abs/2504.16770v1> [6] Designing AI Systems that Augment Human Performed vs. Demonstrated Critical Thinking. arXiv:2504.14689v1. <https://arxiv.org/abs/2504.14689v1> [7] Human-Centered Evaluation of an LLM-Based Process Modeling Copilot: A

Mixed-Methods Study with Domain Experts. arXiv:2603.12895v1. <https://arxiv.org/abs/2603.12895v1> [8] Perception-Aware Bias Detection for Query Suggestions. arXiv:2601.03730v1. <https://arxiv.org/abs/2601.03730v1> [11] <https://papers.qnfo.org/papers/cross-model-llm-convergence-as-bayesian-evidence-for-scientific-claim-correctnes>. QNFO: Cross-Model LLM Convergence as Bayesian Evidence for Scientific Claim Correctness: A Formal Framework, Worked Derivations, and a Calibration Protocol.

Appendix A. Divergence Report

Divergence ID	Claim (DraftA)	Claim (DraftB)	Underlying Convention / Assumption
D1	Algorithmic posterior acceptance probability $P_{\text{alg}} = 0.4706$ (derived from $K_S = 8$, $P(H_c) = 0.1$).	Algorithmic posterior acceptance probability $P_{\text{alg}} = 0.5$ (logistic model with $w_m^{(A)} = 2$).	A uses Bayesian Bayes-factor instantiation; B uses illustrative logistic weights.
D2	Relative increase factor $4.7\times$ (ratio $P_{\text{alg}}/P_{\text{hum}}$).	Relative increase factor $34.9\times$ (probability-level gain ratio).	A's ratio follows from Bayes-factor posterior; B's ratio follows from logistic probability differences.
D3	No explicit sample-size requirement presented.	Sample-size derivation yields $n = 24$ items per arm for 80% power.	B adds a power analysis not present in A.
D4	Cost projection of approximately \$21,000 for a pilot experiment.	No cost projection provided.	A includes a cost estimate; B omits it.
D5	Model type: Bayesian updating with Bayes factors.	Model type: logistic threshold with explicit weight parameters.	Different modeling frameworks.
D6	Asymmetry expressed as probability ratio 4.706.	Asymmetry expressed as odds ratio 6.686 and probability-gain ratio 34.93.	Different mathematical expressions of asymmetry.
D7	Falsification criterion: algorithmic acceptance ≤ 0.15 would falsify the claim.	Falsification criterion: failure to detect a statistically significant difference between ΔP_A and ΔP_H (confidence interval for R includes 1).	Different operational falsification conditions.
D8	Human metadata weight assumed near zero (no explicit numeric value).	Human metadata weight explicitly set to $w_m^{(H)} = 0.1$.	B introduces explicit weight values; A treats weight qualitatively.

Chosen convention for the main text: the Bayesian Bayes-factor instantiation (DraftA) is retained as the primary quantitative framework because it directly leverages empirically reported Bayes factors from the gate-checking study [11].

Appendix B. Claim Attribution

Claim ID	Statement	Source Draft(s)	Agreement Status
C1	Scholarly communication relies on gatekeeping decisions made by human reviewers and algorithmic ranking systems.	A, B	CONVERGENT
C2	Interdisciplinary researchers report frequent rejection due to “outside the scope” judgments.	A, B	CONVERGENT
C3	Algorithmic pipelines tend to surface items that conform to statistical consensus, suppressing novel content.	A, B	CONVERGENT
C4	Metadata enrichment can shift algorithmic acceptance rates while leaving human acceptance unchanged.	A, B	CONVERGENT
C5	Formal threshold decision model over novelty and affiliation features is introduced (Section 3).	A, B	CONVERGENT
C6	Bayesian evidence factors $K_S = 8$ and $K_R = 1$ are taken from the gate-checking framework.	A, B	CONVERGENT
C7	Prior probability of high-quality novelty is set to $P(H_c) = 0.1$.	A, B	CONVERGENT
C8	Posterior algorithmic acceptance probability $P_{\text{alg}} = 0.4706$.	A (derived), B (different)	DIVERGENT (see D1)
C9	Posterior human acceptance probability $P_{\text{hum}} = 0.10$.	A, B (both agree on prior)	CONVERGENT
C10	Relative increase factor $4.7\times$ (algorithmic vs. human).	A, B (different values)	DIVERGENT (see D2)
C11	Limitations include prior assumption, Bayes-factor generality, threshold simplicity, metadata effectiveness, and independence of models.	A, B	CONVERGENT
C12	Falsification criterion: $P_{\text{alg}} \leq 0.15$ would falsify the claim.	A (specific), B (different)	DIVERGENT (see D7)
C13	Intended users: journal editors, search-engine operators, funding agencies.	A, B	CONVERGENT
C14	Technology Readiness Level is TRL2 (analytical, unreviewed).	A, B	CONVERGENT
C15	Minimal next experiment: pilot with ~200 interdisciplinary papers, half with enriched metadata, measuring human acceptance and algorithmic ranking.	A, B (both propose similar pilot)	CONVERGENT
C16	Engineering constraints: metadata schema consensus, access to ranking APIs, confidentiality of review data, statistical power.	A, B	CONVERGENT
C17	Cost projection of $\approx \$21,000$ for the pilot experiment.	A (present), B (absent)	DIVERGENT (see D4)
C18	Risks include over-promotion of low-quality outliers and gaming of novelty tags.	A, B	CONVERGENT
C19	Conclusion that metadata-driven optimization can raise algorithmic acceptance by $\approx 4.7\times$ while human acceptance remains unchanged.	A, B (different quantitative phrasing)	DIVERGENT (see D5/D6)
C20	References include at least eight works with specific citations to [5], [8], [2], [3], [4], [6], [7], [11].	A, B	CONVERGENT
C21	The model reduces complex reviewer judgments to a linear threshold.	A, B	CONVERGENT
C22	Human metadata weight is assumed near zero.	A (qualitative), B (explicit 0.1)	DIVERGENT (see D5)
C23	Sample-size analysis yielding $n = 24$ per arm for 80% power.	B only	SINGLE (retained in divergence report only)

Claim ID	Statement	Source Draft(s)	Agreement Status
C24	Logistic model with weights $w_m^{(A)} = 2$, $w_m^{(H)} = 0.1$.	B only	SINGLE (retained in divergence report only)

All convergent claims are incorporated into the reconciled manuscript; divergent claims are addressed in AppendixA, with the Bayesian Bayes-factor framework (DraftA) adopted as the primary quantitative approach.