An Overview of Challenges, Experiments, and Computational Solutions in Peer Review (Extended Version)

Nihar B. Shah
Machine Learning Department and Computer Science Department
Carnegie Mellon University
nihars@cs.cmu.edu

Abstract

In this overview article, we survey a number of challenges in peer review, understand these issues and tradeoffs involved via insightful experiments, and discuss computational solutions proposed in the literature. The survey is divided into seven parts: mismatched reviewer expertise, dishonest behavior, miscalibration, subjectivity, biases pertaining to author identities, incentives, and norms and policies.

Contents

1 Introduction **3**

2 An overview of the review process **4**

3 Mismatched reviewer expertise **5**
3.1 Computing similarity scores . . . 5
3.1.1 Subject-area selection . . . 5
3.1.2 Text matching . . . 5
3.1.3 Bidding . . . 6
3.1.4 Combining data sources . . . 7
3.2 Computing the assignment . . . 7

4 Dishonest behavior **8**
4.1 Lone wolf . . . 8
4.2 Collusions . . . 9
4.3 Other issues . . . 12

5 Miscalibration **13**

6 Subjectivity **15**
6.1 Commensuration bias . . . 15
6.2 Confirmation bias and homophily . . . 17
6.3 Acceptance via obfuscation ("Dr. Fox effect") . . . 17
6.4 Surprisingness and hindsight bias . . . 17
6.5 Hindering novelty . . . 18
6.6 Positive-outcome bias . . . 18
6.7 Interdisciplinary research . . . 19

[*] This is an extended version of an article in the Communications of the ACM [Sha22a]. This is an evolving draft. The first version appeared on July 8, 2021. The current version is dated August 7, 2025.

7 Biases pertaining to author identities **19**
7.1 Studies in computer science 20
7.2 Studies outside computer science 20
7.3 Design of experiments 21
7.4 Use of author identities in non-anonymized review 21
7.5 De-anonymization of authors in double blind 21

8 Incentives **22**
8.1 Author incentives 22
8.1.1 Open Review 22
8.1.2 Resubmission policies 22
8.1.3 Rolling deadlines 23
8.1.4 Authors' perspectives 23
8.1.5 Citation-based incentives 24
8.2 Reviewer incentives 24
8.2.1 Materialistic and non-materialistic incentives 24
8.2.2 Reviewing the reviews 25
8.2.3 Game-theoretic approaches 25
8.2.4 Signed reviews 25

9 Norms and policies **26**
9.1 Review quality 26
9.1.1 Reviewer training. 26
9.1.2 Following peer-review guidelines 26
9.1.3 Review timeliness 27
9.2 Author rebuttal 27
9.3 Discussions and group dynamics 28
9.3.1 Do panel discussions improve consistency? 28
9.3.2 Influence of other reviewers 29
9.3.3 Herding effects 29
9.3.4 Anonymous versus non-anonymous discussions 29
9.3.5 A survey of reviewers. 30
9.4 Ensuring reviewer anonymity 30
9.5 Time spent by reviewers 30
9.6 Other aspects 31

10 Peer-review objectives **31**
10.1 Ensuring correctness 31
10.1.1 Outside computer science 31
10.1.2 In computer science 32
10.1.3 Overall remarks 33
10.2 Selecting the 'best' research 33
10.2.1 Inter-reviewer agreement 33
10.2.2 Inter-coauthor agreement 34
10.2.3 Relation between reviews and subsequent impact 34
10.3 Constructive feedback to authors 34

11 AI reviewing **35**
11.1 Predicting peer-review scores 35
11.2 Human evaluations of AI reviews 35
11.3 Evaluating correctness 36
11.3.1 Short papers with deliberately inserted flaws 36
11.3.2 Full paper, and direct comparison with human reviewers 36
11.3.3 A chimera test 36
11.4 Identifying better abstracts 36

11.5 Biases . . . 37
11.6 Adversarial attacks on LLM reviewers . . . 37
11.7 Detecting illegitimate LLM-generated reviews . . . 37
11.8 LLM feedback to human reviewers . . . 37

12 Discussion **37**

1 Introduction

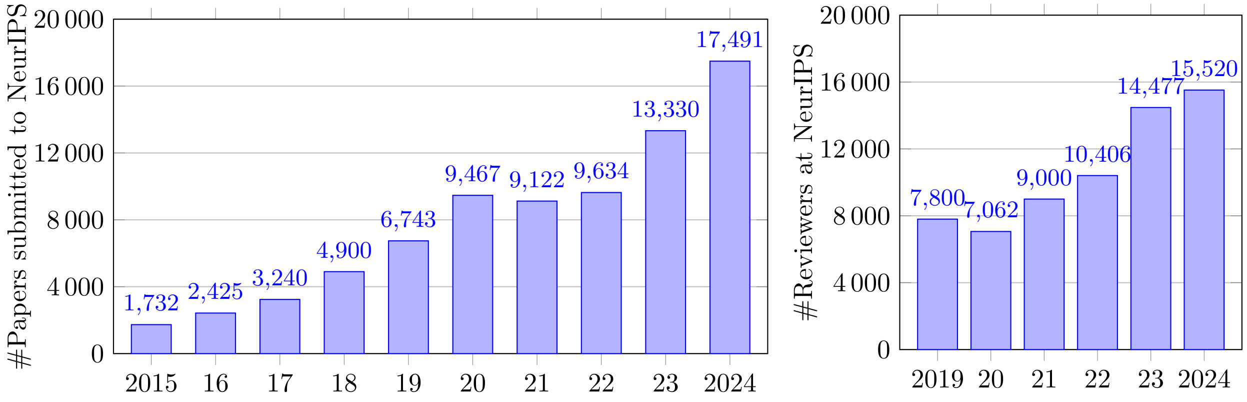


Figure 1: Size of the NeurIPS conference's peer-review process.

Peer review is a cornerstone of scientific research [PF17]. Although quite ubiquitous today, peer review in its current form became popular only in the middle of the twentieth century [Spi02; Bal18]. Peer review looks to assess research in terms of its competence, significance and originality [Bro04]. It aims to ensure quality control to reduce misinformation and confusion [Ben+07] thereby upholding the integrity of science and the public trust in science [WC11; Jam18; OJ21; Kha+21; TA23]. It also helps in improving the quality of the published research [Jef+02]. In the presence of an overwhelming number of papers written, peer review also has another role [Smi97]: "Readers seem to fear the firehose of the internet: they want somebody to select, filter, and purify research material."

Surveys [War16; Tay15; War08; MHR13; Nic+15] of researchers in a number of scientific fields find that peer review is highly regarded by the vast majority of researchers. A majority of researchers believe that peer review gives confidence in the academic rigor of published articles and that it improves the quality of the published papers. These surveys also find that there is a considerable and increasing desire for improving the peer-review process.

Peer review is assumed to provide a "mechanism for rational, fair, and objective decision making" [Jef+02]. For this, one must ensure that evaluations are "independent of the author's and reviewer's social identities and independent of the reviewer's theoretical biases and tolerance for risk" [Lee+13]. There are, however, key challenges towards these goals. The following quote from Rennie [Ren16], in a commentary titled "Let's make peer review scientific" summarizes many of the challenges in peer review: *"Peer review is touted as a demonstration of the self-critical nature of science. But it is a human system. Everybody involved brings prejudices, misunderstandings and gaps in knowledge, so no one should be surprised that peer review is often biased and inefficient. It is occasionally corrupt, sometimes a charade, an open temptation to plagiarists. Even with the best of intentions, how and whether peer review identifies high-quality science is unknown. It is, in short, unscientific."*

Problems in peer review have consequences much beyond the outcome for a specific paper or grant proposal, particularly due to the widespread prevalence of the Matthew effect ("rich get richer") in academia [Mer68; TC14; SG12]. As noted in [TT07] *"an incompetent review may lead to the rejection of the submitted paper, or of the grant application, and the ultimate failure of the career of the author."* This raises the important

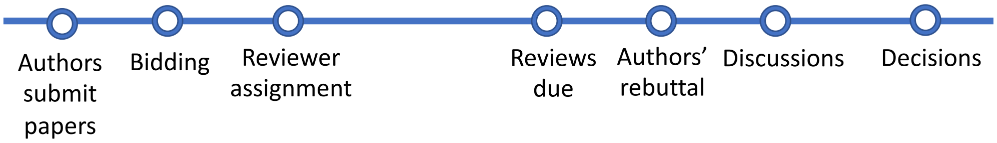


Figure 2: Typical timeline of the review process in computer science conferences.

question [Lee15]: *"In public, scientists and scientific institutions celebrate truth and innovation. In private, they perpetuate peer review biases that thwart these goals... what can be done about it?"* Additionally, the large number of submissions in fields such as machine learning and artificial intelligence (Figure 1) has put a considerable strain on the peer-review process. The increase in the number of submissions is also large in many other fields: *"Submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint"* [McC06]. In medical research, publication of flawed research can cause direct harm to patients [Zar+13; Pou+22]. The annual financial value of the time that researchers dedicate to peer reviewing is estimated to be over two billion dollars [ASH21].

In this overview article on peer review, we discuss several manifestations of the aforementioned challenges, experiments that help understand these issues and the tradeoffs involved, and various (computational) solutions in the literature. For concreteness, our exposition focuses on peer review in scientific conferences.[1] Most points discussed also apply to other forms of peer review such as review of grant proposals used to award billions of dollars worth of grants every year, journal review, and peer evaluation of employees in organizations. Moreover, any progress on this topic has implications for a variety of applications such as crowdsourcing, peer grading, recommender systems, hiring, college admissions, judicial decisions, and healthcare. The common thread across these applications is that they involve distributed human evaluations: a set of people need to evaluate a set of items, but every item is evaluated by a small subset of people and every person evaluates only a small subset of items.

The target audience for this overview article is quite broad. It serves to aid policy makers (such as program chairs of conferences) to design the peer-review process. It can help reviewers understand the inherent biases so that they can actively try to mitigate them. It can help authors and also people outside academia understand what goes on behind the scenes in the peer-review process and the challenges that lie therein.

2 An overview of the review process

We begin with an overview of a representative conference review process. Please see Figure 2 for an illustration. The process is coordinated on an online platform known as a conference management system. Each participant in the peer-review process has one or more of the following four roles: program chairs, who coordinate the entire peer-review process; authors, who submit papers to the conference; reviewers, who read the papers and provide feedback and evaluations; and meta reviewers, who are intermediaries between reviewers and program chairs. There is often a large overlap between the set of authors and the set of (meta) reviewers.

Authors must submit their papers by a pre-decided deadline. The submission deadline is immediately followed by "bidding", where reviewers can indicate which papers they are willing or unwilling to review. The papers are then assigned to reviewers for review. Each paper is reviewed by a handful (typically 3 to 6) of reviewers. The number of papers per reviewer varies across conferences and can range from a handful (3 to 8 in the field of artificial intelligence) to a few dozen papers. Each meta reviewer is asked to handle a few dozen papers, and each paper is handled by one meta reviewer.

Each reviewer is required to provide reviews for their assigned papers before a pre-specified deadline. The reviews comprise an evaluation of the paper and suggestions to improve the paper. The authors may then provide a rebuttal to the review, which could clarify any inaccuracies or misunderstandings in the reviews. Reviewers are asked to read the authors' rebuttal (as well as other reviews) and update their

[1] For those unfamilar with the computer science peer-review culture, unlike many other fields, computer science conferences review full papers, are a venue for archival publication, and are typically rated at par or higher than journals.

reviews accordingly. A discussion for each paper then takes place between its reviewers and meta reviewer. Based on all of this information, the meta reviewer then recommends to the program chairs a decision about whether or not to accept the paper to the conference. The program chairs eventually make the decisions on all papers.

While this description is representative of many conferences (particularly large conferences in the field of artificial intelligence), individual conferences may have some deviations in their peer-review process. For example, many smaller-sized conferences do not have meta reviewers, and the final decisions are made via an in-person or online discussion between the reviewers and program chairs. That said, most of the content to follow in this article is applicable broadly. With this background, we now discuss some challenges and solutions in peer review.

3 Mismatched reviewer expertise

The assignment of the reviewers to papers determines whether reviewers have the necessary expertise to review a paper. The importance of the reviewer-assignment stage of the peer-review process well known: *"one of the first and potentially most important stage is the one that attempts to distribute submitted manuscripts to competent referees"* [RBS07]. Time and again, a top reason for authors to be dissatisfied with reviews is the mismatch of the reviewers' expertise with the paper [McC89].

For small conferences, the program chairs may assign reviewers themselves. However, this approach does not scale to conferences with hundreds or thousands of papers. One may aim to have meta reviewers assign reviewers, but this approach has two problems. First, papers handled by meta reviewers who do the assignment later in time fare worse since the best reviewers for these papers may already be taken for other papers. Second, the question of assigning papers to meta reviewers still remains and is a daunting task if done manually. As a result, reviewer assignments in most moderate-to-large-sized conferences are performed in an automated manner (sometimes with a bit of manual tweaking). Here we discuss automated assignments from the perspective of assigning reviewers, noting that it also applies to assigning meta reviewers.

There are two stages in the automated assignment procedure: the first stage computes "similarity scores" and the second stage computes an assignment using these similarity scores.

3.1 Computing similarity scores

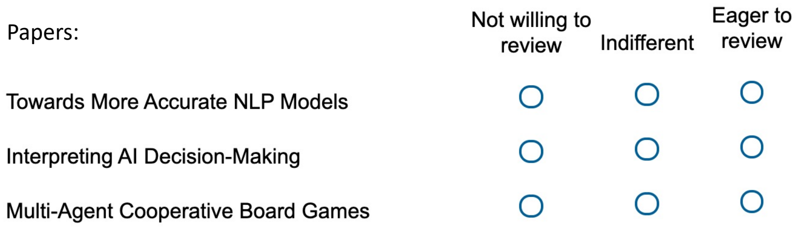
The first stage of the assignment process involves computing a "similarity score" for every reviewer-paper pair. The similarity score $s_{p,r}$ between any paper p and any reviewer r is a number between 0 and 1 that captures the expertise match between reviewer r and paper p. A higher similarity score means a better-envisaged quality of the review. The similarity is computed based on one or more of the following sources of data.

3.1.1 Subject-area selection

When submitting a paper, authors are required to indicate one or more subject areas to which the paper belongs. Before the review process begins, each reviewer also indicates one or more subject areas of their expertise. Then, for every paper-reviewer pair, a score is computed as the amount of intersection between the paper's and reviewer's chosen subject areas.

3.1.2 Text matching

The text of the reviewer's previous papers is matched with the text of the submitted papers using natural language processing techniques [DN92; Bas+99; Fer+06; HP06; MM07; PFS10; CZ13; LSM14; RB08; TCH17; Anj+19; Ker19; Wie+19; Coh+20; Sin+22; ORA+22; Mys+23]. We overview a couple of approaches here [MM07; CZ13]. One approach is to use a language model. At a high level, this approach assigns a higher text-score similarity if (parts of) the text of the submitted paper has a higher likelihood of appearing in the corpus of the reviewer's previous papers under an assumed language model. A simple incarnation of this approach assigns a higher text-score similarity if the words that (frequently) appear in the submitted paper also appear frequently in the papers in the reviewer's previous papers.



Papers:	Not willing to review	Indifferent	Eager to review
Towards More Accurate NLP Models	○	○	○
Interpreting AI Decision-Making	○	○	○
Multi-Agent Cooperative Board Games	○	○	○

Figure 3: A sample interface for bidding.

A second common approach uses "topic modeling". Each paper or set of papers is converted to a vector. Each coordinate of this vector represents a topic that is extracted in an automated manner from the entire set of papers. For any paper, the value of a specific coordinate indicates the extent to which the paper's text pertains to the corresponding topic. The text-score similarity is the dot product of the submitted paper's vector and a vector corresponding to the aggregate of the reviewer's past papers.

These approaches, however, face some shortcomings. For example, suppose all reviewers belong to one of two subfields of research, whereas a submitted paper makes a connection between these two subfields. Then, since only about half of the paper matches any individual reviewer, the similarity of this paper with any reviewer will only be a fraction of the similarity of another paper that lies in exactly one subfield. This discrepancy can systematically disadvantage such a paper in the downstream bidding and assignment processes as discussed later.

Some systems such as the widely employed Toronto Paper Matching System (TPMS) [CZ13] additionally use reviewer-provided confidence scores for each review to improve the similarity computation via supervised learning. The paper [Coh+20] builds language models using citations as a form of supervision.

The paper [Ste+23a] publicly released a "gold standard" dataset for text matching, and in evaluations on this dataset they found that all algorithms incur a significant amount of error (12%-30% error in easy cases to 36%-43% in hard cases). Developing improved text-matching algorithms that incur lower errors is an important open problem.

3.1.3 Bidding

Many conferences employ a "bidding" procedure where reviewers are shown the list of submitted papers and asked to indicate which papers they are willing or unwilling to review. A sample bidding interface is shown in Figure 3.

Cabanac and Preuss [CP13] analyze the bids made by reviewers in several conferences. In these conferences, along with each review, the reviewer is also asked to report their confidence in their evaluation. They find that assigning papers for which reviewers have made positive (willing) bids is associated with higher confidence reported by reviewers for their reviews. This observation suggests the importance of assigning papers to reviewers who bid positively for the paper. Such suggestions are corroborated elsewhere [PF17], noting that the absence of bids from some reviewers can reduce the fairness of assignment algorithms.

Many conferences suffer from the lack of adequate bids on a large fraction of submissions. For instance, 146 out of the 264 submissions at the ACM/IEEE Joint Conference on Digital Libraries (JCDL) 2005 had zero positive bids [RBS07]. In IMC 2010, 68% of the papers had no positive bids [BA12]. The Neural Information Processing Systems (NeurIPS) 2016 conference in the field of machine learning aimed to assign 6 reviewers and 1 meta-reviewer to each of the 2425 papers, but 278 papers received at most 2 positive bids and 816 papers received at most 5 positive bids from reviewers, and 1019 papers received zero positive bids from meta reviewers [Sha+18]. One reason is a lack of reviewer engagement in the review process: 11 out of the 76 reviewers at JCDL 2005 and 148 out of 3242 reviewers at NeurIPS 2016 did not give any bid information.

Cabanac and Preuss [CP13] also uncover a problem with the bidding process. The conference management systems there assigned each submitted paper a number called a "paperID". The bidding interface then ordered the papers according to the paperIDs, that is, each reviewer saw the paper with the smallest paperID at the top of the list displayed to them, and increasing paperIDs thereafter. They found that the number of bids placed on submissions generally decreased with an increase in the paperID value. This phenomenon is explained by well-studied serial-position effects [MHM06] that humans are more likely to interact with an item if shown at the top of a list rather than down the list. Hence, this choice of interface results in a systematic bias against papers with greater values of assigned paper IDs.

Cabanac and Preuss suggest exploiting serial-position effects to ensure a better distribution of bids across papers by ordering the papers shown to any reviewer in increasing order of bids already received. However, this approach can lead to a high reviewer dissatisfaction since papers of the reviewer's interest and expertise may end up significantly down the list, whereas papers unrelated to the reviewer may show up at the top. An alternative ordering strategy used commonly in conference management systems today is to first compute a similarity between all reviewer-paper pairs using other data sources, and then order the papers in decreasing order of similarities with the reviewer. Although this approach addresses reviewer satisfaction, it does not exploit serial-position effects like the idea of Cabanac and Preuss. Moreover, papers with only moderate similarity with all reviewers (e.g., if the paper is interdisciplinary) will not be shown at the top of the list to anyone.

These issues motivate an algorithm [FSR20] that dynamically orders papers for every reviewer by trading off reviewer satisfaction (showing papers with higher similarity at the top, using metrics like the discounted cumulative gain or DCG) with balancing paper bids (showing papers with fewer bids at the top). A second paper [Mei+21] also looks to address the problem of imbalanced bids across papers, but via a different approach. Specifically, it proposes a market-style bidding scheme where it is more "expensive" for reviewer to bid on a paper which has already received many bids. These approaches are empirically evaluated in [Roz+23], which finds that both these approaches [FSR20; Mei+21] help in balancing the bids.

3.1.4 Combining data sources

The data sources discussed above are then merged into a single similarity score. One approach is to use a specific formula for merging, such as $s_{p,r} = 2^{\text{bid-score}_{p,r}}(\text{subject-score}_{p,r} + \text{text-score}_{p,r})/4$ used in the NeurIPS 2016 conference [Sha+18], or $(\frac{1}{2}\text{subject-score}_{p,r} + \frac{1}{2}\text{text-score}_{p,r})^{\frac{1}{\text{bid-score}_{p,r}}}$ used in the AAAI 2021 conference [LBM21], or a convex combination of the data sources as done in the OpenReview.net platform. A second approach involves program chairs trying out various combinations, eyeballing the resulting assignments, and picking the combination that seems to work best. Finally and importantly, if any reviewer r has a conflict with an author of any paper p (that is, if the reviewer is an author of the paper or is a colleague or collaborator of any author of the paper), then the similarity $s_{p,r}$ is set as -1 to ensure that this reviewer is never assigned this paper.

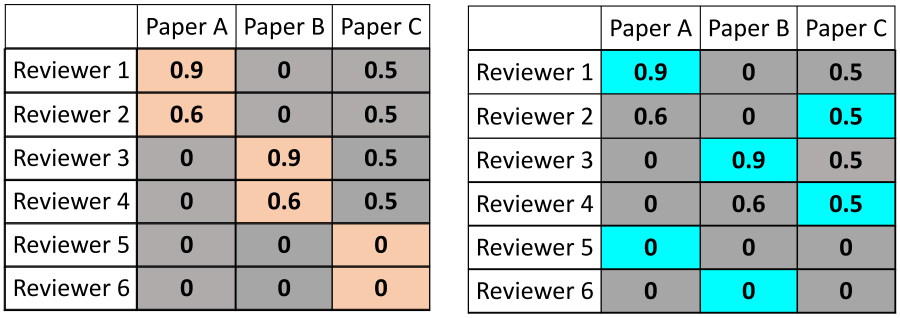
3.2 Computing the assignment

The second stage assigns reviewers to papers in a manner that maximizes some function of the similarity scores of the assigned reviewer-paper pairs. The most popular approach is to maximize the total sum of the similarity scores of all assigned reviewer-paper pairs [CZ13; GS07; Tay08; TTT10; CZB12; Lon+13; LH16]:

$$\underset{\text{assignment}}{\text{maximize}} \sum_{\text{papers } p} \sum_{\substack{\text{reviewers } r \\ \text{assigned to paper } p}} s_{p,r},$$

subject to load constraints that each paper is assigned a certain number of reviewers and no reviewer is assigned more than a certain number of papers.

This approach of maximizing the sum of similarity scores can lead to unfairness to certain papers [SSS21b]. As a toy example illustrating this issue, consider a conference with three papers and six reviewers, where each paper is assigned two reviewers and each reviewer is assigned one paper. Suppose the similarities are given by the table on the left-hand side of Figure 4. Here {paper A, reviewer 1, reviewer 2} belong to one research discipline, {paper B, reviewer 3, reviewer 4} belong to a second research discipline, and paper C's content is split across these two disciplines. Maximizing the sum of similarity scores results in the assignment shaded light/orange in the left-hand side of Figure 4. Observe that in this example, the assignment for paper C is quite poor: all assigned reviewers have a zero similarity with paper C. This is because this method assigns better reviewers to papers A and B at the expense of paper C. Such a phenomenon is indeed found to occur in practice. The paper [KSM19] analyzes data from the Computer Vision and Pattern Recognition



	Paper A	Paper B	Paper C
Reviewer 1	🟧 **0.9**	**0**	**0.5**
Reviewer 2	🟧 **0.6**	**0**	**0.5**
Reviewer 3	**0**	🟧 **0.9**	**0.5**
Reviewer 4	**0**	🟧 **0.6**	**0.5**
Reviewer 5	**0**	**0**	🟧 **0**
Reviewer 6	**0**	**0**	🟧 **0**

	Paper A	Paper B	Paper C
Reviewer 1	🟦 **0.9**	**0**	**0.5**
Reviewer 2	**0.6**	**0**	🟦 **0.5**
Reviewer 3	**0**	🟦 **0.9**	**0.5**
Reviewer 4	**0**	**0.6**	🟦 **0.5**
Reviewer 5	🟦 **0**	**0**	**0**
Reviewer 6	**0**	🟦 **0**	**0**

Figure 4: Assignment in an fictitious example conference using the popular sum-similarity optimization method (left) and a more balanced approach (right).

(CVPR) 2017 and 2018 conferences, which have several thousand papers. The analysis reveals that there is at least one paper each to which this method assigns all reviewers with a similarity score of zero with the paper, whereas other assignments (discussed below) can ensure that every paper has at least some reasonable reviewers.

The right-hand side of Figure 4 depicts the same similarity matrix. The cells shaded light/blue depict an alternative assignment. This assignment is more balanced: it assigns papers A and B reviewers of lower similarity as compared to earlier, but paper C now has reviewers with a total similarity of 1 rather than 0. This assignment is an example of an alternative approach [Gar+10; SSS21b; KSM19; Lia+18] that optimizes for the paper which is worst-off in terms of the similarities of its assigned reviewers:

$$\underset{\text{assignment}}{\text{maximize}} \quad \underset{\text{papers } p}{\text{minimum}} \quad \sum_{\substack{\text{reviewers } r \\ \text{assigned to paper } p}} \quad s_{p,r},$$

The approach then optimizes for the paper that is the next worst-off and so on. Evaluations [KSM19; SSS21b] of this approach on several conferences reveal that it significantly mitigates the problem of imbalanced assignments, with only a moderate reduction in the sum-similarity score value as compared to the approach of maximizing sum-similarity scores. Furthermore, the assignment algorithm [SSS21b] is found to also have desirable properties such as low “envy”, high “Nash social welfare’, and a high similarity on the bottom 10% and the bottom 25% papers [PZ22]. This approach is now adopted in conferences such as the International Conference on Machine Learning (ICML) 2020 [SSS21b] and for peer review of proposals in astronomy [CCS25].

Recent work also incorporates various other desiderata in the reviewer-paper assignments such as geographic diversity [LBM21], envy-freeness [PZ22], and addressing uncertainty in the similarities [CPZ23]. See the paper [JR22] for a survey of researchers on the importance they place on various desiderata in the assignments. An emerging concern when doing the assignment is that of dishonest behavior, as we discuss next.

4 Dishonest behavior

The outcomes of peer review can have a considerable influence on the career trajectories of authors. While we believe that most participants in peer review are honest, the stakes can unfortunately incentivize dishonest behavior. In the next two subsections, we focus on two issues that are more closely tied to conference peer review. In the third and final subsection, we overview other issues of dishonest behavior.

4.1 Lone wolf

Conference peer review is competitive, that is, a roughly pre-determined number (or fraction) of submitted papers are accepted. Moreover, many authors are also reviewers. Thus a reviewer could increase the chances of acceptance of their own papers by manipulating the reviews (e.g., providing lower ratings) for other papers.

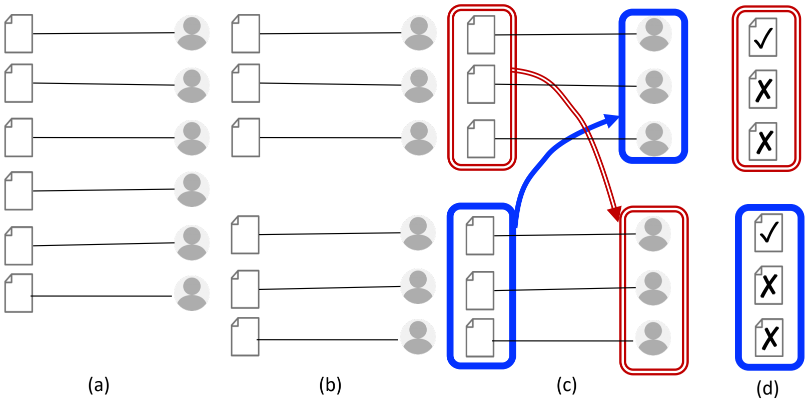


Figure 5: Partition-based method for strategyproofness.

A controlled study by Balietti et al. [BGH16] examined the behavior of participants in competitive peer review. Participants were randomly divided into two conditions: one where their own review did not influence the outcome of their own work, and the other where it did. Balietti et al. observed that the ratings given by the latter group were drastically lower than those given by the former group. They concluded that *"competition incentivizes reviewers to behave strategically, which reduces the fairness of evaluations and the consensus among referees."* The study also found that the number of such strategic reviews increased over time, indicating a retribution cycle in peer review.

Similar concerns of strategic behavior have been raised in the NSF review process [NL13]. See [Aks10; And+07; Lan08] for more anecdotes and [SSS21a] for a dataset comprising such strategies. The paper [TH11] posits that even a small number of selfish, strategic reviewers can drastically reduce the quality of scientific standard.

This motivates the requirement of "strategyproofness": no reviewer must be able to influence the outcome of their own submitted paper by manipulating the reviews they provide. A simple yet effective idea to ensure strategyproofness is called the partition-based method introduced in [Alo+11] and studied subsequently in many papers [HM13; BNV14; FK15; Kur+15; Kah+18; Xu+19; Azi+19; Dhu+22] (see [OW22] for an extensive survey on strategyproofing). The key idea of the partition-based method is illustrated in Figure 5. Consider the "authorship" graph in Figure 5a whose vertices comprise the submitted papers and reviewers, and an edge exists between a paper and reviewer if the reviewer is an author of that paper. The partition-based method first partitions the reviewers and papers into two (or more) groups such that all authors of any paper are in the same group as the paper (Figure 5b). Each paper is then assigned for review to reviewers in the other group(s) (Figure 5c). Finally, the decisions for the papers in any group are made independent of the other group(s) (Figure 5d). This method is strategyproof since any reviewer's reviews influence only papers in other groups, whereas the reviewer's own authored papers belong to the same group as the reviewer.

The partition-based method is largely studied in the context of peer-grading-like settings. In peer grading, one may assume each paper (homework) is authored by one reviewer (student) and each reviewer authors one paper, as is the case in Figure 5. Conference peer review is more complex: papers have multiple authors and authors submit multiple papers. Consequently, in conference peer review it is not clear if there even exists a partition. Secondly, peer grading is more homogeneous where any paper can be assigned to any reviewer, whereas papers and reviewers in peer review are much more specialized (Section 3). Hence, even if such a partition exists, the partition-based constraint on the assignment could lead to a considerable reduction in the assignment quality. Such questions about realizing the partition-based method in conference peer review are still open, with promising initial results [Xu+19; Dhu+22] showing that such partitions do exist in practice and the reduction in quality of assignment may not be too drastic.

4.2 Collusions

Various investigations have uncovered dishonest collusions in peer review (e.g., [Vij20a; Vij20b; Lit21; Lau20; Lau19] and many more). Here a reviewer and an author come to an understanding: the reviewer manipulates the system to try to be assigned the author's paper (or proposal), then accepts the paper if assigned, and the author offers quid pro quo either in the same conference or elsewhere. There may be collusions between more than two people, where a group of reviewers (who are also authors) illegitimately push for each others' papers.[2]

The first line of defense against such behavior is conflicts of interest: one may suspect that colluders may know each other well enough to also have co-authored papers. Then treating previous co-authorship as a conflict of interest, and ensuring to not assign any paper to a reviewer who has a conflict with its authors, may seem to address this problem. It turns out that even if colluders collaborate, they may go to great lengths to enable dishonest behavior [Vij20a]: *"There is a chat group of a few dozen authors who in subsets work on common topics and carefully ensure not to co-author any papers with each other so as to keep out of each other's conflict lists (to the extent that even if there is collaboration they voluntarily give up authorship on one paper to prevent conflicts on many future papers)."* Separately, there are also cases where authors and reviewers have provided fake information, which if left unchecked, can circumvent conflict-of-interest defenses.

A second line of defense addresses attacks where two or more reviewers (who have also submitted their own papers) aim to review each other's papers. This has motivated the design of assignment algorithms [Guo+18; BBN21] with an additional constraint of disallowing any loops in the assignment, that is, ensuring to not assign two people each others' papers. Such a condition of forbidding loops of size two was also used in the reviewer assignment for the Association for the Advancement of Artificial Intelligence (AAAI) 2021 conference [LBM21]. This defence prevents colluders engaging in a quid pro quo in the same venue. However, this defense can be circumvented by colluders who avoid forming a loop, for example, where a reviewer helps an author in a certain conference and the author reciprocates elsewhere. Moreover, it has been uncovered that, in some cases, an author pressures a certain reviewer to get assigned and accept a paper [Lau20]. This line of defense does not guard against such situations where there is no quid pro quo within the conference.

A third line of defense is based on the observation that the bidding stage of peer review is perhaps the most easily manipulable: reviewers can significantly increase the chances of being assigned a paper they may be targeting by bidding strategically [Jec+20; Wu+21]. This suggests curtailing or auditing bids, and this approach is followed in the paper [Wu+21]. This work uses the bids from all reviewers as labels to train a machine learning model which predicts bids based on the other sources of data. This model can then be used as the similarities for making the assignment. It thereby mitigates dishonest behavior by de-emphasizing bids that are significantly different from the remaining data. In parallel, it is also found that popular fraud-detection algorithms from other domains fail to detect malicious bids in peer review [Jec+24].

A challenge with the aforementioned method [Wu+21], however, is that there remains only little influence of the bids (of honest reviewers) on the choice of papers assigned to them [Jec+22b]. Consequently, this may hinder the very purpose of bidding (of correcting any issues in the other similarities computed) and may reduce the incentive for honest reviewers to engage in the bidding process.

Dishonest collusions may also be executed without bidding manipulations. For example, the reviewer/paper subject areas and reviewer profiles may be strategically selected to increase the chances of getting assigned the target papers, or the use of rare keywords [Ail+19]. Such methods of collusion have been found to occur in practice.

Security researchers have demonstrated the vulnerability of paper assignment systems to attacks on the text-matching part. The next two attacks we discuss are of this form. To be clear, the existence or extent of such attacks in practice are unknown, and these are created by security researchers to understand possible vulnerabilities.

The first line of attacks on automated text matching involves an author manipulating the PDF (portable document format) of their submitted paper so that a certain reviewer gets assigned [Mar+17; TJ19]. These attacks insert text in the PDF of the submitted paper in a manner that satisfies three properties: (1) the inserted text matches keywords from a target reviewers' paper; (2) this text is not visible to the human reader; and (3) this text is read by the (automated) parser which computes the text-similarity-score between the submitted paper and the reviewer's past papers. These three properties guarantee a high similarity for the colluding reviewer-paper pair, while ensuring that no human reader detects it. These attacks are

[2] A related reported problem involves settings where a reviewer for any paper can see the identities of the other reviewers for that paper. Here a colluding reviewer reveals the identities of other (honest) reviewers to the colluding author. Then outside the review system, the author pressures one or more of the honest reviewers to accept the proposal or paper.

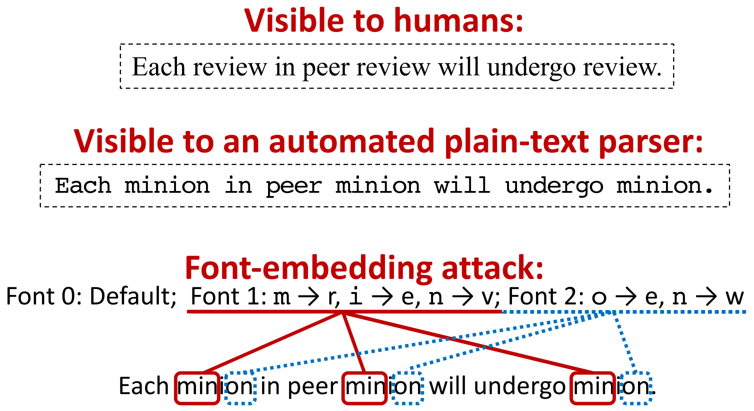

Visible to humans:
Each review in peer review will undergo review.
Visible to an automated plain-text parser:
Each minion in peer minion will undergo minion.
Font-embedding attack:
Font 0: Default; Font 1: m → r, i → e, n → v; Font 2: o → e, n → w
Each minion in peer minion will undergo minion.

Figure 6: An vulnerability in the assignment system, exposed in [Mar+17; TJ19], via attacks that exploit font embedding in the PDF of the submitted paper. Suppose the colluding reviewer has the word “minion” as most frequently occurring in their previous papers, whereas the paper submitted by the colluding author has “review” as most commonly occurring. The author creates two new fonts that map the plain text to rendered text as shown. The author then chooses fonts for each letter in the submitted paper in such a manner that the word “minion” in plain text renders as “review” in the PDF. A human reader will now see “review” but an automated parser will read “minion”. The submitted paper will then be assigned to the target reviewer by the assignment system, whereas no human reader will see “minion” in the submitted paper.

accomplished by targeting the font embedding in the PDF, as illustrated in Figure 6. Empirical evaluations on the reviewer-assignment system used at the International Conference on Computer Communications (INFOCOM) demonstrate the high efficacy of these attacks by being able to get papers matched to target reviewers.

A second line of attacks on automated text matching [Eis+23; HRS+24] uses ideas from adversarial machine learning. In particular, the proposed attack carefully changes the wording of the text of the submitted paper in order to manipulate the text similarities with the reviewers in a desired manner. The two aforementioned vulnerabilities discovered by security researchers suggests the possibility of other novel attacks in practice that may be used by malicious participants beyond what program chairs and security researchers have found to date.

In some cases, the colluding reviewers may naturally be assigned to the target papers without any manipulation of the assignment process [Vij20a]: *“They exchange papers before submissions and then either bid or get assigned to review each other’s papers by virtue of having expertise on the topic of the papers.”*

The next defence we discuss imposes geographical diversity among reviewers of any paper, thereby mitigating collusions occurring among geographically co-located individuals. The paper [Jec+20] considers reviewers partitioned into groups, and designs algorithms which ensures that no paper be assigned multiple reviewers from the same group. The AAAI 2021 conference imposed a related (soft) constraint that each paper should have reviewers from at least two different continents [LBM21].

The final defense we discuss [Jec+20] makes no assumptions on the nature of manipulation, and uses randomized assignments to mitigate the ability of participants to conduct such dishonest behavior. Here the program chairs specify a value between 0 and 1. The randomized assignment algorithm chooses the best possible assignment subject to the constraint that the probability of assigning any reviewer to any paper be at most that value. (The algorithm also allows to customize the value for each individual reviewer-paper pair.) The upper bound on the probability of assignment leads to a higher chance that an independent reviewer will be assigned to any paper, irrespective of the manner or magnitude of manipulations by dishonest reviewers.[3] Naturally, such a randomized assignment may also preclude honest reviewers with appropriate expertise

[3] This assignment procedure also mitigates potential “torpedo reviewing” [Lan12a] where a reviewer intentionally tries to get assigned a paper to reject it, possibly because it is a competing paper or if it is from an area the reviewer does not like. Also interestingly, in the SIGCOMM 2006 conference, the assignments were done randomly among the reviewers who were qualified in the paper topic area to “improve the confidence intervals” [And09] of the evaluation of any paper.

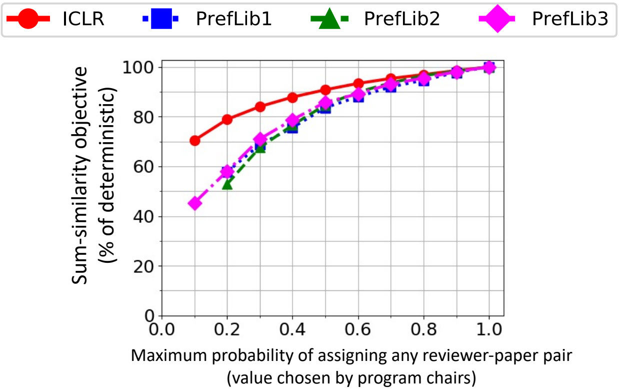


Figure 7: Trading off the quality of the assignment (sum similarity on y-axis) with the amount of randomness (value specified by program chairs on x-axis) to mitigate dishonest collusions [Jec+20]. The similarity scores for the "ICLR" plot are reconstructed [Xu+19] via text-matching from the International Conference on Learning Representations (ICLR conference) 2018 which had 911 submissions. The "Preflib" plots are computed on bidding data from three small-sized conferences (with 54, 52 and 176 submissions), obtained from the Preflib database [MW13].

from getting assigned. Consequently, the program chairs can choose the probability values at run-time by inspecting the tradeoff between the amount of randomization and the quality of the assignment (Figure 7). A subsequent randomized algorithm is available in [Xu+24]. Such randomized assignment defences have been used in AAAI, KDD and NeurIPS conferences.

An attack on this defense – discovered recently to occur in practice – involves colluders forming their colluding agreements after they are assigned papers to review.

There are various tradeoffs between the aforementioned approaches, discussed in [Jec+22b]. Designing algorithms to detect or mitigate such dishonest behavior in peer review is an emerging area of research, with a number of technical problems yet to be solved. This direction of research is however hampered by the lack of publicly available information or data about dishonest behavior. To this end, a small-scale dataset from a controlled experiment is available in [Jec+23].

The recent discoveries of dishonest behavior also pose important questions of law, policy, and ethics for dealing with such behavior: Should algorithms be allowed to flag "potentially malicious" behavior? Should any human be able to see such flags, or should the assignment algorithm just disable suspicious bids? How should program chairs deal with suspicious behavior, and what constitutes appropriate penalties? A case that has led to widespread debate is an ACM investigation [ACM21] which banned certain guilty parties from participating in ACM venues for several years without publicly revealing the names of all guilty parties. Furthermore, some conferences only impose the penalty of rejection of a paper if an author is found to indulge in dishonest behavior including blatant plagiarism. This raises concerns of lack of transparency [Fal+21], and that guilty parties may still participate and possibly continue dishonest behavior in other conferences or grant reviews. Note that such challenges of reporting improper conduct and having action taken are not unique to computer science [Hil21; Els21].

4.3 Other issues

We now enumerate a number of other issues of dishonesty in peer review.

- **Identity theft:** OpenReview, the prominent platform for peer review in machine learning and artificial intelligence, uncovered 94 fake reviewer profiles. Unlike earlier cases involving author-suggested reviewers [FMO14; Coh+16], these incidents occurred despite disallowing authors to suggest reviewers. Most concerningly, the fake accounts were created using verified email addresses from reputable domains. The dishonest researcher posing as the fake reviewer would then try to get assigned the papers submitted by the researchers' genuine identity (using techniques discussed in Section 4.2 to get these papers accepted. The article [Sha+25] details how the scheme operated and proposes several safeguards.

- **Submit and switch:** We discuss an incident in computer science that is not documented elsewhere. This incident involved an author who wrote a paper and wished to get it accepted to a certain conference. The author took somebody else's unpublished paper from the preprint server arXiv (`arxiv.org`), and submitted it as their own paper to a conference (with possibly some changes to prevent discovery of the arXiv version via online search). This submitted paper got accepted. Subsequently when submitting the final version of the paper, the author switched the submitted version with the author's own paper. And voila the author's paper got accepted to the conference! How did this author get caught? The title of the (illegitimate) submission was quite different from what would be apt for their own paper. The author thus tried to change the title in the final version of the paper, but the program chairs had instated a rule that any changes in the title must individually be approved by the program chairs. The author thus contacted the program chairs to change the title, and then the program chairs noticed the inconsistency.

- **Selling submission slots:** Many computer science conferences require authors to submit an abstract by an initial deadline, followed by the full paper about a week later. Submitting an abstract is mandatory to submit a full paper and get it reviewed. This two-step process gives organizers time to perform administrative tasks. Some conferences allow changes to the title, author list, and abstract between deadlines. As a result, a black market has emerged where individuals submit numerous placeholder abstracts and sell them to unscrupulous researchers who missed the abstract deadline but still want to submit a paper.

- **Selling authorship** [Hvi13]

- **Bribing journal editors** [Joe24]

- **Plagiarism of papers** [MADV05; HE15]

- **Plagiarism of reviews** [Pin+24]

- **Data fabrication** [Woo16; Fan09; AM+05; TVS14]

- **Fake paper mills** [EVN21]

- **Multiple submissions** [Bow99]

- **Stealing confidential information from grant proposals submitted for review** [TW18; Mur20]

- **Breach of confidentiality** [RGFP08; Ras+24a]

- **Other issues** [MADV05; ER17; FW17].

5 Miscalibration

Reviewers are often asked to provide assessments of papers in terms of ratings, and these ratings form an integral part of the final decisions. However, it is well known [Mit+11; Fre+03; Sie91; Rag+13; AS12; GB08; Har+09] that the same rating may have different meanings for different individuals: *"A raw rating of 7 out of 10 in the absence of any other information is potentially useless"* [Mit+11]. In the context of peer review, some reviewers are lenient and generally provide high ratings whereas some others are strict and rarely give high ratings; some reviewers are more moderate and tend to give borderline ratings whereas others provide ratings at the extremes; etc.

Miscalibration causes arbitrariness and unfairness in the peer-review process [Sie91]: *"the existence of disparate categories of reviewers creates the potential for unfair treatment of authors. Those whose papers are sent by chance to assassins/demoters are at an unfair disadvantage, while zealots/pushovers give authors an unfair advantage."*

Miscalibration may also occur if there is a mismatch between the conference's overall expectations and reviewers' individual expectations. As a concrete example, the NeurIPS 2016 conference asked reviewers to rate papers according to four criteria on a scale of 1 through 5 (where 5 is best), and specified an expectation regarding each value on the scale. However, as shown in Table 1, there was a significant difference between the

	1 (low or very low)	2 (sub-standard)	3 (poster level: top 30%)	4 (oral level: top 3%)	5 (award level: top 0.1%)
Impact	6.5 %	36.1 %	45.7 %	10.5 %	1.1 %
Quality	6.7 %	38.0 %	44.7 %	9.5 %	1.1 %
Novelty	6.4 %	34.8 %	48.1 %	9.7 %	1.1 %
Clarity	7.1 %	28.0 %	48.6 %	14.6 %	1.8 %

Table 1: Distribution of review ratings in NeurIPS 2016 [Sha+18]. The column headings contain the guidelines provided to reviewers.

expectations and the ratings given by reviewers [Sha+18]. For instance, the program chairs asked reviewers to give a rating of 3 or better if the reviewer considered the paper to lie in the top 30% of all submissions, but the actual number of reviews with the rating 3 or better was nearly 60%. Eventually the conference accepted approximately 22% of the submitted papers.

A frequently-discussed problem that contrasts with the aforementioned general leniency of reviewers is that of “hypercriticality” [Var10; Win11]. Hypercriticality refers to tendency of reviewers to be extremely harsh. This problem is found particularly prevalent in computer science, for instance, with proposals submitted to the computer science directorate of the U.S. National Science Foundation (NSF) receiving reviews with ratings about 0.4 lower (on a 1-to-5 scale) than the average NSF proposal. Another anecdote [Nau10] pertains to the Special Interest Group on Management of Data (SIGMOD) 2010 conference where, out of 350 submissions, there was only one paper with all reviews “accept” or higher, and only four papers with average review of “accept” or higher.

There are other types of miscalibration as well. For instance, an analysis of several conferences [Rag+13] found that the distribution across the rating options varies highly with the scale used. For instance, in a conference that used options $\{1, 2, 3, ..., 10\}$ for the ratings, the amount of usage of each option was relatively smooth across the options. On the other hand, in a conference that used options $\{1, 1.5, 2, 2.5, ..., 5\}$, the “.5” options were rarely used by the reviewers.

There are two popular approaches towards addressing the problem of miscalibration of individual reviewers. The first approach [Fla+10; RRS11; GWG13; Pau81; BK13; Spa+14; Mac+17] is to make simplifying assumptions on the nature of the miscalibration, for instance, assuming that miscalibration is linear or affine. Most works taking this approach assume that each paper p has some “true” underlying rating θ_p, that each reviewer r has two “miscalibration parameters” $a_r > 0$ and b_r, and that the rating given by any reviewer r to any paper p is given by

$$a_r\theta_p + b_r + \text{noise}.$$

These algorithms then use the ratings to estimate the “true” paper ratings θ, and possibly also reviewer parameters.[4]

The simplistic assumptions described above are frequently violated in the real world [BGK05; GB08]; see Figure 8 for an illustration. Algorithms based on such assumptions were tried in some conferences, but based on manual inspection by the program chairs, were found to perform poorly. For instance: *“We experimented with reviewer normalization and generally found it significantly harmful”* in ICML 2012 [Lan12b].

One exception to the simplistic-modeling approach is the paper [Tan+21] which considers more general forms of miscalibration. In more detail, it assumes that the rating given by reviewer r to any paper p is given by $f_r(\theta_p + \text{noise})$, where f_r is a function that captures the reviewer’s miscalibration and is assumed to lie in certain specified classes. Their algorithm then finds the values of θ_p and f_r which best fit the review data.

A second popular approach [Mit+11; Fre+03; Rok68; Har+09; AS12; NOS12] towards handling miscalibrations is via rankings: either ask reviewers to give a ranking of the papers they are reviewing (instead of providing ratings), or alternatively, use the rankings obtained by converting any reviewer’s ratings into a ranking of their reviewed papers. Using rankings instead of ratings *“becomes very important when we*

[4] The paper [CL21] considers this model but assumes $a_r = 1$, treats the noise term as the reviewer’s subjective opinion, and estimates $\theta_p + \text{noise}$ as a calibrated review score.

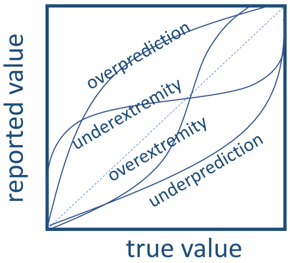


Figure 8: A caricature of a few types of miscalibration [BGK05]. The diagonal line represents perfect calibration. An affine (or linear) miscalibration would result in a straight line.

combine the rankings of many viewers who often use completely different ranges of scores to express identical preferences" [Fre+03].

Ratings can provide some information even in isolation. It was shown recently [WS19b] that even if the miscalibration is arbitrary or adversarially chosen, unquantized ratings can yield better results than rankings alone. While the algorithms designed in the paper [WS19b] are largely of theoretical interest, we note that their guarantees are based on randomized decisions.[5]

Rankings also have their benefits. In NeurIPS 2016, out of all pairs of papers reviewed by the same reviewer, the reviewer gave an identical rating to both papers for 40% of the pairs [Sha+18]. In such situations, rankings can help break ties among these papers, and this approach was followed in the ICML 2021 conference. A second benefit of rankings is to check for possible inconsistencies. For instance, the NeurIPS 2016 conference elicited rankings from reviewers on an experimental basis. They then compared these rankings with the ratings given by the reviewers. They found that 96 (out of 2425) reviewers had rated some paper as strictly better than another on all four criteria, but reversed the pair in the overall ranking [Sha+18]. Given the tradeoffs between rankings and ratings, the papers [PE22; Liu+22] develop methods to exploit benefits of both rankings and ratings by eliciting and then combining these two forms of data.

Addressing miscalibration in peer review is a wide-open problem. The small per-reviewer sample sizes due to availability of only a handful of reviews per reviewer is a key obstacle: for example, if a reviewer reviews just three papers and gives low ratings, it is hard to infer from this data as to whether the reviewer is generally strict. This impediment calls for designing protocols or privacy-preserving algorithms [Din+22] that allow conferences to share some reviewer-specific calibration data with one another in order to calibrate better.

6 Subjectivity

A number of issues pertaining to reviewers' personal subjective preferences exist in peer review. We begin with a discussion on commensuration bias towards which several approaches have been proposed, including a computational mitigating technique. We then discuss other issues pertaining to subjectivity which may benefit from the design of computational mitigating methods and/or human-centric approaches of better reviewer guidelines and training.

6.1 Commensuration bias

Program chairs of conferences often provide criteria to reviewers for judging papers. However, different reviewers have different, subjective opinions about the relative importance of various criteria in judging

[5] Interestingly, randomized decisions are used in practice by certain funding agencies to allocate grants [Liu+20; Cha21; Hey+22]. Such randomized decision-making has found support among researchers [Phi21] as long as it is combined with the peer review process and is not pure randomness. Identified benefits of such randomization include overcoming ambiguous decisions for similarly-qualified proposals, decreasing reviewer effort, circumventing old-boys' networks, and increasing chances for unconventional research [Phi21].

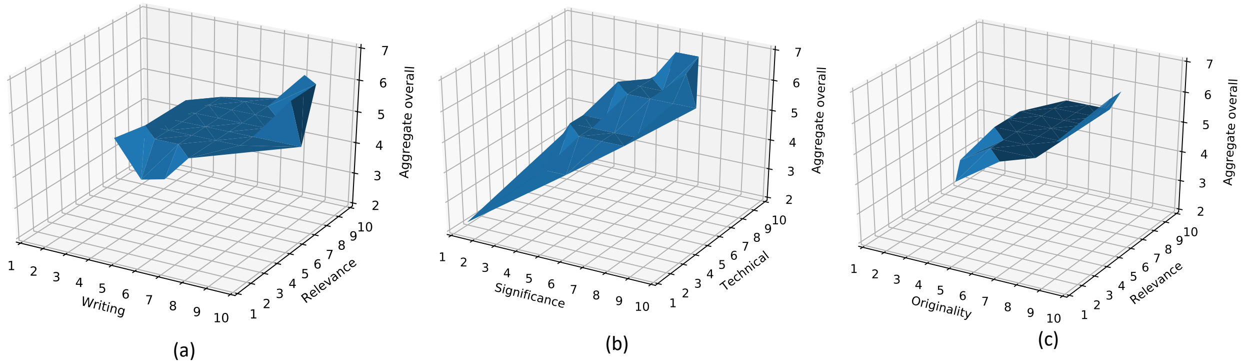


Figure 9: Mapping of individual criteria to overall ratings by reviewers in IJCAI 2017 [NSP21]. The conference used five criteria, and hence the mapping is five-dimensional. The figure plots representative two-dimensional cross sections of the mapping of the following pairs of criteria to overall ratings: (a) writing and relevance, (b) significance and technical quality, and (c) originality and relevance.

papers [Mah77; Chu05; Lam09; BMS87; HO21]. The overall evaluation of a paper then depends on the individual reviewer's preference on how to aggregate the evaluations on the individual criteria. This dependence on factors exogenous to the paper's content results in arbitrariness in the review process. On the other hand, in order to ensure fairness, all (comparable) papers should be judged by the same yardstick. This issue is known as "commensuration bias" [Lee15].

As a toy example, suppose three reviewers consider empirical performance of any proposed algorithm as most important, whereas most others highly regard novelty. Then a novel paper whose proposed algorithm has a modest empirical performance is rejected if reviewed by these three reviewers but would have been accepted by any other set of reviewers. Concretely, as revealed in a survey of reviewers [KTP77], more than 50% of reviewers say that even if the community thinks a certain characteristic of a manuscript is good, if the reviewer's own opinion is negative about that characteristic, it will count against the paper; about 18% say this can also lead them to reject the paper. The paper's fate thus depends on the subjective preference of the assigned reviewers.

The program chairs of the AAAI 2013 conference recognized this problem of commensuration bias. With an admirable goal of ensuring a uniform policy of how individual criteria are aggregated into an overall recommendation across all papers and reviewers, they announced specific rules on how reviewers should aggregate their ratings on the 8 criteria into an overall rating. The goal was commendable, but unfortunately, the proposed rules had shortcomings. For example [NSP21], on a scale of 1 to 6 (where 6 is best), one rule required giving an overall rating of "strong accept" if a paper received a rating of 5 or 6 for some criterion, and did not get a 1 for any criteria. This may seem reasonable at first, but looking at it more carefully, it implies a strong acceptance for any paper that receives a 5 for the criterion of clarity, but receives a low rating of 2 in every other criterion. More generally, specifying a set of rules for aggregation of 8 criteria amounts to specifying an 8-dimensional function, which can be challenging to craft by hand.

Due to concerns about commensuration bias, the NeurIPS 2016 conference did not ask reviewers to provide any overall ratings. A similar recommendation has been made in the natural language processing community [RA20]. NeurIPS 2016 instead asked reviewers to only rate papers on certain criteria and left the aggregation to meta reviewers. This approach can however lead to arbitrariness due to the differences in the aggregation approaches followed by different meta reviewers.

Noothigattu et al. [NSP21] propose an algorithmic solution to this problem. They consider an often-recommended [OTD07; MWC19; KSS21; CL21; Mar+22] interface that asks reviewers to rate papers on a pre-specified set of criteria alongside their overall rating. Commensuration bias implies that each reviewer has their own subjective mapping of criteria to overall ratings. The key idea behind the proposed approach is to use machine learning and social choice theory to learn how the body of reviewers—at an aggregate level—map criteria to overall ratings. The algorithm then applies this learned mapping to the criteria ratings in each review in order to obtain a second set of overall ratings. The conference management system would then augment the reviewer-provided overall ratings with those computed using the learned mapping, with the primary benefit that the latter ratings are computed via the same mapping for all (comparable) papers. This method was used in the AAAI 2022 conference to identify reviews with significant commensuration bias.

The aforementioned method [NSP21] can also be used to understand the reviewer pool's emphasis on various criteria. As an illustration, the mapping learned via this method from the International Joint Conference on Artificial Intelligence (IJCAI conference) 2017 is shown in Figure 9. Observe that interestingly, the criteria of significance and technical quality have a high (and near-linear) influence on the overall recommendations whereas writing and relevance have a large plateau in the middle. A limitation of this approach is that it assumes that reviewers first think about ratings for individual criteria and then merge them to give an overall rating; in practice, however, some reviewers may first arrive at an overall opinion and reverse engineer ratings for individual criteria that can justify their overall opinion.

6.2 Confirmation bias and homophily

A randomized controlled trial by Mahoney [Mah77] asked reviewers to each assess a fictitious manuscript. The contents of the manuscripts sent to different reviewers were identical in their reported experimental procedures but differed in their reported results. The study found that in terms of the overall scores, reviewers were strongly biased against papers with results that contradicted the reviewers' own prior views. Interestingly, the difference in the results section also manifested in other aspects: a manuscript whose results agreed with the reviewer's views was more likely to be rated as methodologically better and as having a better data presentation, even though these components were identical across the manuscripts.[6] Confirmation biases have also been found in subsequent studies [ER94; Koe93].

A related challenge is that of "homophily," that is, reviewers often favor topics which are familiar to them [PR85; Don+19; Li17]. For instance, a study [PR85] found that "Where reviewer and [submission] were affiliated with the same general disciplinary category, peer ratings were better (mean = 1.73 [lower is better]); where they differed, peer ratings were significantly worse (mean = 2.08; p = 0.008)". According to [TC91], reviewers "simply do not fight so hard for subjects that are not close to their hearts". In contrast, the paper [Bou+16] ran a controlled study where they observed an opposite effect that reviewers gave lower scores to topics closer to their own research areas.

6.3 Acceptance via obfuscation ("Dr. Fox effect")

A controlled study [Arm80] asked evaluators to each rate one passage, where the readability of these passages was varied across reviewers but the content remained the same. The study found that the passages which were harder to read were rated higher in research competence. The author cheekily concludes with the phrase "If you can't convince them, confuse them."

6.4 Surprisingness and hindsight bias

One criteria that reviewers often use in their judgment of a paper is the paper's informativeness or surprisingness. Anecdotally, it is not uncommon to see reviews criticizing a paper as "the results are not surprising." But are the results as unsurprising as the reviewers claim them to be? Slovic and Fischhoff [SF77] conducted a controlled study to investigate reviewers' perceptions of surprisingness. They divided the participants in the study randomly into two groups: a "foresight" group and a "hindsight" group. Each participant in the foresight group was shown a fictitious manuscript which contained the description of an experiment but not the results. There results could take two possible values. Each participant in the hindsight group were shown the manuscript containing the description as well as the result. The result of the manuscript shown to any participant was chosen randomly as one of the two possible values. The foresight participants were then asked to assess how surprising each of the two possible results would seem were they obtained, whereas the hindsight subjects were asked to assess the surprisingness of the result obtained.

[6] According to Mahoney [PC82], for this study, "the emotional intensity and resistance of several participants were expressed in the form of charges of ethical misconduct and attempts to have me fired. Several editors later informed me that correspondence from my office was given special scrutiny for some time thereafter to ascertain whether I was secretly studying certain parameters of their operation."

The study found that the participants in the hindsight group generally found the results less surprising than the foresight group. The hindsight subjects also found the study as more replicable. There is thus a downward bias in the perception of surprisingness when a reviewer has read the results, as compared to what they would have prior to doing so. The study also found that the difference between hindsight and foresight reduces if the hindsight participants are additionally asked a counterfactual question of what they would have thought had the reported result been different. Slovic and Fischhoff thus suggest that when writing manuscripts, authors may stress the unpredictability of the results and make the reader think about the counterfactual.

6.5 Hindering novelty

Peer review is said to hinder novel research [Chu05]: *"Reviewers love safe (boring) papers, ideally on a topic that has been discussed before (ad nauseam)... The process discourages growth"*. Naughton makes a noteworthy point regarding one reason for this problem: *"Today reviewing is like grading: When grading exams, zero credit goes for thinking of the question. When grading exams, zero credit goes for a novel approach to solution. (Good) reviewing: acknowledges that the question can be the major contribution. (Good) reviewing: acknowledges that a novel approach can be more important than the existence of the solution"* [Nau10].

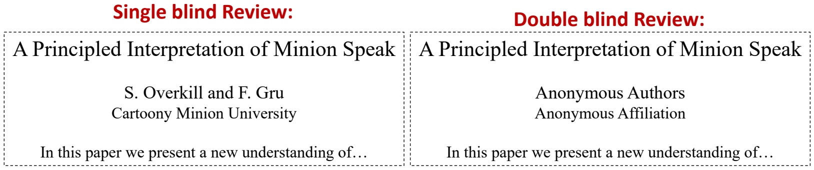
The paper [Bou+16] presents an evaluation of the effects of novelty of submitted grant proposals on the reviews. A key question in conducting such an evaluation is how to define novelty? This study defines novelty in terms of the combination of keywords given by a professional science librarian (not affiliated with the authors) to each submission, relative to the literature. They find a negative relationship between review scores and novelty. Delving deeper, they find that this negative relationship is largely driven by the most novel proposals. On the other hand, at low levels of novelty they observe an increase in scores with an increase in novelty.

The paper [REG00] performs a randomized controlled trial where reviewers were shown one of two versions of a fictitious manuscript whose content pertained to a treatment of obesity. One version claimed to study an orthodox drug and the other version claimed to study an unconventional drug; the two manuscripts were identical otherwise. With a focus on the reviewer's ratings of the manuscript on the criteria of "importance," they observed a significant difference in favor of the orthodox with an odds ratio of 3.01.

The study [Tep+22] examined the reviews of manuscripts submitted to 49 journals in the life and physical sciences. They defined the 'novelty' of a manuscript as being high if the manuscript cited multiple journals that were not conventionally cited together. Under linear modeling assumptions, they found no evidence of reviewer bias against novelty.

6.6 Positive-outcome bias

A positive-outcome bias pertains to the peer review of scientific studies where studies with positive outcomes are more likely to be accepted than those with negative outcomes. A study [Eme+10] investigated the existence of a positive-outcome bias via a randomized controlled trial. The authors of this study created a fictitious manuscript with two versions: the two versions were identical except that one version had a positive outcome (that is, the data showed a difference between two conditions being tested) and the other version had a negative outcome (that is, the data did not show such a difference). They sent one of the two versions at random to each of over 200 reviewers. They found that 97.3% of the reviews of the positive-outcome version recommended acceptance, whereas the acceptance rate was only 80.0% for the negative-outcome version. The authors had also deliberately injected errors into the fictitious manuscript, and they found that reviewers detected roughly twice as many errors in the negative-outcome version. Finally, they asked reviewers to evaluate the methods in the paper (which were identical in the two version) and found that reviewers gave significantly higher scores to the methods in the positive-outcome version. This controlled study thus does find evidence of a positive-outcome bias. As a consequence of this bias, some venues solicit papers with only the study question and methods but without the results, and the acceptance decision of the paper is evaluated based on this information [Smu13].



Single blind Review:

A Principled Interpretation of Minion Speak

S. Overkill and F. Gru
Cartoony Minion University

In this paper we present a new understanding of…

Double blind Review:

A Principled Interpretation of Minion Speak

Anonymous Authors
Anonymous Affiliation

In this paper we present a new understanding of…

Figure 10: An illustration of a paper as seen by a reviewer under single blind versus double blind peer review.

6.7 Interdisciplinary research

Interdisciplinary research is considered a bigger evaluation challenge, and at a disadvantage, as compared to disciplinary research [PR85; TC91; Lau06; Lam09; Huu10; Ano13; PF17; Don+19; Fro21]. There are various reasons for this (in addition to algorithmic challenges discussed in Section 3). First, it is often hard to find reviewers who individually have expertise in each of the multiple disciplines of the submission [PR85; Fro21]. Second, if there are such reviewers, there may be only a few in that interdisciplinary area, thereby "leading to dangers of inbreeding" [PR85]. Third, reviewers often favor topics that are familiar to them ("homophily" discussed in Section 6.2). For disciplinary reviewers, the other discipline of an interdisciplinary paper may be unfamiliar. Fourth, if a set of reviewers is chosen simply to ensure "coverage" where there is one reviewer for each discipline in the submission, then each reviewer has a veto power because their scientific opinions cannot be challenged by other reviewers [Lau06]. Moreover, a multidisciplinary review team can have difficulties reconciling different perspectives [Lau06]. A fifth challenge is that of expectations. To evaluate interdisciplinary research, the "most common approach is to prioritize disciplinary standards, premised on the understanding that interdisciplinary quality is ultimately dependent on the excellence of the contributing specialized component" [Huu10]. Consequently, "interdisciplinary work needs to simultaneously satisfy expert criteria in its disciplines as well as generalist criteria" [Lam09].

In order to mitigate these issues in evaluating interdisciplinary proposals, program chairs, meta-reviewers and reviewers can be made aware of these issues in evaluating interdisciplinary research. One should try, to the extent possible, to assign reviewers that individually span the breadth of the submission [PR85]. In cases where that is not possible, one may use computational tools (Section 3) to inform meta-reviewers and program chairs of submissions that are interdisciplinary and the relationship of reviewers to the submission (e.g., that reviewers as a whole cover all disciplines of the paper, but no reviewer individually does so). The criteria of acceptance may also be reconsidered: program chairs and meta-reviewers sometimes emphasize accepting a paper only when at least one reviewer champions it (and this may naturally occur in face-to-face panel discussions where a paper is favored only if some panelist speaks up for it) [Nie00]. The aforementioned discussion suggests this approach will disadvantage interdisciplinary papers [PR85]. Instead, the decisions should incorporate the bias that reviewers in any individual discipline are less likely to champion an interdisciplinary paper than a paper of comparable quality that is fully in their own discipline.

7 Biases pertaining to author identities

In 2015, two women researchers, Megan Head and Fiona Ingleby submitted a paper to the PLOS ONE journal. A review they received read: *"It would probably be beneficial to find one or two male researchers to work with (or at least obtain internal peer review from, but better yet as active co-authors)"* [Ber15]. This is an example of how a review can take into consideration the authors' identities even when we expect it to focus exclusively on the scientific contribution.

Such biases with respect to author identities are widely debated in computer science and elsewhere. These debates have led to two types of peer-review processes: single-blind reviewing where reviewers are shown authors' identities, and double-blind reviewing where author identities are hidden from reviewers (see Figure 10 for an illustration). In both settings, the reviewer identities are not revealed to authors.

A primary argument against single-blind reviewing is that it may cause the review to be biased with respect to the authors' identities. On the other hand, arguments against double blind include: effort to make a manuscript double blind, efficacy of double blinding (since many manuscripts are posted with author identities on preprint servers and social media), hindrance in checking (self-)plagiarism and conflicts of

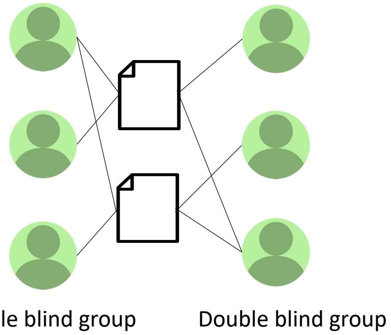


Figure 11: The WSDM 2017 experiment [TZH17] comparing single and double blind reviewing.

interest, and the use of author identities as a guarantee of trust for the details that reviewers have not been able to check carefully. We refer the reader to [Sha22b, Section 4] for a variety of arguments for and against anonymizing authors by researchers in theoretical computer science.

7.1 Studies in computer science

The debate over single-vs-double blind reviewing often rests on the frequently-asked question: "Where is the evidence of bias in single-blind reviewing in my field of research?" There are several experiments that provide a quantitiative answer to this question.

In the conference-review setting, a remarkable experiment was conducted at the Web Search and Data Mining (WSDM) 2017 conference [TZH17] which had 500 submitted papers and 1987 reviewers. The reviewers were split randomly into two groups: a single-blind group and a double-blind group. Every paper was assigned two reviewers each from both groups (see Figure 11). This experimental design allowed for a direct comparison of single blind and double blind reviews for each paper without requiring any additional reviewing for the purpose of the experiment. The study found a significant bias in favor of famous authors, top universities, and top companies. Moreover, it found a non-negligible effect size but not statistically significant bias against papers with at least one woman author; the study also included a meta-analysis combining other studies, and this meta-analysis found this gender bias to be statistically significant. The study did not find evidence of bias with respect to papers from the United States, nor when reviewers were from the same country as the authors, nor with respect to academic (versus industrial) institutions. The WSDM conference moved to double-blind reviewing the following year.

Another study [MS21] did not involve a controlled experiment, but leveraged the fact that the ICLR conference switched from single blind to double blind reviewing in 2018. Analyzing both reviewer-provided ratings and the text of reviews, the study found evidence of bias with respect to the affiliation of authors but not with respect to gender.[7]

Finally, studies [WS19a; Mat+20; FK20a] have found a significant gender skew in terms of representation in computer science conferences.

7.2 Studies outside computer science

These results augment a vast body of literature in various scientific fields outside of computer science investigating biases pertaining to author identities. The study [BMD07] finds gender bias, [WW01] finds biases with respect to gender and personal connections, the studies [Gin+11; Nak+21] find bias with respect to race, whereas the study [For+19] finds little evidence of gender or racial bias. Several studies [Oki+16; Bla91; Ros+06; Gar+94] find bias in favor of authors' status. In particular, [Gar+94] observes a significant bias for brief reports but not for major papers. This observation suggests that reviewers tend to use author characteristics more when less information about the research is available. The study [Nie+21] finds weak evidence of country and institution bias when scientists evaluate abstracts. Bias with respect to author fame is also investigated in the paper [Bla91], which finds that the top and bottom institutions' papers unaffected,

[7] We refer the reader to the paper [Kuz+24] for a survey of literature on natural language processing for peer review.

but those in the middle were affected. In a similar vein, the study [Fou+00] suggests that "evaluation of absolutely outstanding articles will not be biased, but articles of ambiguous merit may be judged based on the author's gender." A randomized controlled trial [FFS94] found that authors with more past papers were given better scores by blinded reviewers. The paper [Bud+08] finds an increased representation of women authors following a policy change from single to double blind. The study [GW99] finds that blinding reviewers to the author's identity does not usefully improve the quality of reviews. The study [Hub+22] found a significant preference of reviewers towards a manuscript that had the name of a Nobel laureate as author than an identical manuscript with a junior researcher's name. Surveys of researchers [War08; MHR13] reveal that double blind review is preferred and perceived as most effective.

7.3 Design of experiments

The experiments on biases have also prompted a focus on careful design of experimental methods and measurement algorithms to evaluate biases in peer review, while mitigating confounding factors that may arise due to the complexity of the peer-review process. For instance, an investigation [MD06] of bias with respect to authors' fame in the SIGMOD conference did not reveal bias, but subsequently an analysis on the same dataset using the same methods except for using medians instead of means revealed existence of fame biases [Tun06]. The paper [SSS19] discusses some challenges in the methods employed in the aforementioned WSDM experiment and provides a framework for design of such experiments. The paper [Jec+22a] considers the splitting of the reviewer pool in two conditions in terms of the tradeoff between experimental design and the assignment quality. A uniform random split of reviewers is natural for experimental design, they find that such a random split is also nearly optimal in terms of the assignment quality as compared to any other way of splitting the reviewer pool.

7.4 Use of author identities in non-anonymized review

Proponents of single-blind reviewing state various uses of author identities in the reviewing, such as using author identities for trust in hard-to-verify mathematical proofs or wacky ideas, or to add extra scrutiny for papers by authors with a history of erroneous papers. The Innovations in Theoretical Computer Science (ITCS) 2023 conference conducted an experiment [Sha22b] where author identities were initially hidden from reviewers, and were subsequently made visible after they submitted their initial review. Reviewers were then allowed to change their reviews, and the study then analyzed the change in the reviews. They found that the amount of change was quite small: only 7.1% reviews changed their overall scores.

7.5 De-anonymization of authors in double blind

Making reviewing double blind can mitigate these biases, but may not fully eliminate them. A study [Sha22b] in the ITCS 2023 conference found that more than half of the reviewers reported having "no idea" about the identities of the authors. Reviewers in three double-blind conferences were asked to guess the authors of the papers they were reviewing [Le +18]. The reviewers were asked to provide this information separately with their reviews, and this information would be visible only to the program chairs. No author guesses were provided alongside 70%-86% of the reviews (it is not clear whether an absence of a guess indicates that the reviewer did not have a guess or if they did not wish to answer the question). However, among those reviews which did contain an author guess, 72%-85% guessed at least one author correctly.

In many research communities, it is common to upload papers on preprint servers such as arXiv before it is reviewed. For instance, 54% of all submissions to the NeurIPS 2019 conference were posted on arXiv and 21% of these submissions were seen by at least one reviewer [Bey+19]. These preprints contain information about the authors, thereby potentially revealing the identities of the authors to reviewers.

In a survey by two double-blind conferences — the ACM Economics and Computation (EC) 2021 conference and the ICML 2021 conference — over a third of its reviewers (anonymously) reported that they had actively searched online for the paper they were reviewing [Ras+22]. Furthermore, the study [Ras+22] also found that better ranks of the authors' affiliations were weakly correlated with visibility of a preprint (to reviewers who did not search for it online).

Based on these observations, one may be tempted to disallow authors from posting their manuscripts to preprint servers or elsewhere before they are accepted. However, one must tread this line carefully. First, such an embargo can hinder the progress of research. Second, the effectiveness of such prohibition is unclear. Studies have shown that the content of the submitted paper can give clues about the identity of the authors. Several papers [HJP03; CUD19; MS20] design algorithms that can identify author identity or affiliations to a moderate degree based on the content of the paper. The aforementioned survey [Le +18] forms an example where humans could guess the authors. Third, due to such factors, papers by famous authors may still be accepted at higher rates, while disadvantaged authors' papers neither get accepted nor can be put up on preprint servers like arXiv. In fast-moving fields, this could also result in their work being scooped while they await a conference acceptance.

These studies provide valuable quantitative information towards policy choices and tradeoffs on blinded reviewing.

8 Incentives

Ensuring appropriate incentives for participants in peer review is a critical open problem: incentivizing reviewers to provide high-quality reviews and incentivizing authors to submit papers only when they are of suitably high quality.

8.1 Author incentives

It is said that authors submitting a below-par paper have little to lose but lots to gain: very few people will see the below-par version if it gets rejected, whereas the arbitrariness in the peer-review process gives it some chance of acceptance. The rapid increase in the number of submissions in various conferences has prompted policies that incentivize authors to submit papers only when they are of suitably high quality [And09].

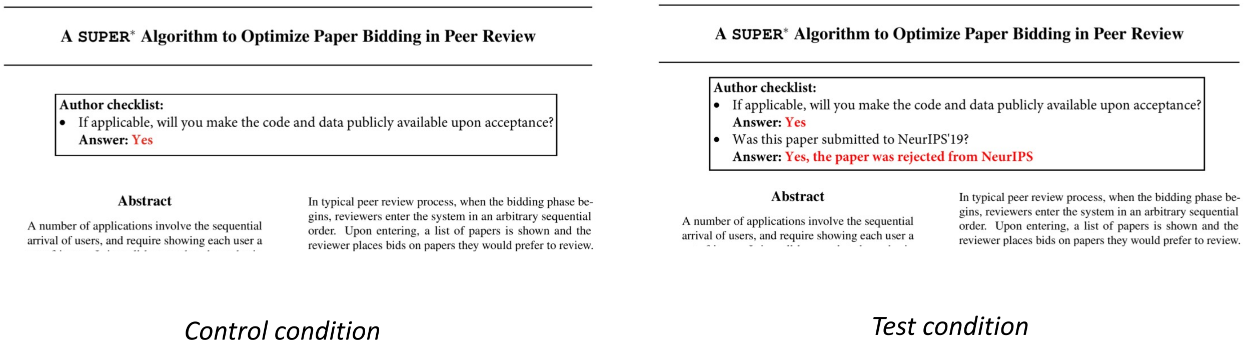
8.1.1 Open Review

Some conferences are adopting an "open review" approach to peer review, where all submitted papers and their reviews (but not reviewer identities) are made public. A prominent example is the OpenReview.net conference management system in computer science. Examples outside computer science include scipost.org, f1000research.com, and eLife where the latter two are among the few venues that also publishe reviewer identities. A survey [SSM13] of participants at the ICLR 2013 conference, which was conducted on OpenReview.net and was one of the first to adopt the open review format, pointed to increased accountability of authors as well as reviewers in this open format. An open reviewing approach also increases the transparency of the review process, and provides more information to the public about the perceived merits/demerits of a paper rather than just a binary accept/reject decision [And09]. Additionally, the public nature of the reviews has yielded useful datasets for research on peer review [Kan+18; MS21; Tra+20; Bha+20; YLN21; MT13; War10].

Alongside these benefits, the open-review format may also result in some challenges such as threats to anonymity of reviewers (see Section 9.4 for more details). We discuss one other issue next, related to public visibility of rejected papers.

8.1.2 Resubmission policies

A non-negligible fraction of papers accepted at top conferences are previously rejected at least once. In a 2017 survey of computer science reviewers, 47.9% reported having previously reviewed 1–2 of the papers assigned to them, 14.6% had already reviewed 3–4 of them, while 37.5% indicated that none of their assigned papers were ones they had reviewed before. A survey of authors of *accepted* papers across many computer science conferences finds that the mean number of prior submissions before acceptance is at least 0.7 [FK20b]. One respondent reported as many as 12 attempts before acceptance. A survey in the biological sciences found that "75% of published articles were submitted first to the journal that would publish them, and high-impact journals published proportionally more articles that had been resubmitted from another journal" [Cal+12].



A SUPER* Algorithm to Optimize Paper Bidding in Peer Review

Author checklist:

- If applicable, will you make the code and data publicly available upon acceptance?
  Answer: Yes

Abstract

A number of applications involve the sequential arrival of users, and require showing each user a

In typical peer review process, when the bidding phase begins, reviewers enter the system in an arbitrary sequential order. Upon entering, a list of papers is shown and the reviewer places bids on papers they would prefer to review.

Control condition

A SUPER* Algorithm to Optimize Paper Bidding in Peer Review

Author checklist:

- If applicable, will you make the code and data publicly available upon acceptance?
  Answer: Yes
- Was this paper submitted to NeurIPS'19?
  Answer: Yes, the paper was rejected from NeurIPS

Abstract

A number of applications involve the sequential arrival of users, and require showing each user a

In typical peer review process, when the bidding phase begins, reviewers enter the system in an arbitrary sequential order. Upon entering, a list of papers is shown and the reviewer places bids on papers they would prefer to review.

Test condition

Figure 12: Experiment to evaluate resubmission bias [Ste+21b]: paper shown to reviewers in the control and test conditions.

To reuse review effort, many conferences are adopting policies where authors of a paper must provide past rejection information along with the submission. For instance, the IJCAI 2020 conference required authors to prepend their submission with details of any previous rejections including prior reviews and the revisions made by authors. While these policies are well-intentioned towards ensuring that authors do not simply ignore reviewer feedback, the information of previous rejection could bias the reviewers.

A controlled experiment [Ste+21b] in conjunction with the ICML 2020 conference tested for such a "resubmission bias" on a population of novice reviewers. Each reviewer was randomly shown one of two versions of a paper to review (Figure 12): one version indicated that the paper was previously rejected at another conference while the other version contained no such information. Reviewers gave almost one point lower rating on a 10-point scale for the overall evaluation of a paper when they were told that a paper was a resubmission. In terms of specific review criteria, reviewers underrated "Paper Quality" the most. The existence of such a resubmission bias has prompted a rethinking of the resubmission-related policies about who (reviewers or meta-reviewers or program chairs) has information about the resubmission and when (from the beginning or after they submit their initial review).

8.1.3 Rolling deadlines

In conferences with a fixed deadline, a large fraction of submissions are made on or very near the deadline [SSM13]. This observation suggests that removing deadlines (or in other words, having a "rolling deadline"), wherein a paper is reviewed whenever it is submitted, may allow authors ample time to write their paper as best as they can before submission, instead of cramming right before the fixed deadline. The flexibility offered by rolling deadlines may have additional benefits such as helping researchers better deal with personal constraints, and allowing a more balanced sharing of resources such as compute (otherwise everyone attempts to use the compute clusters right before the deadline).

The U.S. National Science Foundation experimented with this idea in certain programs [Han16]. The number of submitted proposals reduced drastically from 804 in one year in which there were two fixed deadlines, to just 327 in the subsequent 11 months when there was a rolling deadline. Thus in addition to providing flexibility to authors, rolling deadlines may also help reduce the strain on the peer-review process.

8.1.4 Authors' perspectives

The paper [Su21] presents a novel idea of asking authors to provide a ranking their submitted papers, and using the authors' ranking to "denoise" reviews. However, several challenges remain to make this interesting approach practical. For instance, the method cannot handle disagreements between co-authors [Ras+24b] about the rankings. It can incentivize authors to falsely report their ranking of their own papers, which can in turn lead to poorer quality papers being accepted. Furthermore, this method requires authors to predict the ranking in terms of what reviewers will think, which may add to the stress of submitting papers.

8.1.5 Citation-based incentives

Citations are frequently used in hiring and promotion related activities for researchers, and are often considered as a measure of the prestige of a researcher. This creates an interplay of incentives across both authors and reviewers. In peer review, this has also led to unscrupulous citation-coercion by reviewers: reviewers coaxing authors to (unnecessarily) cite the reviewers' own paper. Two surveys [FW17; RGFP08] of authors found that 14% and 22% of authors respectively report having been coerced to add such citations. In one extreme case [VN20], for many years, the editor of a journal kept asking authors of every submitted paper to add (on average) over 30 citations to the editor's own papers. In computer science, a study [Ste+23c] in the ICML 2020 and EC 2021 conferences found that if a paper happens to cite the reviewer, it receives on average roughly one point higher score on a five or six point scale, even after controlling for various confounders like reviewer expertise, seniority, reviewer preferences, paper quality, and genuine citations.

Naturally, these incentives on the reviewer side create incentives for authors to cite (potential) reviewers. For instance, the survey [FW17] finds that over 40% of authors preemptively pad their papers with non-critical citations for these reasons. Finally, as for the aforementioned extreme example of citation-coercion [VN20], it was found that the authors obeyed the citation requests with "apparently amazing frequency" – the acceptance of their paper was on the line after all.

8.2 Reviewer incentives

We now discuss incentives for reviewers to provide high-quality reviews.

8.2.1 Materialistic and non-materialistic incentives

Many researchers have suggested introducing policies in which reviewers earn materialistic incentives such as points (or possibly money) for reviewing, and these points count for promotional evaluations or can be a required currency to get their own papers reviewed. As with any other real-world deployment, the devil would lie in the details. If not done carefully, an introduction of any such system can significantly skew the motivations for reviewing [NBH16] and lead to other problems.

A range of initiatives have been introduced to encourage a higher volume of reviews. One of the most well-known among these was "Publons," which gained traction across several research communities. Publons rewarded reviewers with points, badges, and also gave out awards to those who completed the most reviews. However, its effectiveness remains debatable, as some worry that it might lead reviewers to chase points by delivering superficial or poor reviews [SAK17]. A study [PDR19] on the Publons platform indicated that the majority of reviews were conducted by comparatively less experienced researchers; for some, peer reviewing might even be their primary research-related activity. In contrast, top-tier researchers were scarcely seen on the leaderboards. In fact, an examination of the top 250 reviewers on Publons revealed that they carried out an astounding average of over 180 reviews annually.

Many venues make it compulsory for qualified submitting authors to also review papers. We are not aware of any studies quantitatively comparing the quality of compulsory versus volunteer reviews of scientific papers. A related application is that of peer evaluation of lab reports in a classroom setting. The study [PBB23] conducted an experiment asking students to provide such evaluations on a compulsory basis in one year and on a voluntary basis in the adjacent year. They found the reviews to be of higher quality when students participated voluntarily.

Squazzoni et al. [SBT13] empirically evaluate the effects of various incentive mechanisms via "investment game" that mimics various characteristics of incentives in peer review. Within this game, they conduct a controlled trial that compares a setting with no payoffs for reviewers, an incentive comprising a fixed payoff for reviewers, and two incentive structures involving a variable payoff for reviewers. They quantitatively find that the no-payoff setting results in the most effective peer review. Surveys of participants point to the trust and cooperation in the no-payoff setting as the key to more effective peer review in this setting in the experiment.

In economics, Chetty et al. [CSS14] provided a monetary incentive of $100 for timely reviews in their journal, and in a randomized controlled trial, they find that it did significantly reduce the latency of review without reducing its quality as compared to a control condition of simply requesting a quicker review. They found a substantial reduction in the latency of review submission. In biomedicine, Cotton et al. [Cot+25] conducted a quasi randomized controlled trial to compare two conditions in their journal: one where each reviewer invitation was accompanied with an offer of $250 for completing the review, and another where no monetary offer was made. It was found that the monetary condition had marginally lower latency in the review submissions, and that editor evaluations of review quality found no difference in quality between the two conditions. Note that such reviews of reviews also come with their own challenges, as we will see in the next section.

Several venues outside of computer science implemented policies whereby authors were financially responsible for compensating reviewers at the time of paper submission. However, this approach did not gain widespread acceptance as authors were seldom willing to pay upon submission rather than on acceptance of their papers [Dav17b; Dav17a].

Among non-materialistic incentives, a survey [NBH16] of researchers in human computer interaction found that the three top motivations for reviewing were: "I want to know what is new in my field," "I receive reviews from the community, so I feel I should review for the community," and "I want to encourage good research."

8.2.2 Reviewing the reviews

An often-made suggestion is to ask meta-reviewers or other reviewers to review the reviewers [Aro+21] in order to allocate points for high-quality reviews. The NeurIPS 2022 conference conducted a randomized controlled trial and other experiments to evaluate the reliability of peer reviewing the peer reviews [Gol+23]. In the randomized controlled trial, evaluators in the control condition were shown the original review to be evaluated, and evaluators in the treatment condition were shown a version of the review that was made longer without any additional information. The experiment found that evaluators are biased positively towards longer reviews. The experiment also found that the amount of inconsistency, miscalibration, and subjectivity in evaluations of reviews is similar or higher than in reviews of papers.

An alternative option is to ask authors to evaluate the reviews. Indeed, one may argue that authors have the best understanding of their papers and that authors have skin in the game. For instance, the Journal of Systems Research (JSys) asks authors to evaluate reviews, and states the policy that reviewers with a history of poor reviews will be removed from the editorial board. Unfortunately, another bias comes into play here: authors are far more likely to positively evaluate a review when the review expresses a positive opinion of the paper. The aforementioned NeurIPS 2022 experiment [Gol+23] finds significant evidence of this bias, after controlling for various other factors. See also [Roo+99a; Ker+20; Web+02; Pap07; KHB13] for more evidence of this bias, [DI15] for a case where no such bias was found, and [Wan+21] for some initial work on debiasing this bias.

8.2.3 Game-theoretic approaches

The papers [XDS14; XDVDS18; SM21; Uga23] present theoretical investigations of incentive structures in peer review. It is not clear whether the assumptions underlying the proposed methods are met nor if the relatively complex mechanisms will work in practice. Designing incentives with mathematical guarantees and practical applicability remains an important and challenging open problem.

8.2.4 Signed reviews

An approach to incentivize higher-quality reviews is to have reviewers "sign" their reviews, that is, to release the reviewer identities either publicly or at least to the authors. The proposed incentives are aligned with researchers' incentives to build their reputation (via high-quality reviews) and not spoil it (hence avoid low-quality reviews), and furthermore, can mitigate various types of dishonest behavior (Section 4). However, if required to sign the reviews, some researchers may be afraid to criticize a paper for fear of retribution from the paper's authors.

To quantify these aspects, a study [Roo+99b] conducted a randomized controlled trial to evaluate the effects of signing reviews. They found that asking reviewers to consent to their identities being released did not affect the quality of the reviews or the overall acceptance recommendations, but a significantly higher fraction of reviewers declined to review. Another similar [GGM98] randomized controlled trial also did not find a significant difference in the review quality. The study [RDE10] conducted a randomized controlled trial investigating differences between revealing reviewer identity to only the authors versus revealing reviewer identity publicly did not find any significant difference in the review quality.

Another randomized controlled trial [Wal+00] did find a difference. Among reviewers who agreed to participate (knowing that their name might be released), the experiment found that signed reviews were more courteous and deemed to be of higher quality, and furthermore, signed reviews were also more lenient.

Some peer-review venues have implemented signing of reviews in practice. Nature journals allowed reviewers to optionally sign their reviews, but less than 1% of reviewers actually did so [McC06]. f1000research.com is one of the few venues currently that publishes reviewer identities.

9 Norms and policies

The norms and policies in any community or conference can affect the efficiency of peer review and the ability to achieve its goals. We discuss a few of them here.

9.1 Review quality

We discuss some other aspects pertaining to the quality of the reviews.

9.1.1 Reviewer training.

While researchers are trained to do research, there is little training for peer review. As a consequence, a sizeable fraction of reviews do not conform to basic standards, such as reviewing the paper and not the authors, supporting criticisms with evidence, and being polite.

Several initiatives and experiments have looked to address this challenge. Shadow program committee programs have been conducted alongside several conferences such as the Special Interest Group on Data Communication (SIGCOMM) 2005 conference [Fel05] and IEEE Symposium on Security and Privacy (S&P) 2017 [PEE17]. The ICML 2020 conference adopted a method to select and then mentor junior reviewers, who would not have been asked to review otherwise, with a motivation of expanding the reviewer pool in order to address the large volume of submissions [Ste+21a]. An analysis of their reviews revealed that the junior reviewers were more engaged through various stages of the process as compared to conventional reviewers. Moreover, the conference asked meta reviewers to rate all reviews, and 30% of reviews written by junior reviewers received the highest rating by meta reviewers, in contrast to 14% for the main pool.

Training reviewers at the beginning of their careers is a good start, but may not be enough. There is some evidence [CM11; JD20] that quality of an individual's review falls over time, at a slow but steady rate, possibly because of increasing time constraints or in reaction to poor-quality reviews they themselves receive.

The study [Sat+15] found that for both novice and experienced reviewers, a training video increased the inter-reviewer agreement, improved alignment with the scoring rubrics, and also resulted in reviewers spending more time to read the review criteria. A randomized controlled trial by Schroter et al. [Sch+04] found that reviewer performance can initially be better by training them, but the quality of trained and untrained reviewers becomes indistinguishable six months after the training. Moreover, past studies [CT07] find that there are no easily identifiable types of formal training or experience that could predict reviewers' review quality.

9.1.2 Following peer-review guidelines

More generally, one would hope that reviewers would follow the guidelines set by the peer-review venue (conference program chairs or journal editors). A study [Cha+15] surveyed reviewers of biomedical research journals to investigate the alignment of the tasks that reviewers deem important and that requested by the journal editors. They found that the task that was most frequently requested by editors (to provide recommendations for publication), was rated in the first tertile of importance by only 21% of reviewers, whereas the task considered to be of highest importance by reviewers (that of evaluating the risk of bias) was clearly requested by only 5% of editors. The study thus finds a misalignment between the reviewers' importance on tasks and the editors' guidelines.

9.1.3 Review timeliness

Review timeliness is a major issue in journals due to the (perceived) flexibility of the review-submission timeline [Cor12; BS13], and there are also concerns about reviewers working on a competing idea unethically delaying their review [Aks10; Ben+07; RGFP08]. In contrast, the review timeline is much more strict in conferences, with a fixed deadline for all reviewers to submit their reviews. However, even in conference peer review, a non-trivial fraction of reviews are not submitted by the deadline, and furthermore, an analysis [CL21] of the NeurIPS 2014 conference reviews found evidence that the reviews that were submitted after the deadline were shorter in length, gave higher quality scores, but with lower confidence.

9.2 Author rebuttal

Many conferences allow authors to provide a rebuttal to the reviews. The reviewers are supposed to accommodate these rebuttals and revise their reviews accordingly. There is considerable debate regarding the usefulness of this rebuttal process. The upside of rebuttals is that they allow authors to clarify misconceptions in the review and answer any questions posed by reviewers. The downsides are the time and effort by authors, that reviewers may not read the rebuttal, and that they may be reluctant to change their mind. We discuss a few studies that investigate the rebuttal process.

An analysis of the NAACL 2015 conference found that the rebuttal did not alter reviewers' opinions much [DI15]. Most (87%) review scores did not change after the rebuttals, and among those which did, scores were nearly as likely to go down as up. Furthermore, the review text did not change for 80% of the reviews. The analysis further found that the probability of acceptance of a paper was nearly identical for the papers which submitted a rebuttal as compared to the papers for which did not. An analysis of NeurIPS 2016 found that fewer than 10% of reviews changed scores after the rebuttal [Sha+18]. An analysis of ACL 2017 found that the scores changed after rebuttals in about 15-20% of cases and the change was positive in twice as many cases as negative [Kan17].

The paper [Liu+23] investigates one potential reason behind the limited change in review scores after the author rebuttal. To examine if the phenomenon is influenced by *anchoring* effects [TK74], they design and execute a randomized controlled trial. Within a laboratory setting, researchers were asked to review a paper. Those in the experimental group were first presented with a flawed draft, only to be corrected after their initial review. It was clarified that the flaw was due to a browser glitch rather than the authors' oversight. They were then given the chance to modify their reviews based on the corrected document. Meanwhile, the control group only reviewed the corrected version. Comparing final scores from both groups, the study found no substantial evidence pointing to anchoring.

The paper [Gao+19] designs a model to predict post-rebuttal scores based on initial reviews and the authors' rebuttals. They find that the rebuttal has a marginal (but statistically significant) influence on the final scores, particularly for borderline papers. They also find that the final score given by a reviewer is largely dependent on their initial score and the scores given by other reviewers for that paper.[8]

Two surveys find researchers to have favorable views of the rebuttal process. In a survey [FK20b] of authors of accepted papers at 56 computer systems conferences, 89.7% of respondents found the author rebuttal process helpful. Non-native English speakers found it helpful at a slightly higher rate. Interestingly, the authors who found the rebuttal process as helpful are only half as experienced (in terms of publication records, career stage, as well as program committee participation) as compared to the set of authors who did not find it helpful.

A survey [PEE17] at the IEEE S&P 2017 conference asked authors whether they feel they could have convinced the reviewers to accept the paper with a rebuttal or by submitting a revised version of the paper. About 75% chose revision whereas 25% chose rebuttal. Interestingly, for a question asking authors whether they would prefer a new set of reviewers or the same set if they were to revise and resubmit their manuscript, about 40% voted for a random mix of new and same, little over 10% voted for same, and a little over 20% voted for new reviewers.

In order to improve the rebuttal process, a suggestion was made long ago by Porter and Rossini [PR85] in the context of evaluating interdisciplinary papers. They suggested that reviewers should not be asked to

[8] The paper [Gao+19] concludes *"Peer pressure"* to be *"the most important factor of score change"*. This claim should be interpreted with caution as there is no evidence presented for this causal claim. The reader may instead refer to the controlled experiment [Lan+22] on this topic, discussed in Section 9.3.

provide a rating with their initial reviews, but only after reading the authors' rebuttal. This suggestion may apply more broadly to all papers, but current low reviewer-participation rates in discussions and rebuttals surfaces the concern that some reviewers may not return to the system to provide the final rating (or perhaps optimistically, might incentivize reviewers to return to provide the ratings). Some conferences such as ICLR take a different approach to rebuttals by allowing a continual discussion between reviewers and authors rather than a single-shot rebuttal.

The rebuttal process is immediately followed by a discussion among the reviewers. One may think that the submission of a rebuttal by authors of a paper would spur more discussion for the paper, as compared to when authors choose to not submit a rebuttal. The NAACL 2015 analysis [DI15] suggests absence of such a relation. This brings us to the topic of discussions among reviewers.

9.3 Discussions and group dynamics

After submitting the initial reviews, reviewers of a paper are often allowed to see each others' reviews. The reviewers and the meta reviewer then engage in a discussion in order to arrive at a final decision. These discussions could occur either over video conferencing, or a typed forum, or in person, with various tradeoffs [NIH20] between these modes. We discuss a few studies on this topic.

9.3.1 Do panel discussions improve consistency?

Several studies [OTD07; Fog+12; Pie+17] conduct controlled experiments in the peer review of grant proposals to quantify the reliability of the process. The peer-review process studied here involves discussions among reviewers in panels. In each panel, reviewers first submit independent reviews, following which the panel engages in a discussion about the proposal, and reviewers can update their opinions. These studies reveal the following three findings. First, reviewers have quite a high level of disagreement with each other in their independent reviews. Second, the inter-reviewer disagreement within a panel decreases considerably after the discussions (possibly due to implicit or explicit pressure on reviewers to arrive at a consensus). This observation seems to suggest that discussions actually improve the quality of peer review. After all, it appears that the wisdom of all reviewers is being aggregated to make a more "accurate" decision. To quantify this aspect, these studies form multiple panels to evaluate each proposal, where each panel independently conducts the entire review process including the discussion. The studies then measure the amount of disagreement in the outcomes of the different panels for the same proposal. Their third finding is that, surprisingly, the level of disagreement across panels does *not* decrease after discussions, and instead often increases. Please see Figure 13 for more details.

The paper [Hof+00] performed a similar study in the peer review of hospital quality, and reached similar conclusions: *"discussion between reviewers does not improve reliability of peer review."*

In computer science, an experiment was carried out at the NeurIPS 2014 conference [LC14; CL21] to measure the inconsistency in the peer-review process. In this experiment, 10% of the submissions were assigned to two independent committees, each tasked with the goal of accepting 22% of the papers. It was found that 57% of papers accepted by one committee were rejected by the other. However, details of relative inter-committee disagreements before and after the discussions are not known. A similar experiment at NeurIPS 2021 [Bey+23] found that the levels of inconsistency were consistent with 2014 despite an order of magnitude increase in the number of submissions. A similar (albeit smaller scale) experiment [Bas20] at the European Symposium on Algorithms (ESA) 2018 conference found that the amount of overlap between the sets of accepted papers in the two independent committees was 58%. There was some agreement across the two panels regarding clear rejects, but very little regarding clear accepts. The discussions pertaining to papers were conducted on a per-paper basis led to an increase in the amount of agreement across the two panels.

These observations indicate the need for a careful look at the efficacy of the discussion process and the protocols used therein. We discuss two experiments investigating potential reasons for the surprising reduction in the inter-panel agreement after discussions.

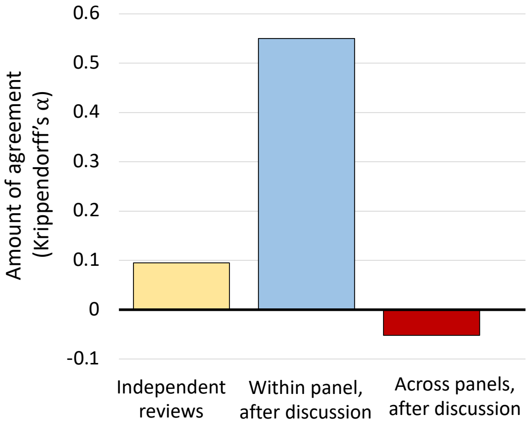


Figure 13: Amount of agreement before and after discussions [Pie+17] in terms of the Krippendorff's alpha coefficient $\alpha = 1 - \frac{\text{amount of observed disagreement}}{\text{amount of disagreement expected by chance}}$. (Left bar) independent reviews submitted by reviewers have a low agreement, (middle bar) the agreement among reviewers within any panel significantly increases after discussions among reviewers within the panel, and (right bar) after discussions within each panel, the agreement *across* panels is negative indicating a slight disagreement.

9.3.2 Influence of other reviewers

Teplitskiy et al. [Tep+19] conducted a controlled study that exposed reviewers to artificial ratings from other (fictitious) reviews. They found that 47% of the time, reviewers updated their ratings. Women reviewers updated their ratings 13% more frequently than men, and more so when they worked in male-dominated fields. Ratings that were initially high were updated downward 64% of the time, whereas ratings that were initially low were updated upward only 24% of the time. An extended version of this study is available at [Lan+22].

9.3.3 Herding effects

Past research on human decision-making finds that the decision of a group can be biased towards the opinion of the group member who initiates the discussions. Such a "herding" effect in discussions can undesirably influence the final decisions in peer review. In ML/AI conferences, there is no specified policy on who initiates the discussions, and this decision can be at the discretion of the meta reviewer or reviewers. A large-scale controlled experiment conducted at the ICML 2020 conference studied the existence of a "herding" effect [Ste+23b]. The study investigated the question: Does the final decision of the paper depend on the order in which reviewers join the discussion? They partitioned the papers at random into two groups. In one group, the most positive reviewer was asked to start the discussion, then later the most negative reviewer was asked to contribute to the discussion. In the second group, the most negative reviewer was asked to start the discussion, then later the most positive reviewer was asked to contribute. The study found no difference in the outcomes of the papers in the two groups. The absence of a "herding" effect in peer review discussions thus suggests that from this perspective, the current absence of any policy for choosing the discussion initiator does not hurt.

9.3.4 Anonymous versus non-anonymous discussions

The Uncertainty in Artificial Intelligence (UAI) 2022 conference conducted a randomized trial to understand the effects of anonymizing (or not) reviewers to each other in the discussions [Ras+24a]. Specifically, papers and reviewers were divided into two conditions: one condition in which reviewers engaged in discussions anonymized to each other, and the other condition in which reviewers were shown each others' identities. All of these discussions took place asynchronously over a typed forum.

The study included various measurements including a survey of the reviewers. They found that (i) reviewers discussed marginally more in the anonymous condition, (ii) paper acceptance decisions were closer to senior reviewers' opinions in the non-anonymous condition than in anonymous, (iii) there was no significant difference in politeness, (iv) reviewers self-reported differences did not differ across the two conditions, (v) reviewers weakly preferred anonymous discussions, (vi) approximately 7% of reviewers said they had experienced some dishonest behavior either in this conference or some other venue.

9.3.5 A survey of reviewers.

There are various conferences and grant proposal evaluations in which the entire pool of reviewers meets together (either in person or online) to discuss all papers. The IEEE S&P 2017 conference was one such conference, and here we discuss a survey of its reviewers [PEE17]. The survey asked the reviewers how often they participated in the discussions of papers that they themselves did not review. Out of the respondents, about 48% responded that they did not engage in the discussions of any other paper, and fewer than 15% reported engaging in discussions of over two other papers. On the other hand, for the question of whether the meeting contributed to the quality of the final decisions, a little over 70% of respondents thought that the discussions did improve the quality.

9.4 Ensuring reviewer anonymity

As discussed in Section 8.2.4, reviewers do not consent to releasing their identities to authors when given the option to do so. The principle behind this preference lies in the safety and freedom that anonymity provides—allowing reviewers to offer candid feedback without the looming fear of backlash. Such anonymity is considered paramount in the majority of peer-reviewing platforms.

The study detailed in [GFS23] highlights an interesting observation: the timing patterns of posts in reviewer discussion forums might inadvertently allow authors to deduce a reviewer's identity. They analyzed a large conference and found that when a reviewer commented on multiple papers, the reviewer had a 30% chance of making their comments within 5 minutes of one other, whereas a pair of distinct reviewers had only a 0.66% chance of making comments on different papers within 5 minutes of each other. Thus an attack using the times of posting can compromise reviewer anonymity. To counteract this risk, the study proposes the introduction of random delays to the timings of these posts, that are proved to provide differential privacy.

Another potential vulnerability to reviewer anonymity exists in the automated reviewer matching process discussed in Section 3. Some venues rely solely on the text similarities computed from the submitted papers and reviewer profiles via natural language processing techniques. One conceivable strategy an attacker might employ is leveraging the openly accessible database of all submitted papers (as is the case with venues like ICLR) and the list of all potential reviewers (which some conferences release). With these, the attacker could reproduce the reviewer assignment algorithm, especially if it's open-source, to discern which reviewers were assigned to particular papers. However, as pointed out in the paper [Jec+20], introducing randomness to the assignment procedures offers a certain degree of protection against such attempts, provided the random seed remains undisclosed.

9.5 Time spent by reviewers

It is estimated that every year, 15 million hours of researchers' time is spent in reviewing papers that are eventually rejected [The13]. Within computer science, in a 2017 survey [PEE17] of reviewers in the IEEE Security and Privacy conference finds that reviewers self report spending a median of 4 hours (mean of 4.6 hours) of time to review a paper, with the minimum self report being 1 hour and the maximum being 12 hours. In a survey [Ern+21] in the subfield of software engineering, 88% of reviewers for journal papers, 56% of reviewers for conference papers, and 16% of reviewers for workshop papers self reported spending over 2 hours to review a paper.

A 2015 survey of researchers across multiple fields [War16] reports that reviewers claim to spend a median of 5 hours (median 8.4 hours) of time reviewing a paper, and these numbers are comparable to an earlier survey from 2007. Similarly, another 2018 survey also reports a median of 5 hours [Pub18].

9.6 Other aspects

Various other aspects pertaining to peer review are under focus, that are not covered in detail here. This includes the (low) acceptance rates at conferences [Chu05; And09; Zha+22], various problems surrounding the reproducibility crisis [Coc+20; Bak16] (including HARKing [Gen+19] and data withholding [Cam+02]), desk rejection [Bey+19], socio-political issues [HGC03], post-publication review [Kri12; Bor+20] (and deployments in pubpeer.com, openreview.net), reviewer forms [Ste17; Sha+18], two-stage reviewing [Mog13; PLD15; LBM21; LB+22; Jec+22a], alternative modes of reviewing [WC11; Bar16; Mac+19; Emi+22], and others [RA20], including calls to abolish peer review altogether [Was12].

10 Peer-review objectives

In this section, we discuss research on the three key objectives of peer review (discussed in Section 1). We summarize them first, and detail them subsequently. (1) *Ensuring correctness*: Outside computer science, reviewers do reject flawed papers a large fraction of the time, although they find only a subset of flaws. Within computer science, reviews do not focus on correctness. (2) *Highlighting the "best" research:* There is at best weak evidence that peer review can separate the "top-tier" research from the next tier. (3) *Providing constructive feedback to authors*: A majority of authors across several surveys report that they find the reviews helpful, however, authors' opinions are also biased by whether the reviews recommended acceptance.

10.1 Ensuring correctness

An important objective of peer review is to filter out bad or incorrect science. We discuss controlled studies that evaluate how well peer review achieves this objective. We review a few studies that specifically evaluate peer review in terms of its ability to identify errors by deliberately adding errors to manuscripts and measuring the rate at which reviewers catch these errors.

10.1.1 Outside computer science

Schroter et al. [Sch+04] introduce 9 major errors to three papers they sent for review to several hundred reviewers. Across various conditions tested in the study (pertaining to different kinds of reviewer training), the mean number of major errors identified by reviewers ranged from 2.13 to 3.37, and the fraction of reviews recommending rejection ranged from 67% to 92%.

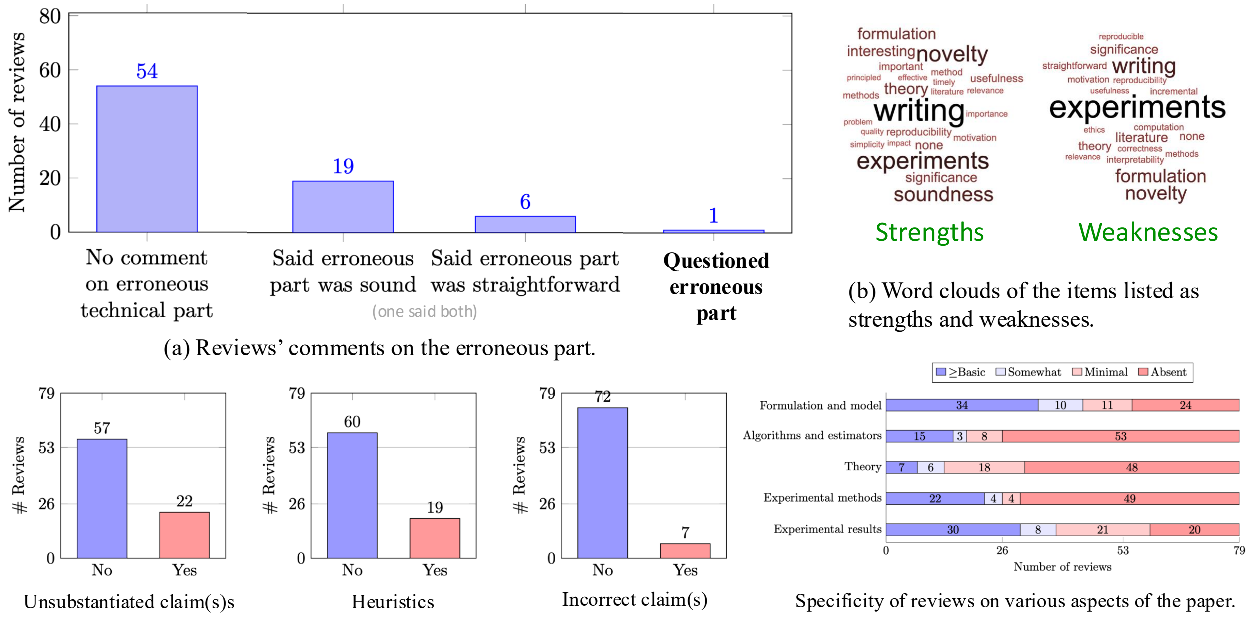
Schroter et al. [Sch+08] assigned three papers to several hundred reviewers, where each paper included a set of 9 deliberately introduced major methodological errors. Reviewers on average identified 2.58, 2.71 and 3 errors in the three papers. They also found that 68% of the reviewers recommended rejection of paper 1, 83% recommended rejection of paper 2, and 81% recommended rejection of paper 3. Providing a short training to reviewers did not affect the outcomes.

Baxt et al. [Bax+98] created a fictitious manuscript and deliberately placed 10 major and 13 minor errors in it. This manuscript was reviewed by about 200 reviewers: 15 recommended acceptance, 117 rejection, and 67 recommended a revision. The reviewers identified one-third of the major errors on average, but failed to identify two-thirds of the major errors. Furthermore, about two-thirds of reviewers did not realize that the conclusions were not supported by the results.

Godlee et al [GGM98] modified a manuscript to deliberately introduce 8 errors. This modified manuscript was reviewed by over 200 reviewers, who on average identified 2 errors. There was no difference in terms of single versus double blind reviewing and in terms of whether reviewer names were revealed publicly.

Emerson et al. [Eme+10] created two versions of a fabricated manuscript, placing 5 errors (2 mathematical, 2 in reference citation, and 1 involving transposition of results in a table).

We note that once a reviewer spots a fatal error, they may not carefully review that paper further, believing that the caught error already provides sufficient grounds for rejection. We were able to analyze the data of [Sch+08] (thanks to the first author Sara Schroter), where we found that 90.94% of reviews detected at least one of the nine major errors. That said, whether the reviewers actually ceased their evaluation upon encountering the first significant error remains unknown. Additionally, it is debatable whether the severity of the error justified this discontinuation of the review.



(a) Reviews' comments on the erroneous part.

(b) Word clouds of the items listed as strengths and weaknesses.

(c) Annotation of review text

Figure 14: Analysis of the 79 reviews from the experiment described in Section 10.1.2.

10.1.2 In computer science

As of July 2024 [Sha24], the author of this survey conducted a study in the review process of a prominent conference in ML/AI whose review process took place some time in 2022-24. In these conferences, there is no policy emphasizing checking correctness, and reviewers have a heavy workload (3-6 reviews) to be completed in short time (3-4 weeks). We evaluated peer review's objective of ensuring correctness under such policies and review workloads. We created three versions of a paper, each containing a significant error in a purported key result in the main text of the paper. One error was regarding a false claim that an optimization problem is convex, a second involved the main algorithm making an inappropriate selection of hyperparameters based on the test set thereby rendering the evaluations of this algorithm incorrect, and the third was an incorrect theorem claiming to derive necessary and sufficient certain conditions for statistical identifiability. The optimization error was the most apparent to spot, the evaluation error could be spotted on reading the pseudocode of the algorithm, and the identifiability error was really subtle in the proof. The experimental papers are available at `https://github.com/niharshah/PaperCorrectnessCheck`.

The study obtained 79 reviews in the review process, working with the program chairs and workflow chairs of the conference, and with an IRB approval. The analysis of the review texts revealed the following:

- Figure 14(a) presents an evaluation of the reviews along the primary outcome variable of detecting the flaw in the paper. One out of the 79 reviews brought the erroneous segment under scrutiny.

- 31% of the reviews recommended accepting the paper (the overall acceptance rate at the conference was approximately 25%).

- The generalized Jaccard similarity between the multisets of the strengths and weaknesses stated in the review texts was 0.47. Figure 14(b) depicts a wordcloud of the stated strengths and weaknesses.

- Figure 14(c) further annotates the reviews according to various quality criteria: 19 reviews employed heuristics, 22 unsubstantiated claims, and 7 made incorrect claims. Review text primarily focused on the formulation and model and the experiments, and little on the algorithms and theory.

- In order to evaluate the quality reviews in a holistic manner, we now consider multiple criteria jointly: we compute the number of reviews that had no unsubstantiated claims, no heuristics, no false claims, and had at least a basic specificity on at least two of the five components of the paper. We tabulate the results in Table 2. Overall, 21 (26.6%) reviews met all these criteria. This performance showed no correlation with the self-reported confidence or expertise in the reviews.

Correlation with self-reported confidence	Coefficient = -0.108; p = 0.98
Correlation with self-reported expertise	Coefficient = -0.002; p = 0.30

Table 2: Kendall tau b correlation of annotated review quality with self reports of confidence and expertise.

There are important limitations to note. The study considered three versions of a single paper on one particular topic in one conference. Consequently, the applicability of the findings to other contexts or disciplines requires verification through additional research. Moreover, due to the absence of established standards of what errors are expected to be detected in peer review, it remains unclear whether the types of errors examined, which vary in their detectability, are generally expected to be identified.

We subsequently also evaluated the ability of large language models in detecting the errors. Please see Section 11.3.2 for details.

10.1.3 Overall remarks

These results indicate that while peer review does filter out some poor-quality science, there is significant room for improvement. The prevalent evaluation methods, which often emphasize subjective aspects like perceived novelty or impact, can detract from the primary goal of ensuring the correctness of published work. An insightful controlled experiment by Lane et al. [Lan+24] found that asking reviewers to evaluate multiple criteria simultaneously can dilute the quality of assessments with respect to each individual criterion.

To enhance peer review's effectiveness in detecting errors, a proposed incremental change is to assign well-defined roles to reviewers. Currently, reviewers are typically asked to simultaneously assess both subjective and objective criteria, with insufficient emphasis on rigor. It may be insightful to experiment with a system where some reviewers are solely responsible for checking correctness. Some journals have taken this approach further by emphasizing rigor as their primary acceptance criterion, such as the Transactions on Machine Learning Research (TMLR), Public Library of Science (PLOS) ONE, Nature Scientific Reports, PeerJ, F1000Research, BioMed Central (BMC) Series journals, Royal Society Open Science, the Journal of Systems Research (JSys), Journal of Open Research Software (JORS), Journal of Open Source Software (JOSS), and many others.

10.2 Selecting the 'best' research

Many venues impose a constraint on the number or fraction of papers that can be accepted. Acceptance or rejection of a paper in such selective or so-called top-tier venues has profound implications. For instance, the study [PRK18] in the field of economics compared researchers' perceptions towards two fictitious curricula vitae (CVs). One CV featured papers exclusively in so-called top-tier venues, whereas the other was identical in all other respects but also listed additional publications in lower-tier venues. The study found that the CV with the additional publications in venues perceived as lower-tier was viewed significantly less favorably. Such perceptions fuel intense competition for coveted spots in selective venues, since rejection in these venues necessitates either shelving the paper or publishing in alternative venues perceived to be lower quality.

In these selective venues, papers are evaluated on criteria such as reviewers' perceptions of novelty and predictions of future impact, in addition to rigor. As discussed in previous sections of this survey, numerous factors such as biases, subjectivity, fraud and arbitrariness can influence the evaluations. This naturally raises the question: how effective is peer review in its mission to discern the 'top' quality research? We discuss this from three perspectives.

10.2.1 Inter-reviewer agreement

Numerous studies across various fields have found that the agreement between reviewers is "poor" [Cic91; Bor15; Jir+17]. In computer science, an analysis [Rag+13] of reviews from 10 conferences found that the intraclass correlation coefficient was at least 0.6 (significant correlation) in six conferences, 0.4 to 0.59 (fair correlation) for three conferences, and smaller (poor correlation) for one conference. The NeurIPS 2016 conference asked reviewers to rate papers on four criteria. The study [Sha+18] measured the amount of agreement between reviewers on relative rankings of papers when two reviewers both reviewed two papers. They found that the reviewers disagreed a fourth to a third of the time. There was no difference in the amount of disagreements within junior versus senior reviewers.

Further, in Section 9.3 we discuss several experiments which have independent sets of reviewers reviewing the same set of submissions. These experiments find a significant amount of disagreement in the subset of papers accepted by the two independent sets of reviewers; some of the experiments in fact find a zero correlation between reviewers across the two sets.

10.2.2 Inter-coauthor agreement

An experiment at the NeurIPS2021 conference [Ras+24b] asked authors who submitted multiple papers to rank their papers in terms of their own perceived scientific contributions of these papers. Interestingly, they found that the amount of disagreement between co-authors about their jointly authored papers was as high as the disagreement with reviewers. Specifically, co-authors disagreed on the relative ranking of a pair of jointly authored papers roughly $1/3^{rd}$ of the time. This result suggests existence of an intrinsic level of noise in identifying the 'best' research, that may be fundamentally hard to overcome.[9]

10.2.3 Relation between reviews and subsequent impact

Finally, we discuss studies analyzing the correlation between the reviews and the future impact of the papers. These studies measure impact in terms of the number of citations received by the papers. The papers [Rag+13; Wei+24] find that for the accepted papers, the reviewer scores are nearly uncorrelated with the number of citations received subsequently. Likewise, the study [Eys22] finds that reviewers' ratings of perceived impact are uncorrelated with citations, but sometimes correlated better with altmetrics (such as social media impressions). Similarly, the study [CMF14] finds that reviewer ratings are uncorrelated with citations or downloads. An analysis [CL21] of the NeurIPS 2014 conference finds no significant correlation: "If we accept that final paper citation counts are some measure of paper quality, then we see that reviewers fail to capture this in their scores." The study [CL21] also analyses the rejected papers and finds a weak correlation between the reviewer scores and future citations. The result on rejected papers may however be taken with a pinch of salt, as authors of rejected papers may have taken different paths for these papers depending on the reviews. The study [Pat+24] finds no significant correlation of the reviewer rating scores with either 2-year citations or altmetrics. Interestingly, they also don't find any significant difference in two-year citations between accepted and rejected manuscripts (although the altmetric scores are significantly higher for accepted papers). Finally, Schroter et al. [Sch+22] explicitly task evaluators to forecast future citations, and find that the evaluators fail to make such predictions accurately.

In conclusion, the available evidence at best only weakly supports the notion that reviewers can consistently and reliably differentiate so-called "top" research worthy of publication in highly selective venues from the next tier of research.

10.3 Constructive feedback to authors

A key objective of peer review is to provide feedback to authors that can help them improve their manuscript, and more generally, perhaps their research as well. There are many surveys of authors about their perceptions of the helpfulness of the reviews. These surveys generally find that a majority of authors do find reviews helpful, but the stated helpfulness is significantly confounded by how positive the review is towards the submission. We provide more details and references in the remainder of this section.

In a survey [PGF18] of authors of software engineering conferences ICSE 2014/15/16, approximately a third of respondents said reviews were good (helpful for the acceptance decision and for the authors and that substantiates all its points), a third said they were reasonable, and the rest said the reviews were unhelpful or grossly faulty.

[9] The experiment [Ras+24b] also asked authors to predict the probability of acceptance of their papers. Authors significantly over-predicted: for an acceptance rate of 21% to 25%, the mean author prediction was 67%.

A survey [FK20b] of authors of accepted papers in computer systems conferences also found that about a third of respondents found the reviews very helpful, about half found them somewhat helpful, and about a sixth found them unhelpful.

In a survey [KHB13] of authors of the CVPR 2012 conference, respondents indicated reviews as helpful 42% of the time, somewhat helpful 37% of the time, and unhelpful 21% of the time. These numbers are significantly confounded by the rating that the reviewer gave the paper, with more positive reviews being considered more helpful.

In reviews of proposals in astronomy, the studies [Pat+19; Ker+20] find that experts seemingly very rarely give unhelpful comments and that non-experts rarely give very helpful comments.

Finally, authors in the NeurIPS 2021 conference were asked a different question – after reading the reviews, how did your perception of the value of your paper change [Ras+24b]? Among both accepted and rejected papers, more than 30% authors reported that their perception about their paper became more positive.

11 AI reviewing

Historically, there have been several partial uses of AI in reviewing of papers, such as ensuring papers adhere to appropriate submission and reporting guidelines, statistical rigor, plagiarism checking, and to mitigate fraud by finding duplicated images or fake papers [NP20; FMG19; VN+22; CLM22; NW23]. In this section, we will discuss more recent work beyond this on fully autonomous AI-based reviewers. We partition the section according to the stated objectives of the developed AI method. Section 11.3.2 also includes new results by the author of this survey that are not published elsewhere.

11.1 Predicting peer-review scores

A number of works [Hua18; Wan+20; YLN21; Che+21; IA24; Shc+24; TY25; Chi+25; Shi+25] either develop new AI methods or test off-the-shelf methods that could predict the scores given by (human) peer reviewers to papers in (past) peer-review processes. While data for such a task is readily available (e.g., on OpenReview), the challenge with this approach is lack of objectivity and that past human reviews themselves have numerous inadequacies as discussed throughout this survey. Moreover, many of these works solely focus on predicting human reviewer scores, while not considering the other rich data associated with the reviews such as the text of the reviews.

11.2 Human evaluations of AI reviews

The papers [Lia+23; D'A+24; Tys+24] propose autonomous reviewers based on LLMs, and conduct surveys of researchers to evaluate the generated reviews.

The study [Lia+23] surveys 308 researchers on their perceptions of the feedback generated by the AI reviewer on their own papers. They find that 57.4% respondents rated the AI reviews helpful or very helpful and 82.4% rated it more beneficial than reviews from at least some human reviewers.

The paper [D'A+24] constructs a multi-agent system for reviewing, where multiple GPT-4 instances interact with each other to generate the review. They conduct a blinded human evaluation in which the reviews generated by their system are rated significantly better than those by a baseline GPT-4 system.

[Tys+24] conduct a blinded experiment in which three researchers evaluated reviews written either by GPT-4 or human reviewers. The findings reveal that the evaluation ratings were similar between the two groups. Additionally, the researchers correctly identified whether a review was written by GPT-4 or a human in approximately 59% of cases.

A shortcoming of the approach of human evaluations of AI generated reviews, given the biases in review evaluation discussed previously: (1) Evaluators are biased positively towards longer reviews [Gol+23]; (2) Asking authors to evaluate the reviews for their own papers [Lia+23] is fraught with the bias that authors provide better ratings for more positive reviews [Gol+23]; (3) The results can be biased in the absence of any blinding or control group [Lia+23].

[Sul+24] conduct a study where researchers are asked to compare reviews provided by human peer reviewers and those generated by GPT-4 on medical research papers. Their findings reveal that 78.5% of the observations made by human reviewers were not mirrored in the comments made by GPT-4. Specifically, the disparity was more pronounced in comments pertaining to the context and methodology of the papers, which showed substantially less concordance compared to more general comments.

11.3 Evaluating correctness

We now move from subjective metrics discussed previously to objective ones. Specifically, we consider what is arguably the primary objective of peer review – ensuring correctness of published research.

11.3.1 Short papers with deliberately inserted flaws

The study [LS23] constructed 13 short papers and one pilot short paper, each no more than a page long, deliberately inserting a fatal flaw in each. These errors ranged from conceptual to mathematical mistakes. While asking LLMs to simply 'review the paper' proved ineffective, when asked pointedly to identify any errors, GPT-4 successfully detected the mistakes in seven out of the 13 papers, as well as in the pilot paper.

11.3.2 Full paper, and direct comparison with human reviewers

The author of this survey evaluated the ability of LLMs to detect errors in a full research paper. We developed three versions of a fictitious paper, each with a significant error intentionally embedded in one of the main claims. These errors were crafted to be fully contained within the paper's main text, relate directly to a result touted as a key contribution, and not be immediately obvious but require a detailed reading of the paper. The papers are available at `https://github.com/niharshah/PaperCorrectnessCheck`. These are the very papers that were used in the experiment described in Section 10.1.2, thereby providing a **direct comparison between the performance of the LLM and the human reviewers in a real-world setting**.

We explored two types of prompts: (i) asking the LLM to evaluate the entire paper and (ii) dividing the paper's results into fine-grained modules and separately asking the LLM to evaluate each individual result. Our findings indicate that asking the LLM to evaluate the entire paper generically is ineffective. However, when the LLM is prompted to evaluate individual results for correctness, it shows some success. Specifically, it consistently detects one of the three errors but fails to identify the second error entirely. For the third error, the LLM did not detect it under our standard prompts. To understand the appropriate level of granularity for error detection, we progressively refined the prompts to be more specific to the part containing the error. The LLM successfully identified the third error on the third iteration of increasingly specific prompts.

11.3.3 A chimera test

As of February 2024, the author of this survey proposed and conducted a "chimera" test. The aim was to *objectively* evaluate automated reviewers. The author put together parts of three of his own papers, each addressing different problems in theoretical statistics, to create a nonsensical paper. This nonsensical paper was then submitted to various AI reviewer systems [Lia+23; Tys+24; D'A+24; CS 25] for evaluation. The results showed that none of these AI reviewers identified the obvious major flaws in the paper.

11.4 Identifying better abstracts

A second evaluative goal of peer review often is to select the better or more intersting research. In general, this goal is somewhat ill specified and quite subjective. Despite this challenge, in order to evaluate AI reviewers on this goal objectively, [LS23] conduct the following experiment. They designed ten pairs of abstracts. In each pair, the two abstracts addressed the same problem, with one abstract presenting superior results. Some pairs included elements such as exaggerated language or prompt injection attacks to test the AI reviewer's robustness. The experiment then presented each pair of abstracts to the AI reviewer, and asked it to pick the one with superior results. The metric of success was the fraction of pairs for which it could correctly pick out the abstract with superior results. They evaluated GPT-4 which was the state of the art model at that time, but found that it did not perform well, failing to correctly identify the superior abstract in five out of the ten instances.

11.5 Biases

The paper [PPM25] evaluates reviews written by GPT4o-mini for papers in Economics, where the experimenters vary the characteristics of the authors of the papers being evaluated. They find biases towards top institutions, male authors, and famous researchers. They also find that the model struggled to distinguished between genuine and AI-generated submissions.

It has also been found that LLM reviewers are generally more lenient as compared to human reviewers. For instance, the mean score given by humans to all papers submitted to the ICLR 2024 conference is 5.11 [Pap24], while the mean score given by LLM reviewers to the same set of papers is 7.0 [Rao+25, Table 14, "without instruction"]. Further, the standard deviation of human reviews is 1.26 as compared to only 0.26 for LLM reviews.

The blog [Wil25] argues that complete automation of reviews can lead to undesirable equilibria due to 'monoculture': it would centralize power in the hands of whoever controls the model, implicitly steering the research agenda.

11.6 Adversarial attacks on LLM reviewers

LLM reviewers are also found susceptible to various attacks. One such attack is executed via an indirect prompt injection – embedding instructions for the LLM reviewer in the PDF of the paper, which is found to be quite successful [Rao+25, Appendix C].

A second such attack [Lin+25] pertains to using methods from adversarial machine learning to modify the text of the submission in a manner that successfully elicits a more positive review. This attack is found to be highly successful in making the recommendation provided by the LLM reviewer more positive.

11.7 Detecting illegitimate LLM-generated reviews

The organizers of peer-review processes require peer reviewers to write their own reviews, and not submit LLM generated reviews. After all, if the organizers wanted LLM-generated reviews, they could query the LLM themselves. Despite these requirements, it is estimated [Lia+24; Lat+24] that a non-negligible fraction of reviewers submit LLM-generated reviews. Rao et al. [Rao+25] propose statistical tests to detect LLM generated reviews under an indirect prompt injection framework. Specifically, in this framework, a command is inserted for the LLM reviewer in the PDF of the paper (e.g., via a font-embedding attack discussed earlier in this survey), which instructs the LLM to insert a watermark that was previously chosen at random by the peer-review organizers. When the reviews arrive, their proposed tests then detect LLM generated reviews via the presence of the (randomly chosen) watermarks, in a manner that can control the family-wise error rate even when the number of reviews is large, and without making assumptions on how the honest human-reviews were written.

11.8 LLM feedback to human reviewers

An experiment at the ICLR conference had LLMs provide feedback to human reviewers on their reviews [Tha+25]. Specifically, they conducted a randomized controlled trial where a subset of reviewers were provided with LLM feedback on their reviews. They found that 27% of reviewers who were provided the LLM feedback updated their reviews, that about 12,000 of the LLM suggestions were incorporated by the human reviewers, and that blinded researchers rated the reviews with the LLM feedback as more informative (although this may be affected bi reviewing review biases 8.2.2).

12 Discussion

Research on peer review faces at least two overarching challenges. The first challenge is about challenges in measuring the outcomes of any policy change or algorithmic use. There is no "ground truth" regarding which papers should have been accepted. Proxies such as subsequent citations (of accepted versus rejected papers) are sometimes employed, but they face a slew of other biases and problems [AR09; And09; DOR12; ALW19; Cha20; Rez+20; DLCRGTS14]. Furthermore, there are no agreed-upon standards of the objectives on how to measure the quality of peer review, thereby making quantitative analyses challenging: *"having precise objectives for the analysis is one of the key and hardest challenges as it is often unclear and debatable to define what it means for peer review to be effective"* [Rag+13; JWD02]. One can sometimes evaluate individual modules of peer review and specific biases, as discussed in this article, but there is no well-defined measure of how a certain solution affected the entire process.

A second challenge is the unavailability of data: *"The main reason behind the lack of empirical studies on peer-review is the difficulty in accessing data"* [BGH16]. Research on improving peer review can significantly benefit from the availability of more data pertaining to peer review. However, a large part of the peer-review data is sensitive since the reviewer identities for each paper and other associated data are usually confidential. For instance, the paper [TZH17] on the aforementioned WSDM 2017 experiment states: *"We would prefer to make available the raw data used in our study, but after some effort we have not been able to devise an anonymization scheme that will simultaneously protect the identities of the parties involved and allow accurate aggregate statistical analysis. We are familiar with the literature around privacy preserving dissemination of data for statistical analysis and feel that releasing our data is not possible using current state-of-the-art techniques."* Designing policies and privacy-preserving computational tools to enable research on such data is an important open problem [DSW20; Jec+20].

Improving scientific review is sometimes characterized as a "fundamentally difficult problem" [LG15]: *"Every program chair who cares tries to tweak the reviewing process to be better, and there have been many smart program chairs that tried hard. Why isn't it better? There are strong nonvisible constraints on the reviewers time and attention."* The current research on improving scientific review, particularly using computational methods, has only scratched the surface of this important application domain. There is a lot more to be done, with numerous open problems which are exciting and challenging, will be impactful when solved, and allow for an entire spectrum of theoretical, applied, and conceptual research.

Acknowledgments

This work is supported by an NSF CAREER Award. This article is based on tutorials on this topic given by the author at several conferences. The author thanks Emma Pierson and Bryan Parno for helpful discussions and feedback on this article. The author thanks the anonymous CACM reviewers whose comments were invaluable in improving this manuscript. The author thinks this demonstrated at least one example that peer review does work. Or perhaps the author is biased since the reviewers recommended acceptance.

References

[ACM21] ACM. *Public Announcement of the Results of the Joint Investigative Committee (JIC) Investigation into Significant Allegations of Professional and Publications Related Misconduct.* 2021. URL: https://www.sigarch.org/wp-content/uploads/2021/02/JIC-Public-Announcement-Feb-8-2021.pdf.

[Ail+19] A. Ailamaki et al. "The SIGMOD 2019 Research Track Reviewing System". In: *ACM SIGMOD Record* (2019).

[Aks10] J. Akst. "I Hate Your Paper. Many say the peer review system is broken. Here's how some journals are trying to fix it". In: *The Scientist* (2010).

[Alo+11] N. Alon et al. "Sum of us: Strategyproof selection from the selectors". In: *Conf. on Theoretical Aspects of Rationality and Knowledge.* 2011.

[ALW19] D. W. Aksnes, L. Langfeldt, and P. Wouters. "Citations, citation indicators, and research quality: An overview of basic concepts and theories". In: *Sage Open* (2019).

[AM+05] S. Al-Marzouki et al. "Are these data real? Statistical methods for the detection of data fabrication in clinical trials". In: *BMJ* (2005).

[And+07] M. S. Anderson et al. "The perverse effects of competition on scientists' work and relationships". In: *Science and engineering ethics* (2007).

[And09] T. Anderson. "Conference reviewing considered harmful". In: *ACM SIGOPS Operating Systems Review* (2009).

[Anj+19] O. Anjum et al. "PaRe: A Paper-Reviewer Matching Approach Using a Common Topic Space". In: *EMNLP-IJCNLP.* 2019.

[Ano13] Anonymous. *Is your PhD a monster?* https://thesiswhisperer.com/2013/09/11/help-i-think-i-have-created-a-monster/. 2013.

[AR09] D. W. Aksnes and A. Rip. "Researchers' perceptions of citations". In: *Research Policy* (2009).

[Arm80] J. S. Armstrong. "Unintelligible management research and academic prestige". In: *Interfaces* (1980).

[Aro+21] I. Arous et al. "Peer Grading the Peer Reviews: A Dual-Role Approach for Lightening the Scholarly Paper Review Process". In: (2021).

[AS12] A. Ammar and D. Shah. "Efficient rank aggregation using partial data". In: *SIGMETRICS*. 2012.

[ASH21] B. Aczel, B. Szaszi, and A. O. Holcombe. "A billion-dollar donation: estimating the cost of researchers' time spent on peer review". In: *Research Integrity and Peer Review* (2021).

[Azi+19] H. Aziz et al. "Strategyproof peer selection using randomization, partitioning, and apportionment". In: *Artificial Intelligence* (2019).

[BA12] R. Beverly and M. Allman. "Findings and implications from data mining the IMC review process". In: *ACM SIGCOMM Computer Communication Review* (2012).

[Bak16] M. Baker. "Reproducibility crisis". In: *Nature* (2016).

[Bal18] M. Baldwin. "Scientific autonomy, public accountability, and the rise of "peer review" in the Cold War United States". In: *Isis* (2018).

[Bar16] B. Barak. "Computer science should stay young". In: *Communications of the ACM* (2016).

[Bas20] H. Bast. *How Objective is Peer Review?* 2020.

[Bas+99] C. Basuyz et al. "Recommending Papers by Mining the Web". In: (1999).

[Bax+98] W. G. Baxt et al. "Who reviews the reviewers? Feasibility of using a fictitious manuscript to evaluate peer reviewer performance". In: *Annals of emergency medicine* (1998).

[BBN21] N. Boehmer, R. Bredereck, and A. Nichterlein. "Combating Collusion Rings is Hard but Possible". In: *arXiv preprint arXiv:2112.08444* (2021).

[Ben+07] D. J. Benos et al. "The ups and downs of peer review". In: *Advances in physiology education* (2007).

[Ber15] R. Bernstein. "PLOS ONE ousts reviewer, editor after sexist peer-review storm". In: *Science* (2015).

[Bey+19] A. Beygelzimer et al. *What we learned from NeurIPS 2019 data.* 2019.

[Bey+23] A. Beygelzimer et al. "Has the Machine Learning Review Process Become More Arbitrary as the Field Has Grown? The NeurIPS 2021 Consistency Experiment". In: *arXiv preprint arXiv:2306.03262* (2023).

[BGH16] S. Balietti, R. Goldstone, and D. Helbing. "Peer review and competition in the Art Exhibition Game". In: *Proceedings of the National Academy of Sciences* (2016).

[BGK05] L. Brenner, D. Griffin, and D. J. Koehler. "Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment". In: *Organizational Behavior and Human Decision Processes* (2005).

[Bha+20] H. Bharadhwaj et al. "De-anonymization of authors through arXiv submissions during double-blind review". In: *arXiv preprint arXiv:2007.00177* (2020).

[BK13] Y. Baba and H. Kashima. "Statistical Quality Estimation for General Crowdsourcing Tasks". In: *KDD*. 2013.

[Bla91] R. M. Blank. "The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review". In: *The American Economic Review* (1991).

[BMD07] L. Bornmann, R. Mutz, and H.-D. Daniel. "Gender differences in grant peer review: A meta-analysis". In: *Journal of Informetrics* (2007).

[BMS87] V. Bakanic, C. McPhail, and R. J. Simon. "The manuscript review and decision-making process". In: *American Sociological Review* (1987).

[BNV14] N. Bousquet, S. Norin, and A. Vetta. "A near-optimal mechanism for impartial selection". In: *International Conference on Web and Internet Economics*. Springer. 2014.

[Bor15] L. Bornmann. "Interrater reliability and convergent validity of F 1000 P rime peer review". In: *Journal of the Association for Information Science and Technology* (2015).

[Bor+20] F. Bordignon et al. "Self-correction of science: a comparative study of negative citations and post-publication peer review". In: *Scientometrics* (2020).

[Bou+16] K. J. Boudreau et al. "Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science". In: *Management science* (2016).

[Bow99] K. Bowyer. "Multiple submission: professionalism, ethical issues, and copyright legalities". In: *IEEE Trans. PAMI* (1999).

[Bro04] T. Brown. *Peer Review and the Acceptance of New Scientific Ideas: Discussion Paper from a Working Party on Equipping the Public with an Understanding of Peer Review: November 2002-May 2004*. Sense About Science, 2004.

[BS13] B.-C. Björk and D. Solomon. "The publishing delay in scholarly peer-reviewed journals". In: *Journal of informetrics* (2013).

[Bud+08] A. E. Budden et al. “Double-blind review favours increased representation of female authors”. In: *Trends in ecology & evolution* (2008).

[Cal+12] V. Calcagno et al. “Flows of research manuscripts among scientific journals reveal hidden submission patterns”. In: *Science* (2012).

[Cam+02] E. G. Campbell et al. “Data withholding in academic genetics: evidence from a national survey”. In: *JAMA* (2002).

[CCS25] J. M. Carpenter, A. Corvillón, and N. B. Shah. “Enhancing Peer Review in Astronomy: A Machine Learning and Optimization Approach to Reviewer Assignments for ALMA”. In: *Publications of the Astronomical Society of the Pacific* (2025).

[Cha+15] A. Chauvin et al. “The most important tasks for peer reviewers evaluating a randomized controlled trial are not congruent with the tasks most often requested by journal editors”. In: *BMC medicine* (2015).

[Cha20] D. S. Chawla. *Improper publishing incentives in science put under microscope around the world*. Chemistry World `https://www.chemistryworld.com/news/improper-publishing-incentives-in-science-put-under-microscope-around-the-world/4012665.article`. 2020.

[Cha21] D. S. Chawla. “Swiss funder draws lots to make grant decisions”. In: *Nature* (2021).

[Che+21] A. Checco et al. “AI-assisted peer review”. In: *Humanities and Social Sciences Communications* (2021).

[Chi+25] M. P. Chitale et al. “AutoRev: Automatic Peer Review System for Academic Research Papers”. In: *arXiv preprint arXiv:2505.14376* (2025).

[Chu05] K. Church. “Reviewing the reviewers”. In: *Computational Linguistics* (2005).

[Cic91] D. V. Cicchetti. “The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation”. In: *Behavioral and brain sciences* (1991).

[CL21] C. Cortes and N. D. Lawrence. “Inconsistency in Conference Peer Review: Revisiting the 2014 NeurIPS Experiment”. In: *arXiv preprint arXiv:2109.09774* (2021).

[CLM22] G. Cabanac, C. Labbé, and A. Magazinov. “The’Problematic Paper Screener’automatically selects suspect publications for post-publication (re) assessment”. In: *arXiv e-prints* (2022).

[CM11] M. Callaham and C. McCulloch. “Longitudinal trends in the performance of scientific peer reviewers”. In: *Annals of emergency medicine* (2011).

[CMF14] R. Connolly, J. Miller, and R. Friedman. “A longitudinal examination of SIGITE conference submission data, 2007-2012”. In: *Proceedings of the 15th Annual Conference on Information technology education*. 2014.

[Coc+20] A. Cockburn et al. “Threats of a replication crisis in empirical computer science”. In: *Communications of the ACM* (2020).

[Coh+16] A. Cohen et al. “Organised crime against the academic peer review system”. In: *British Journal of Clinical Pharmacology* (2016).

[Coh+20] A. Cohan et al. “SPECTER: Document-level Representation Learning using Citation-informed Transformers”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

[Cor12] J. L. Cornelius. “Reviewing the review process: Identifying sources of delay”. In: *The Australasian medical journal* (2012).

[Cot+25] C. S. Cotton et al. “Effect of monetary incentives on peer review acceptance and completion: A quasi-randomized interventional trial”. In: *Critical Care Medicine* (2025).

[CP13] G. Cabanac and T. Preuss. “Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees”. In: *Journal of the Association for Information Science and Technology* (2013).

[CPZ23] C. Cousins, J. Payan, and Y. Zick. “Into the Unknown: Assigning Reviewers to Papers with Uncertain Affinities”. In: *arXiv preprint arXiv:2301.10816* (2023).

[CS 25] CS Paper Reviews. *Test Paper’s Fate at Top Computer Science Research Conferences*. `https://review.cspaper.org/` Last accessed 07-07-2025. 2025.

[CSS14] R. Chetty, E. Saez, and L. Sándor. “What policies increase prosocial behavior? An experiment with referees at the Journal of Public Economics”. In: *Journal of Economic Perspectives* (2014).

[CT07] M. L. Callaham and J. Tercier. “The relationship of previous training and experience of journal peer reviewers to subsequent review quality”. In: *PLoS medicine* (2007).

[CUD19] C. Caragea, A. Uban, and L. P. Dinu. “The myth of double-blind review revisited: ACL vs. EMNLP”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

[CZ13] L. Charlin and R. S. Zemel. “The Toronto Paper Matching System: An automated paper-reviewer assignment system”. In: *ICML Workshop on Peer Reviewing and Publishing Models*. 2013.

[CZB12] L. Charlin, R. S. Zemel, and C. Boutilier. “A Framework for Optimizing Paper Matching”. In: *CoRR* (2012).

[D’A+24] M. D’Arcy et al. “MARG: Multi-Agent Review Generation for Scientific Papers”. In: *arXiv preprint arXiv:2401.04259* (2024).

[Dav17a] P. Davis. *Portable Peer Review RIP*. en-US. 2017.

[Dav17b] P. Davis. *Wither Portable Peer Review*. en-US. 2017.

[Dhu+22] K. Dhull et al. "Strategyproofing Peer Assessment via Partitioning: The Price in Terms of Evaluators' Expertise". In: *HCOMP*. 2022.

[DI15] H. Daumé III. *Some NAACL 2013 statistics on author response, review quality, etc.* `https://nlpers.blogspot.com/2015/06/some-naacl-2013-statistics-on-author.html`. 2015.

[Din+22] W. Ding et al. "Calibration with Privacy in Peer Review". In: *ISIT*. 2022.

[DLCRGTS14] E. Delgado López-Cózar, N. Robinson-García, and D. Torres-Salinas. "The Google scholar experiment: How to index false papers and manipulate bibliometric indicators". In: *Journal of the Association for Information Science and Technology* (2014).

[DN92] S. T. Dumais and J. Nielsen. "Automating the assignment of submitted manuscripts to reviewers". In: *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. 1992.

[Don+19] P. Dondio et al. "The "invisible hand" of peer review: the implications of author-referee networks on peer review in a scholarly journal". In: *Journal of Informetrics* (2019).

[DOR12] DORA. *San Francisco Declaration on Research Assessment*. `https://sfdora.org/read/`. 2012.

[DSW20] W. Ding, N. B. Shah, and W. Wang. "On the Privacy-Utility Tradeoff in Peer-Review Data Analysis". In: *AAAI Privacy-Preserving Artificial Intelligence (PPAI-21) workshop*. 2020.

[Eis+23] T. Eisenhofer et al. "No more Reviewer# 2: Subverting Automatic Paper-Reviewer Assignment using Adversarial Learning". In: *arXiv preprint arXiv:2303.14443* (2023).

[Els21] H. Else. "Scientific image sleuth faces legal action for criticizing research papers". In: *Nature* (2021).

[Eme+10] G. B. Emerson et al. "Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial". In: *Archives of internal medicine* (2010).

[Emi+22] S. H. Emile et al. "Types, limitations, and possible alternatives of peer review based on the literature and surgeons' opinions via Twitter: a narrative". In: (2022).

[ER17] M. A. Edwards and S. Roy. "Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition". In: *Environmental engineering science* (2017).

[ER94] E. Ernst and K.-L. Resch. "Reviewer bias: a blinded experimental study". In: *The Journal of laboratory and clinical medicine* (1994).

[Ern+21] N. A. Ernst et al. "Understanding peer review of software engineering papers". In: *Empirical Software Engineering* (2021).

[EVN21] H. Else and R. Van Noorden. "The fight against fake-paper factories that churn out sham science". In: *Nature* (2021).

[Eys22] G. Eysenbach. "Association Between Peer Reviewers' Priority Ratings of Impact of Research Manuscripts With Citations and Altmetric Scores of Subsequently Published Articles in the Journal of Medical Internet Research". In: *Peer Review Congress (abstract)*. 2022.

[Fal+21] B. Falsafi et al. *Questions About Policies & Processes in the Wake of JIC*. Computer Architecture Today `https://www.sigarch.org/questions-about-policies-processes-in-the-wake-of-jic/`. 2021.

[Fan09] D. Fanelli. "How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data". In: *PloS one* (2009).

[Fel05] A. Feldmann. "Experiences from the SIGCOMM 2005 European shadow PC experiment". In: *ACM SIGCOMM Computer Communication Review* (2005).

[Fer+06] S. Ferilli et al. "Automatic topics identification for reviewer assignment". In: *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer. 2006.

[FFS94] M. Fisher, S. B. Friedman, and B. Strauss. "The effects of blinding on acceptance of research papers by peer review". In: *JAMA* (1994).

[FK15] F. Fischer and M. Klimm. "Optimal impartial selection". In: *SIAM Journal on Computing* (2015).

[FK20a] E. Frachtenberg and R. Kaner. *Representation of women in high-performance computing conferences*. Tech. rep. EasyChair, 2020.

[FK20b] E. Frachtenberg and N. Koster. "A survey of accepted authors in computer systems conferences". In: *PeerJ Computer Science* (2020).

[Fla+10] P. Flach et al. "Novel Tools to Streamline the Conference Review Process: Experiences from SIGKDD'09". In: *SIGKDD Explor. Newsl.* (2010).

[FMG19] T. Foltỳnek, N. Meuschke, and B. Gipp. "Academic plagiarism detection: a systematic literature review". In: *ACM Computing Surveys (CSUR)* (2019).

[FMO14] C. Ferguson, A. Marcus, and I. Oransky. "Publishing: The peer-review scam". In: *Nature News* (2014).

[Fog+12] M. Fogelholm et al. “Panel discussion does not improve reliability of peer review for medical research grant proposals”. In: *Journal of clinical epidemiology* (2012).

[For+19] P. S. Forscher et al. “Little race or gender bias in an experiment of initial review of NIH R01 grant proposals”. In: *Nature human behaviour* (2019).

[Fou+00] N. Fouad et al. “Women in academe: Two steps forward, one step back”. In: *Report of the Task Force on Women in Academe, American Psychological Association* (2000).

[Fre+03] Y. Freund et al. “An Efficient Boosting Algorithm for Combining Preferences”. In: *Journal of Machine Learning Research* (2003).

[Fro21] S. Frolov. *Quantum computing’s reproducibility crisis: Majorana fermions.* 2021.

[FSR20] T Fiez, N Shah, and L Ratliff. “A SUPER* Algorithm to Optimize Paper Bidding in Peer Review”. In: *Conference on Uncertainty in Artificial Intelligence.* 2020.

[FW17] E. A. Fong and A. W. Wilhite. “Authorship and citation manipulation in academic research”. In: *PloS one* (2017).

[Gao+19] Y. Gao et al. “Does My Rebuttal Matter? Insights from a Major NLP Conference”. In: *Proceedings of NAACL-HLT.* 2019.

[Gar+10] N. Garg et al. “Assigning Papers to Referees”. In: *Algorithmica* (2010).

[Gar+94] J. M. Garfunkel et al. “Effect of institutional prestige on reviewers’ recommendations and editorial decisions”. In: *JAMA* (1994).

[GB08] D. Griffin and L. Brenner. “Perspectives on Probability Judgment Calibration”. In: *Blackwell Handbook of Judgment and Decision Making.* Wiley-Blackwell, 2008. Chap. 9.

[Gen+19] O. Gencoglu et al. “HARK Side of Deep Learning–From Grad Student Descent to Automated Machine Learning”. In: *arXiv preprint arXiv:1904.07633* (2019).

[GFS23] A. Goldberg, G. Fanti, and N. B. Shah. “Batching of Tasks by Users of Pseudonymous Forums: Anonymity Compromise and Protection”. In: *ACM SIGMETRICS.* 2023.

[GGM98] F. Godlee, C. R. Gale, and C. N. Martyn. “Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial”. In: *JAMA* (1998).

[Gin+11] D. K. Ginther et al. “Race, ethnicity, and NIH research awards”. In: *Science* (2011).

[Gol+23] A. Goldberg et al. “Peer Reviews of Peer Reviews: A Randomized Controlled Trial and Other Experiments”. In: *arXiv preprint arXiv:2311.09497* (2023).

[GS07] J. Goldsmith and R. Sloan. “The AI conference paper assignment problem”. In: (2007).

[Guo+18] L. Guo et al. “K-Loop Free Assignment in Conference Review Systems”. In: *ICNC.* 2018.

[GW99] S. Goldbeck-Wood. “Evidence on peer review—scientific quality control or smokescreen?” In: *BMJ: British Medical Journal* (1999).

[GWG13] H. Ge, M. Welling, and Z. Ghahramani. *A Bayesian model for calibrating conference review scores.* Manuscript. Available online `http://mlg.eng.cam.ac.uk/hong/unpublished/nips-review-model.pdf` Last accessed: April 4, 2021. 2013.

[Han16] E. Hand. “No pressure: NSF test finds eliminating deadlines halves number of grant proposals”. In: *Science* (2016).

[Har+09] A.-W. Harzing et al. “Rating versus ranking: What is the best way to reduce response and language bias in cross-national research?” In: *International Business Review* (2009).

[HE15] G. Helgesson and S. Eriksson. “Plagiarism in research”. In: *Medicine, Health Care and Philosophy* (2015).

[Hey+22] R. Heyard et al. “Rethinking the Funding Line at the Swiss National Science Foundation: Bayesian Ranking and Lottery”. In: *Statistics and Public Policy* (2022).

[HGC03] M. Hojat, J. S. Gonnella, and A. S. Caelleigh. “Impartial judgment by the “gatekeepers” of science: fallibility and accountability in the peer review process”. In: *Advances in Health Sciences Education* (2003).

[Hil21] J. Hilgard. *Crystal Prison Zone: I tried to report scientific misconduct. How did it go?* 2021.

[HJP03] S. Hill and F. J. Provost. “The myth of the double-blind review? Author identification using only citations”. In: *SIGKDD Explorations* (2003).

[HM13] R. Holzman and H. Moulin. “Impartial nominations for a prize”. In: *Econometrica* (2013).

[HO21] S. E. Hug and M. Ochsner. “Do peers share the same criteria for assessing grant applications?” In: *arXiv preprint arXiv:2106.07386* (2021).

[Hof+00] T. P. Hofer et al. “Discussion between reviewers does not improve reliability of peer review of hospital quality”. In: *Medical care* (2000).

[HP06] S. Hettich and M. J. Pazzani. “Mining for proposal reviewers: lessons learned at the national science foundation”. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.* 2006.

[HRS+24] J. Hsieh, A. Raghunathan, N. B. Shah, et al. "Vulnerability of Text-Matching in ML/AI Conference Reviewer Assignments to Collusions". In: *arXiv preprint arXiv:2412.06606* (2024).

[Hua18] J.-B. Huang. "Deep paper gestalt". In: *arXiv preprint arXiv:1812.08775* (2018).

[Hub+22] J. Huber et al. "Nobel and novice: Author prominence affects peer review". In: *Proceedings of the National Academy of Sciences* (2022).

[Huu10] K. Huutoniemi. *Evaluating interdisciplinary research.* Oxford University Press Oxford, 2010.

[Hvi13] M. Hvistendahl. *China's publication bazaar.* 2013.

[IA24] M. Idahl and Z. Ahmadi. "OpenReviewer: A Specialized Large Language Model for Generating Critical Scientific Paper Reviews". In: *arXiv preprint arXiv:2412.11948* (2024).

[Jam18] K. H. Jamieson. "Crisis or self-correction: Rethinking media narratives about the well-being of science". In: *Proceedings of the National Academy of Sciences* (2018).

[JD20] D. Joyner and A. Duncan. *Eroding Investment in Repeated Peer Review: A Reaction to Unrequited Aid?* `http://lucylabs.gatech.edu/b/wp-content/uploads/2020/07/Eroding-Investment-in-Repeated-Peer-Review-A-Reaction-to-Unrequited-Aid.pdf`. 2020.

[Jec+20] S. Jecmen et al. "Mitigating Manipulation in Peer Review via Randomized Reviewer Assignments". In: *NeurIPS*. 2020.

[Jec+22a] S. Jecmen et al. "Near-Optimal Reviewer Splitting in Two-Phase Paper Reviewing and Conference Experiment Design". In: *HCOMP*. 2022.

[Jec+22b] S. Jecmen et al. "Tradeoffs in Preventing Manipulation in Paper Bidding for Reviewer Assignment". In: *ICLR workshop on ML Evaluation Standards*. 2022.

[Jec+23] S. Jecmen et al. "A Dataset on Malicious Paper Bidding in Peer Review". In: *TheWebConf*. 2023.

[Jec+24] S. Jecmen et al. "On the Detection of Reviewer-Author Collusion Rings From Paper Bidding". In: *arXiv preprint arXiv:2402.07860* (2024).

[Jef+02] T. Jefferson et al. "Effects of editorial peer review: a systematic review". In: *JAMA* (2002).

[Jir+17] J. Jirschitzka et al. "Inter-rater reliability and validity of peer reviews in an interdisciplinary field". In: *Scientometrics* (2017).

[Joe24] F. Joelving. *Paper trail: In the latest twist of the publishing arms race, firms churning out fake papers have taken to bribing journal editors.* `https://www.science.org/content/article/paper-mills-bribing-editors-scholarly-journals-science-investigation-finds`. 2024.

[JR22] T. S. T. Jakobsen and A. Rogers. "What Factors Should Paper-Reviewer Assignments Rely On? Community Perspectives on Issues and Ideals in Conference Peer-Review". In: *arXiv preprint arXiv:2205.01005* (2022).

[JWD02] T. Jefferson, E. Wager, and F. Davidoff. "Measuring the quality of editorial peer review". In: *JAMA* (2002).

[Kah+18] A. Kahng et al. "Ranking wily people who rank each other". In: *AAAI*. 2018.

[Kan17] M.-Y. Kan. *Author Response: Does it help?* ACL 2017 PC Chairs Blog `https://acl2017.wordpress.com/2017/03/27/author-response-does-it-help/`. 2017.

[Kan+18] D. Kang et al. "A dataset of peer reviews (peerread): Collection, insights and NLP applications". In: *arXiv preprint arXiv:1804.09635* (2018).

[Ker19] W. E. Kerzendorf. "Knowledge discovery through text-based similarity searches for astronomy literature". In: *Journal of Astrophysics and Astronomy* (2019).

[Ker+20] W. E. Kerzendorf et al. "Distributed peer review enhanced with natural language processing and machine learning". In: *Nature Astronomy* (2020).

[Kha+21] E. D. Kharasch et al. *Peer review matters: research quality and the public trust.* 2021.

[KHB13] A. Khosla, D. Hoiem, and S. Belongie. "Analysis of Reviews for CVPR 2012". In: (2013).

[Koe93] J. J. Koehler. "The influence of prior beliefs on scientific judgments of evidence quality". In: *Organizational behavior and human decision processes* (1993).

[Kri12] N. Kriegeskorte. "Open evaluation: a vision for entirely transparent post-publication peer review and rating for science". In: *Frontiers in computational neuroscience* (2012).

[KSM19] A. Kobren, B. Saha, and A. McCallum. "Paper Matching with Local Fairness Constraints". In: *ACM KDD*. 2019.

[KSS21] D. Kahneman, O. Sibony, and C. R. Sunstein. *Noise: a flaw in human judgment.* Little, Brown, 2021.

[KTP77] S. Kerr, J. Tolliver, and D. Petree. "Manuscript characteristics which influence acceptance for management and social science journals". In: *Academy of Management Journal* (1977).

[Kur+15] D. Kurokawa et al. "Impartial Peer Review". In: *IJCAI*. 2015.

[Kuz+24] I. Kuznetsov et al. *What Can Natural Language Processing Do for Peer Review?* 2024.

[Lam09] M. Lamont. *How professors think.* Harvard University Press, 2009.

[Lan08] J. Langford. *Adversarial Academia*. 2008.

[Lan12a] J. Langford. *Bidding Problems*. `https://hunch.net/?p=407` [Online; accessed 6-Jan-2019]. 2012.

[Lan12b] J. Langford. *ICML acceptance statistics*. `http://hunch.net/?p=2517` [Online; accessed 14-May-2021]. 2012.

[Lan+22] J. N. Lane et al. "Conservatism gets funded? A field experiment on the role of negative information in novel project evaluation". In: *Management science* (2022).

[Lan+24] J. Lane et al. "Greenlighting Innovative Projects: How Evaluation Format Shapes the Perceived Feasibility of Novel Ideas". In: *Harvard Business School Technology & Operations Mgt. Unit Working Paper* (2024).

[Lat+24] G. R. Latona et al. "The AI review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates". In: *arXiv preprint arXiv:2405.02150* (2024).

[Lau06] G. Laudel. "Conclave in the Tower of Babel: how peers review interdisciplinary research proposals". In: *Research Evaluation* (2006).

[Lau19] M. Lauer. "Case Study in Review Integrity: Undisclosed Conflict of Interest". In: *NIH Extramural Nexus* (2019).

[Lau20] M. Lauer. "Case Study in Review Integrity: Asking for Favorable Treatment". In: *NIH Extramural Nexus* (2020).

[LB+22] K. Leyton-Brown et al. "Matching Papers and Reviewers at Large Conferences". In: *arXiv preprint arXiv:2202.12273* (2022).

[LBM21] K. Leyton-Brown and Mausam. *AAAI 2021 - Introduction*. `https://slideslive.com/38952457/aaai-2021-introduction?ref=account-folder-79533-folders`; minute 8 onwards in the video. 2021.

[LC14] N. Lawrence and C. Cortes. *The NIPS Experiment*. `http://inverseprobability.com/2014/12/16/the-nips-experiment`. [Online; accessed 11-June-2018]. 2014.

[Le +18] C. Le Goues et al. "Effectiveness of anonymization in double-blind review". In: *CACM* (2018).

[Lee+13] C. J. Lee et al. "Bias in peer review". In: *Journal of the Association for Information Science and Technology* (2013).

[Lee15] C. J. Lee. "Commensuration bias in peer review". In: *Philosophy of Science* (2015).

[LG15] J. Langford and M. Guzdial. "The arbitrariness of reviews, and advice for school administrators". In: *Communications of the ACM* (2015).

[LH16] B. Li and Y. T. Hou. "The new automated IEEE INFOCOM review assignment system". In: *IEEE Network* (2016).

[Li17] D. Li. "Expertise versus Bias in Evaluation: Evidence from the NIH". In: *American Economic Journal: Applied Economics* (2017).

[Lia+18] J. W. Lian et al. "The conference paper assignment problem: Using order weighted averages to assign indivisible goods". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[Lia+23] W. Liang et al. "Can large language models provide useful feedback on research papers? A large-scale empirical analysis". In: *arXiv preprint arXiv:2310.01783* (2023).

[Lia+24] W. Liang et al. "Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews". In: *arXiv preprint arXiv:2403.07183* (2024).

[Lin+25] T.-L. Lin et al. "Breaking the Reviewer: Assessing the Vulnerability of Large Language Models in Automated Peer Review Under Textual Adversarial Attacks". In: *arXiv preprint arXiv:2506.11113* (2025).

[Lit21] M. L. Littman. "Collusion rings threaten the integrity of computer science research". In: *Communications of the ACM* (2021).

[Liu+20] M. Liu et al. "The acceptability of using a lottery to allocate research funding: a survey of applicants". In: *Research integrity and peer review* (2020).

[Liu+22] Y. Liu et al. "Integrating Rankings into Quantized Scores in Peer Review". In: *Transactions on Machine Learning Research* (2022).

[Liu+23] R. Liu et al. "Testing for Reviewer Anchoring in Peer Review: A Randomized Controlled Trial". In: *arXiv preprint arXiv:2307.05443* (2023).

[Lon+13] C. Long et al. "On Good and Fair Paper-Reviewer Assignment". In: *ICDM*. 2013.

[LS23] R. Liu and N. Shah. "ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing". In: *arXiv preprint 2306.00622* (2023). AAAI 2024 Workshop on Scientific Document Understanding.

[LSM14] X. Liu, T. Suel, and N. Memon. "A Robust Model for Paper Reviewer Assignment". In: *ACM Conference on Recommender Systems*. 2014.

[Mac+17] R. S. MacKay et al. "Calibration with confidence: a principled method for panel assessment". In: *Royal Society Open Science* (2017).

[Mac+19] T. K. Mackey et al. "A framework proposal for blockchain-based scientific publishing using shared governance". In: *Frontiers in Blockchain* (2019).

[MADV05] B. C. Martinson, M. S. Anderson, and R. De Vries. “Scientists behaving badly”. In: *Nature* (2005).

[Mah77] M. J. Mahoney. “Publication prejudices: An experimental study of confirmatory bias in the peer review system”. In: *Cognitive therapy and research* (1977).

[Mar+17] I. Markwood et al. “Mirage: Content Masking Attack Against Information-Based Online Services”. In: *USENIX Security Symposium*. 2017.

[Mar+22] A. Marcoci et al. “Reimagining peer review as an expert elicitation process”. In: *BMC Research Notes* (2022).

[Mat+20] S. Mattauch et al. “A bibliometric approach for detecting the gender gap in computer science”. In: *Communications of the ACM* (2020).

[McC06] A. McCook. “Is peer review broken? Submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint about the process at top-tier journals. What’s wrong with peer review?” In: *The scientist* (2006).

[McC89] J. McCullough. “First comprehensive survey of NSF applicants focuses on their concerns about proposal review”. In: *Science, Technology, & Human Values* (1989).

[MD06] S. Madden and D. DeWitt. “Impact of double-blind reviewing on SIGMOD publication rates”. In: *ACM SIGMOD Record* (2006).

[Mei+21] R. Meir et al. “A market-inspired bidding scheme for peer review paper assignment”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021.

[Mer68] R. K. Merton. “The Matthew Effect in Science”. In: *Science* (1968).

[MHM06] J. Murphy, C. Hofacker, and R. Mizerski. “Primacy and recency effects on clicking behavior”. In: *Journal of Computer-Mediated Communication* (2006).

[MHR13] A. Mulligan, L. Hall, and E. Raphael. “Peer review in a changing world: An international study measuring the attitudes of researchers”. In: *Journal of the Association for Information Science and Technology* (2013).

[Mit+11] I. Mitliagkas et al. “User rankings from comparisons: Learning permutations in high dimensions”. In: *Allerton Conference*. 2011.

[MM07] D. Mimno and A. McCallum. “Expertise Modeling for Matching Papers with Reviewers”. In: *KDD*. 2007.

[Mog13] J. C. Mogul. *Towards more constructive reviewing of SIGCOMM papers*. 2013.

[MS20] Y. Matsubara and S. Singh. “Citations Beyond Self Citations: Identifying Authors, Affiliations, and Nationalities in Scientific Papers”. In: *International Workshop on Mining Scientific Publications*. 2020.

[MS21] E. Manzoor and N. B. Shah. “Uncovering Latent Biases in Text: Method and Application to Peer Review”. In: *AAAI*. 2021.

[MT13] E. Mohammadi and M. Thelwall. “Assessing non-standard article impact using F1000 labels”. In: *Scientometrics* (2013).

[Mur20] S. Murrin. “NIH Has Acted To Protect Confidential Information Handled by Peer Reviewers, But It Could Do More”. In: (2020).

[MW13] N. Mattei and T. Walsh. “Preflib: A library for preferences http://www. preflib. org”. In: *International Conference on Algorithmic Decision Theory*. Springer. 2013.

[MWC19] L. Mackenzie, J. Wehner, and S. Correll. *Why Most Performance Evaluations Are Biased, and How to Fix Them*. `https://hbr.org/2019/01/why-most-performance-evaluations-are-biased-and-how-to-fix-them`[Online; accessed 6-Jan-2019]. 2019.

[Mys+23] S. Mysore et al. “Editable User Profiles for Controllable Text Recommendation”. In: *arXiv preprint arXiv:2304.04250* (2023).

[Nak+21] R. K. Nakamura et al. “An experimental test of the effects of redacting grant applicant identifiers on peer review outcomes”. In: *Elife* (2021).

[Nau10] J. Naughton. “DBMS Research: First 50 Years, Next 50 Years”. In: *Keynote at ICDE* (2010). `http://pages.cs.wisc.edu/~naughton/naughtonicde.pptx`.

[NBH16] S. Nobarany, K. S. Booth, and G. Hsieh. “What motivates people to review articles? The case of the human-computer interaction community”. In: *Journal of the Association for Information Science and Technology* (2016).

[Nic+15] D. Nicholas et al. “Peer review: still king in the digital age”. In: *Learned Publishing* (2015).

[Nie00] O. Nierstrasz. “Identify the champion”. In: *Pattern Languages of Program Design* (2000).

[Nie+21] M. W. Nielsen et al. “Meta-Research: Weak evidence of country-and institution-related status bias in the peer review of abstracts”. In: *Elife* (2021).

[NIH20] NIH center for scientific review. *Impact of Zoom Format on CSR Review Meetings*. `https://public.csr.nih.gov/sites/default/files/2021-08/CSR_Analysis_of_Zoom_in_Review_July_2021.pdf`. 2020.

[NL13] P. Naghizadeh and M. Liu. “Incentives, quality, and risks: A look into the NSF proposal review pilot”. In: *arXiv preprint arXiv:1307.6528* (2013).

[NOS12] S. Negahban, S. Oh, and D. Shah. "Iterative ranking from pair-wise comparisons". In: *Advances in Neural Information Processing Systems*. 2012.

[NP20] M. B. Nuijten and J. R. Polanin. ""statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses". In: *Research synthesis methods* (2020).

[NSP21] R. Noothigattu, N. Shah, and A. Procaccia. "Loss Functions, Axioms, and Peer Review". In: *Journal of Artificial Intelligence Research* (2021).

[NW23] M. B. Nuijten and J. Wicherts. "Implementing statcheck during peer review is related to a steep decline in statistical reporting errors". In: (2023).

[OJ21] Y. Ophir and K. H. Jamieson. "The effects of media narratives about failures and discoveries in science on beliefs about and support for science". In: *Public Understanding of Science* (2021).

[Oki+16] K. Okike et al. "Single-blind vs double-blind peer review in the setting of author prestige". In: *JAMA* (2016).

[ORA+22] M Ostendorff, N Rethmeier, I Augenstein, et al. "Neighborhood contrastive learning for scientific document representations with citation embeddings". In: *arXiv preprint arXiv:2202.06671* (2022).

[OTD07] M. Obrecht, K. Tibelius, and G. D'Aloisio. "Examining the value added by committee discussion in the review of applications for research awards". In: *Research Evaluation* (2007).

[OW22] M. Olckers and T. Walsh. "Manipulation and Peer Mechanisms: A Survey". In: *arXiv preprint arXiv:2210.01984* (2022).

[Pap07] K. Papagiannaki. "Author Feedback Experiment at PAM 2007". In: *SIGCOMM Comput. Commun. Rev.* (2007).

[Pap24] Paper copilot. *ICLR 2024 Statistics*. `https://papercopilot.com/statistics/iclr-statistics/iclr-2024-statistics/`. Accessed on 21 July 2025. 2024.

[Pat+19] F. Patat et al. "The distributed peer review experiment". In: *The Messenger* (2019).

[Pat+24] A. Patel et al. *How effective is peer review? Measuring the association between reviewer rating scores, publication status, and article impact*. 2024.

[Pau81] S. R. Paul. "Bayesian methods for calibration of examiners". In: *British Journal of Mathematical and Statistical Psychology* (1981).

[PBB23] T. F. Piccinno, A. Basso, and F. Bracco. "Results of a Peer Review Activity Carried out Alternatively on a Compulsory or Voluntary Basis". en. In: *Journal of Chemical Education* (2023).

[PC82] D. P. Peters and S. J. Ceci. "Peer-review practices of psychological journals: The fate of published articles, submitted again". In: *Behavioral and Brain Sciences* (1982).

[PDR19] F. Pomponi, B. D'Amico, and T. Rye. *Who is (likely) peer-reviewing your papers? A partial insight into the world's top reviewers*. 2019.

[PE22] M. Pearce and E. A. Erosheva. "A Unified Statistical Learning Model for Rankings and Scores with Application to Grant Panel Review". In: *arXiv preprint arXiv:2201.02539* (2022).

[PEE17] B Parno, U Erlingsson, and W Enck. *Report on the IEEE S&P 2017 submission and review process and its experiments*. `http://ieee-security.org/TC/Reports/2017/SP2017-PCChairReport.pdf`. 2017.

[PF17] S. Price and P. A. Flach. "Computational Support for Academic Peer Review: A Perspective from Artificial Intelligence". In: *Communications of the ACM* (2017).

[PFS10] S. Price, P. A. Flach, and S. Spiegler. "SubSift: a novel application of the vector space model to support the academic research process". In: *Proceedings of the First Workshop on Applications of Pattern Analysis*. PMLR. 2010.

[PGF18] L. Prechelt, D. Graziotin, and D. M. Fernández. "A community's perspective on the status and future of peer review in software engineering". In: *Information and Software Technology* (2018).

[Phi21] A. Philipps. "Research funding randomly allocated? A survey of scientists' views on peer review and lottery". In: *Science and Public Policy* (2021).

[Pie+17] E. Pier et al. "Your comments are meaner than your score: score calibration talk influences intra-and inter-panel variability during scientific grant peer review". In: *Research Evaluation* (2017).

[Pin+24] M. Piniewski et al. "Emerging plagiarism in peer-review evaluation reports: a tip of the iceberg?" In: *Scientometrics* (2024).

[PLD15] PLDI. *PLDI 2015 author and reviewer surveys*. `https://conf.researchr.org/track/pldi2015/pldi2015-papers#Surveys`. 2015.

[Pou+22] F. Poutoglidou et al. "Fraud and deceit in medical research: insights and current perspectives". In: *Voices in Bioethics* (2022).

[PPM25] P. Pataranutaporn, N. Powdthavee, and P. Maes. "Can AI Solve the Peer Review Crisis? A Large Scale Experiment on LLM's Performance and Biases in Evaluating Economics Papers". In: 2025.

[PR85] A. L. Porter and F. A. Rossini. "Peer review of interdisciplinary research proposals". In: *Science, technology, & human values* (1985).

[PRK18] N. Powdthavee, Y. E. Riyanto, and J. L. Knetsch. “Lower-rated publications do lower academics’ judgments of publication lists: Evidence from a survey experiment of economists”. In: *Journal of Economic Psychology* (2018).

[Pub18] Publons. *Global State of Peer Review 2018*. Tech. rep. Accessed: 2025-08-03. Publons (a Clarivate Analytics company), 2018.

[PZ22] J. Payan and Y. Zick. “I Will Have Order! Optimizing Orders for Fair Reviewer Assignment”. In: *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 2022.

[RA20] A. Rogers and I. Augenstein. “What Can We Do to Improve Peer Review in NLP?” In: *arXiv preprint arXiv:2010.03863* (2020).

[Rag+13] A. Ragone et al. “On peer review in computer science: Analysis of its effectiveness and suggestions for improvement”. In: *Scientometrics* (2013).

[Rao+25] V. Rao et al. “Detecting LLM-Generated Peer Reviews”. In: *arXiv preprint arXiv:2503.15772* (2025).

[Ras+22] C. Rastogi et al. “To ArXiv or not to ArXiv: A Study Quantifying Pros and Cons of Posting Preprints Online”. In: *arXiv preprint arXiv:2203.17259* (2022).

[Ras+24a] C. Rastogi et al. “A Randomized Controlled Trial on Anonymizing Reviewers to Each Other in Peer Review Discussions”. In: *arXiv preprint arXiv:2403.01015* (2024).

[Ras+24b] C. Rastogi et al. “How do Authors’ Perceptions of their Papers Compare with Co-authors’ Perceptions and Peer-review Decisions?” In: *PLOS ONE* (2024). Short blog: `https://blog.ml.cmu.edu/2022/11/22/neurips2021-author-perception-experiment/`.

[RB08] M. A. Rodriguez and J. Bollen. “An Algorithm to Determine Peer-reviewers”. In: *ACM Conference on Information and Knowledge Management*. 2008.

[RBS07] M. A. Rodriguez, J. Bollen, and H. Van de Sompel. “Mapping the bid behavior of conference referees”. In: *Journal of Informetrics* (2007).

[RDE10] S. van Rooyen, T. Delamothe, and S. J. Evans. “Effect on peer review of telling reviewers that their signed reviews might be posted on the web: randomised controlled trial”. In: *BMJ* (2010).

[REG00] K. Resch, E. Ernst, and J. Garrow. “A randomized controlled study of reviewer bias against an unconventional therapy”. In: *Journal of the Royal Society of Medicine* (2000).

[Ren16] D. Rennie. “Let’s make peer review scientific”. In: *Nature* (2016).

[Rez+20] R. Rezapour et al. “Beyond citations: Corpus-based methods for detecting the impact of research outcomes on society”. In: *Proceedings of The 12th Language Resources and Evaluation Conference*. 2020.

[RGFP08] D. B. Resnik, C. Gutierrez-Ford, and S. Peddada. “Perceptions of ethical problems with scientific journal peer review: an exploratory study”. In: *Science and engineering ethics* (2008).

[Rok68] M. Rokeach. “The Role of Values in Public Opinion Research”. In: *Public Opinion Quarterly* (1968).

[Roo+99a] S. van Rooyen et al. “Effect of blinding and unmasking on the quality of peer review”. In: *Journal of general internal medicine* (1999).

[Roo+99b] S. van Rooyen et al. “Effect of open peer review on quality of reviews and on reviewers’ recommendations: a randomised trial”. In: *BMJ* (1999).

[Ros+06] J. S. Ross et al. “Effect of blinded peer review on abstract acceptance”. In: *JAMA* (2006).

[Roz+23] I. Rozencweig et al. “Mitigating Skewed Bidding for Conference Paper Assignment”. In: *arXiv preprint arXiv:2303.00435* (2023).

[RRS11] M. Roos, J. Rothe, and B. Scheuermann. “How to Calibrate the Scores of Biased Reviewers by Quadratic Programming”. In: *AAAI*. 2011.

[SAK17] J. Teixeira da Silva and A. Al-Khatib. “The Clarivate Analytics acquisition of Publons–an evolution or commodification of peer review?” In: *Research Ethics* (2017).

[Sat+15] D. N. Sattler et al. “Grant peer review: improving inter-rater reliability with training”. In: *PloS one* (2015).

[SBT13] F. Squazzoni, G. Bravo, and K. Takács. “Does incentive provision increase the quality of peer review? An experimental study”. In: *Research Policy* (2013).

[Sch+04] S. Schroter et al. “Effects of training on quality of peer review: randomised controlled trial”. In: *BMJ* (2004).

[Sch+08] S. Schroter et al. “What errors do peer reviewers detect, and does training improve their ability to detect them?” In: *Journal of the Royal Society of Medicine* (2008).

[Sch+22] S. Schroter et al. “Evaluation of editors’ abilities to predict the citation potential of research manuscripts submitted to The BMJ: a cohort study”. In: *BMJ* (2022).

[SF77] P. Slovic and B. Fischhoff. “On the psychology of experimental surprises.” In: *Journal of Experimental Psychology: Human Perception and Performance* (1977).

[SG12] F. Squazzoni and C. Gandelli. “Saint Matthew strikes again: An agent-based model of peer review and the scientific community structure”. In: *Journal of Informetrics* (2012).

[Sha+18] N. Shah et al. "Design and Analysis of the NIPS 2016 Review Process". In: *JMLR* (2018).

[Sha22a] N. B. Shah. "Challenges, experiments, and computational solutions in peer review". In: *Communications of the ACM* (2022).

[Sha22b] N. B. Shah. "The Role of Author Identities in Peer Review". In: *arXiv preprint arXiv:2301.00221* (2022).

[Sha24] N. B. Shah. *Theory and Experiments for Peer Review and Other Distributed Human Evaluations.* Tutorial `https://cs.cmu.edu/~nihars/tutorials/SPCOM2024/SPCOMtutorial2024slides.pdf` at SPCOM 2024. 2024.

[Sha+25] N. B. Shah et al. "Identity Theft in AI Conference Peer Review". In: (2025). arXiv:2508.04024.

[Shc+24] A. Shcherbiak et al. "Evaluating science: A comparison of human and AI reviewers". In: *Judgment and Decision Making* (2024).

[Shi+25] H. Shin et al. "Automatically evaluating the paper reviewing capability of large language models". In: *arXiv e-prints* (2025).

[Sie91] S. Siegelman. "Assassins and zealots: variations in peer review". In: *Radiology* (1991).

[Sin+22] A. Singh et al. "Scirepeval: A multi-format benchmark for scientific document representations". In: *arXiv preprint arXiv:2211.13308* (2022).

[SM21] S. Srinivasan and J. Morgenstern. "Auctions and Prediction Markets for Scientific Peer Review". In: *arXiv preprint arXiv:2109.00923* (2021).

[Smi97] R. Smith. *Peer review: reform or revolution?: Time to open up the black box of peer review.* 1997.

[Smu13] Y. M. Smulders. "A two-step manuscript submission process can reduce publication bias". In: *Journal of clinical epidemiology* (2013).

[Spa+14] A. Spalvieri et al. "Weighting peer reviewers". In: *2014 Twelfth Annual International Conference on Privacy, Security and Trust.* IEEE. 2014.

[Spi02] R. Spier. "The history of the peer-review process". In: *TRENDS in Biotechnology* (2002).

[SSM13] D. Soergel, A. Saunders, and A. McCallum. "Open Scholarship and Peer Review: a Time for Experimentation". In: (2013).

[SSS19] I. Stelmakh, N. Shah, and A. Singh. "On Testing for Biases in Peer Review". In: *NeurIPS*. 2019.

[SSS21a] I. Stelmakh, N. Shah, and A. Singh. "Catch Me if I Can: Detecting Strategic Behaviour in Peer Assessment". In: *AAAI.* 2021.

[SSS21b] I. Stelmakh, N. Shah, and A. Singh. "PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review". In: *Journal of Machine Learning Research* (2021).

[Ste17] A. Stent. *Our New Review Form.* NAACL 2018 Chairs blog `https://naacl2018.wordpress.com/2017/12/14/our-new-review-form`. 2017.

[Ste+21a] I. Stelmakh et al. "A Novice-Reviewer Experiment to Address Scarcity of Qualified Reviewers in Large Conferences". In: *AAAI.* 2021.

[Ste+21b] I. Stelmakh et al. "Prior and Prejudice: The Novice Reviewers' Bias against Resubmissions in Conference Peer Review". In: *CSCW*. 2021.

[Ste+23a] I. Stelmakh et al. "A Gold Standard Dataset for the Reviewer Assignment Problem". In: *arXiv preprint arXiv:2303.16750* (2023).

[Ste+23b] I. Stelmakh et al. "A large scale randomized controlled trial on herding in peer-review discussions". In: *PLOS ONE* (2023).

[Ste+23c] I. Stelmakh et al. "Cite-seeing and reviewing: A study on citation bias in peer review". In: *Plos one* (2023).

[Su21] W. Su. "You Are the Best Reviewer of Your Own Papers: An Owner-Assisted Scoring Mechanism". In: *Advances in Neural Information Processing Systems* (2021).

[Sul+24] A. Suleiman et al. "Assessing ChatGPT's Ability to Emulate Human Reviewers in Scientific Research: A Descriptive and Qualitative Approach". In: *Computer Methods and Programs in Biomedicine* (2024).

[TA23] B. K. Tyson and Alec. *Americans' Trust in Scientists, Positive Views of Science Continue to Decline.* en-US. 2023.

[Tan+21] S. Tan et al. "Least Square Calibration for Peer Reviews". In: *Advances in Neural Information Processing Systems* (2021).

[Tay08] C. J. Taylor. "On the optimal assignment of conference papers to reviewers". In: (2008).

[Tay15] Taylor and Francis group. *Peer review in 2015 A global view.* `https://authorservices.taylorandfrancis.com/publishing-your-research/peer-review/peer-review-global-view/`. 2015.

[TC14] W. Thorngate and W. Chowdhury. "By the Numbers: Track Record, Flawed Reviews, Journal Space, and the Fate of Talented Authors". In: *Advances in Social Simulation.* Springer, 2014.

[TC91] G. D. L. Travis and H. M. Collins. "New light on old boys: Cognitive and institutional particularism in the peer review system". In: *Science, Technology, & Human Values* (1991).

[TCH17] H. D. Tran, G. Cabanac, and G. Hubert. “Expert suggestion for conference program committees”. In: *2017 11th International Conference on Research Challenges in Information Science (RCIS)*. 2017.

[Tep+19] M. Teplitskiy et al. “Social Influence among Experts: Field Experimental Evidence from Peer Review”. In: (2019).

[Tep+22] M. Teplitskiy et al. “Is novel research worth doing? Evidence from peer review at 49 journals”. In: *Proceedings of the National Academy of Sciences* (2022).

[TH11] S. Thurner and R. Hanel. “Peer-review in a world with rational scientists: Toward selection of the average”. In: *The European Physical Journal B* (2011).

[Tha+25] N. Thakkar et al. “Can LLM feedback enhance review quality? A randomized study of 20k reviews at ICLR 2025”. In: *arXiv preprint arXiv:2504.09737* (2025).

[The13] The AJE Team. *Peer Review: How We Found 15 Million Hours of Lost Time*. 2013.

[TJ19] D. Tran and C. Jaiswal. “PDFPhantom: Exploiting PDF Attacks Against Academic Conferences’ Paper Submission Process with Counterattack”. In: *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. IEEE. 2019.

[TK74] A. Tversky and D. Kahneman. “Judgment under Uncertainty: Heuristics and Biases”. In: *Science* (1974).

[Tra+20] D. Tran et al. “An Open Review of OpenReview: A Critical Analysis of the Machine Learning Conference Review Process”. In: *arXiv preprint arXiv:2010.05137* (2020).

[TT07] C. R. Triggle and D. J. Triggle. “What is the future of peer review? Why is there fraud in science? Is plagiarism out of control? Why do scientists do bad things? Is it all a case of: “All that is necessary for the triumph of evil is that good men do nothing?”” In: *Vascular health and risk management* (2007).

[TTT10] W. Tang, J. Tang, and C. Tan. “Expertise Matching via Constraint-Based Optimization”. In: *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 2010.

[Tun06] A. K. Tung. “Impact of double blind reviewing on SIGMOD publication: a more detail analysis”. In: *ACM SIGMOD Record* (2006).

[TVS14] J. K. Tijdink, R. Verbeke, and Y. M. Smulders. “Publication pressure and scientific misconduct in medical scientists”. In: *Journal of Empirical Research on Human Research Ethics* (2014).

[TW18] L. Tabak and M. R. Wilson. “Foreign Influences on Research Integrity”. In: (2018).

[TY25] M. Thelwall and A. Yaghi. “Evaluating the predictive capacity of ChatGPT for academic peer review outcomes across multiple platforms”. In: *Scientometrics* (2025).

[Tys+24] K. Tyser et al. “OpenReviewer: Mitigating Challenges in LLM Reviewing”. In: *submitted to ICLR 2024 on OpenReview.net* (2024).

[TZH17] A. Tomkins, M. Zhang, and W. D. Heavlin. “Reviewer bias in single-versus double-blind peer review”. In: *Proceedings of the National Academy of Sciences* (2017).

[Uga23] A. Ugarov. “Peer Prediction for Peer Review: Designing a Marketplace for Ideas”. In: *arXiv:2303.16855* (2023).

[Var10] M. Y. Vardi. “Hypercriticality”. In: *Communications of the ACM* (2010).

[Vij20a] T. N. Vijaykumar. *Potential Organized Fraud in ACM/IEEE Computer Architecture Conferences*. `https://medium.com/@tnvijayk/potential-organized-fraud-in-acm-ieee-computer-architecture-conferences-ccd61169370d`. 2020.

[Vij20b] T. N. Vijaykumar. *Potential Organized Fraud in On-Going ASPLOS Reviews*. 2020.

[VN20] R. Van Noorden. “Highly cited researcher banned from journal board for citation abuse.” In: *Nature* (2020).

[VN+22] R. Van Noorden et al. “Journals adopt AI to spot duplicated images in manuscripts”. In: *Nature* (2022).

[Wal+00] E. Walsh et al. “Open peer review: a randomised controlled trial”. In: *The British Journal of Psychiatry* (2000).

[Wan+20] Q. Wang et al. “ReviewRobot: Explainable Paper Review Generation based on Knowledge Synthesis”. In: *arXiv preprint arXiv:2010.06119* (2020).

[Wan+21] J. Wang et al. “Debiasing Evaluations that are Biased by Evaluations”. In: *AAAI*. 2021.

[War08] M. Ware. “Peer review: benefits, perceptions and alternatives”. In: *Publishing Research Consortium* (2008).

[War10] D. A. Wardle. “Do’Faculty of 1000’(F1000) ratings of ecological publications serve as reasonable predictors of their future impact?” In: *Ideas in Ecology and Evolution* (2010).

[War16] M. Ware. “Publishing Research Consortium Peer review survey 2015”. In: *Publishing Research Consortium* (2016).

[Was12] L. Wasserman. “A world without referees”. In: *ISBA Bulletin* (2012).

[WC11] J. M. Wing and E. H. Chi. “Reviewing peer review”. In: *Communications of the ACM* (2011).

[Web+02] E. J. Weber et al. “Author perception of peer review: impact of review quality and acceptance on satisfaction”. In: *JAMA* (2002).

[Wei+24] A. S. Weitzner et al. "How predictive is peer review for gauging impact? The association between reviewer rating scores, publication status, and article impact measured by citations in a pain subspecialty journal". In: *Regional Anesthesia & Pain Medicine* (2024).

[Wie+19] J. Wieting et al. "Simple and Effective Paraphrastic Similarity from Parallel Translations". In: *ACL*. Florence, Italy, 2019.

[Wil25] B. Wilder. *Equilibrium Effects of LLM Reviewing*. Online at `https://bryanwilder.github.io/files/llmreviews.html`. Retrieved August 2, 2025. 2025.

[Win11] J. Wing. "Yes, Computer Scientists Are Hypercritical". In: *Communications of the ACM* (2011).

[Woo16] M. Woodhead. *80% of China's clinical trial data are fraudulent, investigation finds*. 2016.

[WS19a] J. Wang and N. Shah. *Gender Distributions of Paper Awards*. Research on Research blog. `https://researchonresearch.blog/2019/06/18/gender-distributions-of-paper-awards/`. 2019.

[WS19b] J. Wang and N. B. Shah. "Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings". In: *AAMAS*. 2019.

[Wu+21] R. Wu et al. "Making Paper Reviewing Robust to Bid Manipulation Attacks". In: *ICML*. 2021.

[WW01] C. Wenneras and A. Wold. "Nepotism and sexism in peer-review". In: *Women, sience and technology: A reader in feminist science studies* (2001).

[XDS14] Y. Xiao, F. Dörfler, and M. van der Schaar. "Rating and matching in peer review systems". In: *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 2014.

[XDVDS18] Y. Xiao, F. Dörfler, and M. Van Der Schaar. "Incentive design in peer review: Rating and repeated endogenous matching". In: *IEEE Transactions on Network Science and Engineering* (2018).

[Xu+19] Y. Xu et al. "On Strategyproof Conference Review". In: *IJCAI*. 2019.

[Xu+24] Y. Xu et al. "A one-size-fits-all approach to improving randomness in paper assignment". In: *Advances in Neural Information Processing Systems* (2024).

[YLN21] W. Yuan, P. Liu, and G. Neubig. "Can We Automate Scientific Reviewing?" In: *arXiv preprint arXiv:2102.00176* (2021).

[Zar+13] R. Zarychanski et al. "Association of hydroxyethyl starch administration with mortality and acute kidney injury in critically ill patients requiring volume resuscitation: a systematic review and meta-analysis". In: *Jama* (2013).

[Zha+22] Y. Zhang et al. "A System-Level Analysis of Conference Peer Review". In: *Proceedings of the 23rd ACM Conference on Economics and Computation*. 2022.