From Detection to Operativity: Defeater Propagation in Hybrid AI–Human Peer Review

Ken Howarth
Mercer County Community College

Revised 24 October 2026

Abstract

Hybrid AI–human review systems are evaluated through model accuracy, reviewer agreement, transparency, citation verification, and human oversight. These properties leave a further issue unresolved: whether detected evidence against an initial assessment receives appropriate treatment in the eventual decision. This paper introduces a formal framework for measuring that transition. It models review as a sequential process in which a candidate defeater must be detected, transmitted, taken up, and connected to a reasoned disposition. Defeater Propagation Adequacy (DPA) measures the weighted proportion of detected candidate defeaters receiving traceable downstream disposition; the Presence-to-Operativity Gap (POG) is its complement.

Three structural results follow under stated assumptions. Downstream bottlenecks bound the benefit of improved detection. Additive scoring can dilute a severe minority objection as panel size increases. Mandatory escalation improves transmission when other transition rates remain fixed, but cannot alone secure substantive uptake. An exact parameter sweep illustrates these relations without estimating any platform's performance.

The publicly documented review architecture of the Institute for Complexity Science and Advanced Computing serves as a motivating case. Its citation checks, disagreement flags, review-quality audit, and requirement for human curation provide identifiable stages for prospective propagation research. The analysis distinguishes these policy commitments from evidence about how individual objections are adjudicated.

The contribution is a measurement protocol connecting adverse evidence to authoritative treatment, together with a proposed procedural answerability floor and event-log schema. DPA measures documented treatment, not decision correctness or the adequacy of a rationale. Independent assessment of defeater validity and adjudicative reasons remains necessary. Transparency can reveal available evidence while leaving its decisional force unresolved.

Keywords: human–AI decision making, peer review, defeaters, AI-assisted review, judgment, human oversight, accountability, auditability, methodological transparency

1 Introduction

An AI-assisted review system can detect a consequential objection and preserve it in a report without establishing how that objection informed the final decision. Evaluation therefore needs to follow the objection through the review process. Detection accuracy and human sign-off describe different points in that process; neither reveals, by itself, whether adverse evidence received appropriate authoritative treatment.

Consider an unsupported inference correctly identified in one reviewer report. The objection remains visible in the record, other reviewers score the manuscript favorably, and a curator accepts it without expressly addressing the inference. The outcome need not be wrong. The objection may be mistaken, answered, or outweighed. The record nevertheless leaves its treatment unresolved. The distinction concerns the difference between information being present and information becoming operative in adjudication.

This paper develops a formal model of that transition. It introduces Defeater Propagation Adequacy (DPA), a measure of documented downstream treatment, and the corresponding Presence-to-Operativity Gap (POG). It derives three elementary structural results, provides an exact sensitivity analysis, and specifies an event-log protocol for prospective study. The proposed answerability floor identifies procedural conditions under which authority can be examined in light of material objections.

The model draws on Bettering Bearings (BB), a judgment-centered framework in which evidence constrains judgment and institutions must preserve opportunities for correction. Here that commitment becomes a limited measurement proposal. The paper does not equate documented disposition with good judgment, nor assume that human discretion improves automated assessment. Its unit of analysis is the pathway through which a candidate reason against an assessment reaches authoritative treatment.

2 Human and algorithm interaction

Research on decision support cautions against treating either algorithmic advice or human oversight as a sufficient unit of evaluation. Alon-Barkat and Busuioc (2023), in three experimental studies of administrative decisions, found no general automation-bias pattern relative to equivalent human advice. Selective adherence appeared in some conditions and was absent in another. These results support attention to how advice interacts with the decision-maker’s dispositions rather than an assumption of uniform deference.

Green and Chen (2019a) explicitly place algorithmic advice within a sociotechnical account of human decision making. Their controlled study examines how lay participants interpret and use risk assessments, while cautioning against directly generalizing the results to judges. Green and Chen (2019b) further examine the principles and limits of algorithm-in-the-loop decision making. Together, these studies motivate evaluation of the combined decision process; predictive performance alone cannot establish its adequacy.

For scholarly review specifically, Ye et al. (2024) report vulnerabilities of LLM reviews to manuscript manipulation and other systematic weaknesses. These findings concern the production of reviews; they do not establish how a particular journal subsequently adjudicates them. They nevertheless show why plausible review text and favorable scores cannot exhaust evaluation.

The present contribution specifies one process property within this broader problem: whether detected candidate defeaters receive a traceable disposition. It complements accuracy, calibration, reviewer agreement, and fairness measures. Its novelty lies in the measurement protocol and its application to review, rather than in the probability chain rule or the elementary arithmetic used below.

3 A formal model of defeater propagation

Fix a manuscript version and a completed adjudication episode. Let $\mathcal{D} = \{d_1, \ldots, d_m\}$ be a set of distinct candidate defeaters. A candidate defeater is an objection that, if supported and material, would count against an assessment, inference, recommendation, or decision. Calling an objection a candidate leaves open its validity. Correctly rebutted objections should not be treated as propagation failures.

Assign each candidate a nonnegative weight w_j before observing its final disposition. A possible ordinal coding uses 1 for minor correctable concerns, 2 for substantial concerns, and 4 for validity-threatening concerns. These are protocol choices, not measurements on a validated interval scale. Report unweighted counts and results by severity alongside any weighted score, and test alternative weightings. Duplicate objections should share an identifier or be clustered so that repeated reports do not inflate their importance.

For each candidate, define four binary indicators. D_j records detection; T_j records delivery to the stage exercising final authority; U_j records explicit adjudicative uptake; and A_j records a terminal disposition connected to the final action or a reason for non-action. Terminal disposition may sustain, rebut, outweigh, or remedy an objection, or defer the manuscript decision pending its resolution. A pending escalation alone is not a terminal disposition.

The successful pathway is nested: $A_j \le U_j \le T_j \le D_j$. Detection by a curator rather than an AI reviewer counts as detection at that point; source and timestamps retain this distinction. The same candidate must be tracked through all stages. Independent rediscovery can complete its pathway, but should not be confused with successful transmission from the original source.

$$DPA = \frac{\Sigma_j w_j D_j T_j U_j A_j}{\Sigma_j w_j D_j}$$

$$POG = 1 - DPA$$

These measures are defined only when the denominator is positive. If no weighted candidate defeater is detected, report DPA and POG as not applicable, rather than assigning perfect propagation. Incomplete records are missing observations; they become zeros only when the observation window has closed and the relevant stage's absence is established.

DPA measures recorded disposition conditional on detection. It cannot identify objections that were never detected. Nor can it establish that a rationale substantively answers an objection. To address those limitations, a controlled benchmark needs independently established defects, detection recall and precision, and separate assessment of the adequacy of adjudicative reasons. In observational data without such a benchmark, a high DPA supports a claim about documented treatment only.

3 1 Conditional transition probabilities

For a randomly sampled candidate in a specified population, define pD = P(D = 1), pT = P(T = 1 | D = 1), pU = P(U = 1 | D = T = 1), and pA = P(A = 1 | D = T = U = 1). The chain rule then gives the probability Q of completing the pathway:

$$Q = pD \; pT \; pU \; pA$$

$$P(A = 1 \mid D = 1) = pT \; pU \; pA$$

No independence assumption is needed for this factorization because the transition rates are conditional probabilities. Defined as marginal rates, they could not generally be multiplied. The numerical analysis below assumes a homogeneous population of equally weighted candidates. Its conditional pathway probability represents population DPA under that simplified model; it is not an estimate of a finite-sample weighted ratio for a heterogeneous review corpus.

3 2 Proposition 1 Downstream bounds

For conditional transition rates in [0, 1], Q ≤ min(pD, pT, pU, pA). With downstream rates held fixed, increasing detection by ΔpD changes Q by ΔpD pT pU pA.

Proof. Multiplying a number in [0, 1] by other numbers in that interval cannot increase it. The change formula follows by subtracting the two products. Thus even perfect detection leaves Q bounded by the downstream product. The relation is linear in detection when other rates remain fixed; it does not establish diminishing marginal returns in pD.

Interpretation. Better detection can produce a small absolute improvement when downstream treatment is unreliable. This algebraic bound is not, by itself, a causal prediction about deploying a new detector. A detector may change the volume or type of objections and thereby change downstream transition rates.

4 Aggregation and severe minority objections

Suppose n reviewers score a manuscript on a scale from 1 to 5, and acceptance depends solely on a mean score exceeding a fixed threshold τ < 5. One reviewer assigns 1 after identifying a validity-threatening defect; the other n − 1 reviewers assign 5. The mean is:

$$\bar{s} = (1 + 5(n - 1)) / n = 5 - 4 / n$$

It is 4.2 for five reviewers and 4.6 for ten. The example holds the number of adverse reviewers fixed while favorable scores are added. It does not model a population in which a constant fraction of reviewers detects the defect.

4 1 Proposition 2 Aggregation dilution

If a fixed minority objection is represented only by a bounded additive score, its influence on the mean can approach zero as favorable panel members are added. In the example, the adverse reviewer lowers the mean by 4/n, which converges to zero. Acceptance occurs whenever n > 4/(5 − τ).

The result is conditional on a score-only rule. It does not apply to an architecture that preserves severe objections for separate adjudication, and it does not establish that the minority reviewer is

correct. Its practical implication is to distinguish ordinary gradable disagreement from a candidate defect that would defeat the manuscript’s central claim if sustained.

An invalid step in a proof, an unreproducible central result, or a citation that reverses its source’s claim cannot be adequately represented solely as a small reduction in overall enthusiasm. Such objections require assessment of their validity and materiality. Separate flags can preserve them without granting a single reviewer an unqualified veto.

5 Escalation and adjudication

A mandatory escalation rule can ensure that every detected high-severity candidate reaches the authority stage, setting pT to 1 for that class. The resulting pathway probability is $Q' = pD\ pU\ pA$.

5 1 Proposition 3 Conditional escalation benefit

With pD, pU, and pA held fixed, replacing pT by 1 weakly increases pathway completion. The increase is $pD(1 - pT)pU\ pA \geq 0$. It is strictly positive when pD, pU, and pA are positive and $pT < 1$.

The fixed-rate assumption matters. Escalation can increase workload, encourage perfunctory responses, or alter the mix of cases. Under those conditions pU or pA may fall, and total completion need not improve. A prospective study should measure these downstream effects rather than infer them from the escalation rule.

Even reliable delivery leaves the distinction between seeing an objection and adjudicating it. A curator may acknowledge an alert without assessing its force. Requiring a terminal disposition improves the record’s examinability, but formulaic explanations can satisfy the recording requirement. Substantive assessment must therefore remain separate from the existence of a recorded disposition.

6 Exact sensitivity analysis

The following calculations evaluate the model rather than an operating review platform. There are no sampled manuscripts, observed outcomes, fitted probabilities, or stochastic simulations. All parameters are stipulated, and every reported value follows from multiplication.

Table 1 Illustrative pathway probabilities

Scenario	pD	pT	pU	pA	Q
Baseline	.90	.80	.60	.70	.30240
Detection increase	.99	.80	.60	.70	.33264
Uptake increase	.90	.80	.90	.70	.45360
Combined change	.90	1.00	.90	.85	.68850

Relative to the baseline, the detection change increases Q by .03024, or 10 percent. The uptake change increases Q by .15120, or 50 percent. The uptake intervention is also larger

proportionally; the comparison does not establish that uptake interventions are cheaper or more effective per unit effort. The final row changes several parameters and cannot isolate an escalation effect.

6 1 Parameter sweep

Evaluate all combinations pT, pU, pA ∈ {.2, .4, .6, .8, 1.0}, with pD = .9. The grid contains 125 equally enumerated configurations. Q ranges from .0072 to .9000, with arithmetic mean .1944. Conditional pathway completion ranges from .0080 to 1.0000, with mean .2160. These means describe the chosen grid, not a distribution of actual systems.

Table 2 is a slice at pA = .7 with pD = .9. Each cell gives Q = .63 pT pU. Population DPA in this homogeneous model is obtained by dividing a cell by .9.

Table 2 Pathway probabilities with detection at .9 and disposition at .7

pT / pU	.2	.4	.6	.8	1.0
0.2	0.0252	0.0504	0.0756	0.1008	0.1260
0.4	0.0504	0.1008	0.1512	0.2016	0.2520
0.6	0.0756	0.1512	0.2268	0.3024	0.3780
0.8	0.1008	0.2016	0.3024	0.4032	0.5040
1.0	0.1260	0.2520	0.3780	0.5040	0.6300

Both the scenario comparison and the grid can be reproduced from the formula without external data. Appendix A supplies executable pseudocode using only a standard Python installation. The calculations illustrate multiplicative bottlenecks. They provide no evidence that any journal occupies a particular point on the grid.

6 2 Sensitivity to severity weights

Consider a closed episode with four distinct detected candidates: two minor concerns, one substantial concern, and one validity-threatening concern. The first three receive terminal disposition; the last does not. All have complete observation records, and the severity classes are fixed before examining disposition. Table 3 compares alternative weighting rules on this same illustrative record.

Table 3 Effect of weighting on a fixed disposition record

Weights for minor substantial severe	Disposed weight	Detected weight	DPA
1 1 1	3	4	75.0%
1 2 3	4	7	57.1%
1 2 4	4	8	50.0%

The percentages differ because the rules attach different numerical importance to the same unresolved severe candidate. They do not show that one rule is empirically correct. The 1–2–4 score is $(1 + 1 + 2)/(1 + 1 + 2 + 4) = .50$; the unweighted score is $3/4 = .75$. Reports should identify the severe unresolved candidate alongside every aggregate, compare prespecified weight schemes, and avoid interpreting an ordinal severity ranking as an independently validated ratio scale. This example changes weights only; the parameter sweep in Section 6.1 changes transition rates.

7 ICSAC as a motivating architecture

ICSAC publicly documents separate AI review passes, citation verification, disagreement flags, a review-quality audit, and human curation (ICSAC 2026). These provisions identify stages for a prospective propagation study. Separate reviewer execution does not establish statistically independent errors.

The review-process policy accessed on October 24, 2026 requires a conflicted curation-team member to recuse. If no non-conflicted member is available, the submission waits for an eligible member or a credentialed, non-conflicted reviewer acting for the team; publication does not proceed on the AI panel recommendation alone (ICSAC 2026). This is a requirement to preserve human adjudication during a conflict, rather than an automated fallback.

These are descriptions of policy, not findings about implementation or review outcomes. Preserving human authority leaves an empirical issue open: whether each material candidate reaches that authority and receives a reasoned disposition. Disagreement flagging addresses the transmission concern raised by score-only aggregation, but does not establish later uptake. The present model assigns no DPA score to ICSAC and makes no claim that its curators fail to engage with objections.

A valid study would link unique candidate identifiers to decision records and compare ordinary review, escalation, and cases delayed by conflicts. A case awaiting a non-conflicted adjudicator remains open; delay alone is not a propagation failure. System-level estimates require records for rejected, returned, and deferred submissions as well as accepted papers, with appropriate confidentiality protections. Researchers should retain dated policy and implementation snapshots and distinguish the version retrieved from the version governing a submission.

The motivating role of this case therefore rests on its identifiable review stages and explicit human-authority requirement. These features make further measurement possible without settling the adequacy of particular reasons. Traceable disposition, substantive engagement, and decision correctness remain distinct objects of evaluation.

8 A proposed procedural answerability floor

Auditability permits examination of processing. Answerability additionally concerns the authority exercised and the reasons that can be demanded of its bearer. The following Internal Answerability Floor (IAF) is a proposed procedural criterion for systems claiming answerable human adjudication. It is not a validated scale or a complete theory of responsible judgment.

Condition	Criterion
F_1	A human or institution exercising decisional authority is identified, with its role and authority basis recorded.
F_2	Material candidate defeaters are preserved and made accessible at that authority stage.
F_3	High-severity candidates receive explicit terminal disposition linked to the decision.
F_4	A specified route permits correction, contestation, or revision after a consequential error.

$$IAF = F_1 F_2 F_3 F_4, F_k \in \{0, 1\}$$

This noncompensatory representation implements a floor: success on some conditions cannot numerically cancel complete absence of another. Unknown conditions should be reported as unknown rather than coded as absence. Passing the floor establishes the specified procedural conditions only. It does not establish that the authority is competent, the reasons are sound, or the outcome is just.

F_1 requires particular care in automated cases. Recording a model identifier or an automated rule is necessary for auditability, but does not identify an answerable human or institution by itself. A system may satisfy F_1 through an institution that explicitly accepts responsibility for the rule and provides a usable route to challenge its application. Whether that arrangement is adequate must be assessed independently of the product score.

Different journals can implement these conditions without disclosing confidential deliberations or using identical appeal procedures. What matters is whether a consequential objection has a route to authoritative treatment and whether an affected author can seek correction on stated grounds.

9 Event logs and prospective testing

A prospective study should fix its manuscript versions, observation window, materiality rubric, and rules for grouping duplicates before coding outcomes. Each candidate record should retain its identifier, manuscript and version, detection source and time, independently coded severity, and evidence locator. Stage events should record delivery, uptake, terminal disposition, authority identity, rationale locator, and final manuscript action. Automated and human authority events must be separately marked.

Disposition	Coding meaning
Sustained	The objection is supported and affects the action or its conditions.
Rebutted	A specified reason rejects the objection’s validity or applicability.
Outweighed	The concern remains but stated reasons support the chosen action.
Remedied	A specified correction resolves the concern in a linked version.

Disposition	Coding meaning
Decision deferred	The final action is to withhold a verdict pending specified resolution.
Pending or unresolved	The candidate lacks terminal disposition at the observation cutoff.

The last category is not successful pathway completion. If an episode remains open, treat the observation as censored; if it is closed and an objection has no disposition, code A = 0. An explicit acceptance of unresolved risk counts only when a reasoned terminal disposition and responsible authority are recorded.

Two independent coders should classify severity and uptake before accessing the final manuscript outcome where feasible. A separate blinded assessment should evaluate whether disposition rationales actually engage the objections. Report agreement, adjudication of coding disputes, and sensitivity to the weight scheme. Do not use an automated summary as the sole evidence that a human considered a candidate.

Estimate DPA by severity and source, alongside detection recall and precision when benchmark defects are known. Report the number of manuscripts, candidates, duplicate clusters, missing records, and censored episodes. Uncertainty should account for clustering within manuscripts; any bootstrap should resample manuscripts rather than treat all objections as independent.

A controlled evaluation could insert known methodological or citation defects into research-only manuscript variants and compare an ordinary interface with an escalation-and-disposition interface. The unit of randomization and adjudicator workload should be prespecified. Separate outcomes should include propagation, rationale adequacy, verdict accuracy, time, and false-positive burden. Any test of a third-party pipeline would require an authorized research environment; this paper does not report such a test.

9 1 Estimating conditional transition rates

For a fixed cohort of fully observed candidate records, let N_D, N_T, N_U, and N_A count detection, transmission, uptake, and terminal disposition, respectively. The records must satisfy the nesting in Section 3, with $N_A \le N_U \le N_T \le N_D$. Using the same cohort for every count, estimate each downstream conditional rate as follows.

Table 4 Conditional count estimators for observed candidate records

Transition	Estimate	Conditioning group
Transmission	N_T / N_D	Detected candidates
Uptake	N_U / N_T	Transmitted candidates
Terminal disposition	N_A / N_U	Candidates taken up

When all denominators are positive, the product telescopes: $(N_T/N_D)(N_U/N_T)(N_A/N_U) = N_A/N_D$. The final ratio is unweighted observed DPA. This is a counting identity, not an independence

assumption or a causal model. If an intermediate conditioning group is empty, its transition-rate estimate is undefined; do not assign it an artificial zero. The aggregate N_A/N_D remains defined whenever $N_D > 0$, including when no candidate completes the pathway.

For the severity-weighted DPA, replace each stage count by the sum of fixed candidate weights over that stage. The weighted conditional ratios then telescope to the weighted DPA when their denominators are positive. They describe a severity-weighted target rather than the frequency for a uniformly sampled candidate. Preserve the weight assigned at cohort entry throughout this calculation; if later evidence changes severity, report that revision separately and assess its effect.

Detection sensitivity requires a separate reference set: K independently established material defects in a controlled benchmark. If B_D of those defects are detected, B_D/K estimates benchmark recall. Match reports to distinct defects so that duplicates do not inflate recall, and report false positives and detection precision separately. An ordinary review log cannot supply K because an undetected defect leaves no log entry. Benchmark recall should enter a full-pathway estimate only when the downstream rates refer to the same target population and protocol.

Exclude open or censored cases from a simple complete-episode estimate and disclose that restriction; it can select for faster cases and should not be described as representative of all submissions. Prespecify a follow-up horizon or use an appropriate time-to-disposition analysis if ongoing cases are retained. Report manuscript-level clustering and missingness as specified above. These procedures define a prospective measurement plan; no transition rates for an operating platform are estimated here.

10 Discussion and limitations

DPA can reach 1 while judgment remains poor. Every objection might receive a superficial rationale, or a careful-looking rationale might be substantively mistaken. Conversely, a correct verdict can coexist with low documented propagation. The measure therefore establishes a property of the recorded process, with rationale quality and outcome correctness requiring additional evidence.

The sequential model also simplifies recurrent review. Author responses can change severity, revised manuscripts can remedy defects, and adjudicators can discover new concerns. A longitudinal extension should model repeated episodes and link candidate identifiers across versions. The baseline is useful because it makes the simplest pathway and its failure points explicit.

Severity weighting creates another risk: a system could raise its score by coding hard objections as minor, splitting easy concerns into many candidates, or increasing documentation without improving reasoning. Prespecified coding, duplicate control, unweighted reporting, and independent assessment limit those incentives. A scalar DPA should always be accompanied by severe unresolved cases and stage-specific loss rates.

Human discretion also needs evaluation. An adjudicator can disregard adverse evidence or selectively accept advice. A system that gives humans more freedom can therefore perform worse. The framework supplies no presumption in favor of either greater automation or greater discretion; it specifies the evidence required to assess treatment of adverse reasons in the combined process.

Applications beyond review would require domain-specific definitions of materiality, authority, and correction. Clinical advice, administrative adjudication, and safety supervision cannot inherit a journal's severity rubric unchanged. The transferable proposal is to track a candidate objection through authoritative disposition while assessing the reasons and outcome separately.

11 Conclusion

Hybrid review evaluation should follow detected objections into the final exercise of authority. The formal model shows how downstream rates bound pathway completion, why score-only aggregation can dilute a fixed severe minority objection, and when escalation improves transmission. These are conditional structural results, not evidence about a platform's actual decisions.

The proposed protocol makes a neglected property observable: whether material adverse information receives traceable authoritative treatment. Its evidentiary limit is equally important. A recorded disposition cannot certify the quality of judgment. DPA, independent evaluation of reasons, and outcome measures together provide a more adequate basis for assessing AI-assisted review.

Declarations

Author interest. The author is founder and principal of Bettering Bearings, LLC, whose judgment-centered framework informs this paper. This professional and conceptual interest is disclosed because the manuscript develops that framework's measurement proposals.

AI assistance. AI tools assisted drafting, editing, computational preparation, and source checking. The manuscript develops concepts from the author's Bettering Bearings work. AI tools are not listed as authors; responsibility for the submitted manuscript rests with its human author.

Methods and materials. This paper reports formal derivations and deterministic illustrative calculations, not an empirical audit. The formula, parameter grid, tables, and reproduction procedure are contained in the manuscript. No external dataset or separately hosted software package is claimed.

Appendix A Reproduction procedure

The following standard-library Python procedure reproduces the 125-point grid and Table 2. Each configuration is evaluated once; no seed is required because no random sampling occurs. Values are rounded only for display.

```
from itertools import product
from statistics import mean

grid = (0.2, 0.4, 0.6, 0.8, 1.0)
pD = 0.9
conditional = [t * u * a for t, u, a in product(grid, repeat=3)]
overall = [pD * q for q in conditional]
print(len(overall))  # 125
print(*(f"{x:.4f}" for x in
      (min(overall), max(overall), mean(overall))))
# 0.0072 0.9000 0.1944
print(*(f"{x:.4f}" for x in
      (min(conditional), max(conditional), mean(conditional))))
# 0.0080 1.0000 0.2160
for t in grid:
   print(f"{t:.1f}", *(f"{pD*t*u*0.7:.4f}" for u in grid))
```

For event-log analysis, group records by unique candidate and manuscript version, determine the final stage indicators within the observation window, and evaluate the weighted ratio in Section 3. Do not substitute the product of marginal stage frequencies for the ratio. Report not applicable when the weighted detection denominator is zero.

References

Alon-Barkat, Saar, and Madalina Busuioc. 2023. “Human–AI Interactions in Public Sector Decision Making: ‘Automation Bias’ and ‘Selective Adherence’ to Algorithmic Advice.” Journal of Public Administration Research and Theory 33 (1): 153–169. https://doi.org/10.1093/jopart/muac007.

Green, Ben, and Yiling Chen. 2019a. “Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments.” In Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 90–99. https://doi.org/10.1145/3287560.3287563.

Green, Ben, and Yiling Chen. 2019b. “The Principles and Limits of Algorithm-in-the-Loop Decision Making.” Proceedings of the ACM on Human-Computer Interaction 3 (CSCW), Article 50. https://doi.org/10.1145/3359152.

Institute for Complexity Science and Advanced Computing (ICSAC). 2026. “Review Process.” Accessed October 24, 2026. https://icsacinstitute.org/review-process/.

Ye, Rui, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. 2024. “Are We There Yet? Revealing the Risks of Utilizing Large Language Models in Scholarly Peer Review.” arXiv:2412.01708. https://doi.org/10.48550/arXiv.2412.01708.