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

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

## 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.

## 1 INTRODUCTION

Peer review is a cornerstone of scientific research [2]. Although quite ubiquitous today, peer review in its current form became popular only in the middle of the twentieth century [3, 4]. Peer review looks to assess research in terms of its competence, significance and originality [5]. It aims to ensure quality control to reduce misinformation and confusion [6] thereby upholding the integrity of science and the public trust in science [7]. It also helps in improving the quality of the published research [8]. In the presence of an overwhelming number of papers written, peer review also has another role [9]: "Readers seem to fear the firehose of the internet: they want somebody to select, filter, and purify research material."

Surveys [10–14] 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" [8]. 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" [15]. There are, however, key challenges towards these goals. The following quote from Rennie [16], 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."*

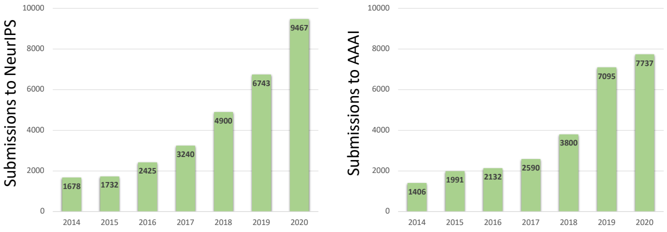


**Figure 1: Number of submissions to two prominent conferences over the past few years.**

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 [17–19]. As noted in [20] *"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 question [21]: *"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"* [22].

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:

This is an extended version of an article in the Communications of the ACM [1]. This is an evolving draft. The first version appeared on July 8, 2021. The current version is dated August 31, 2022. Associated slides are available at https://www.cs.cmu.edu/~nihars/tutorials/Shah_Tutorial_PeerReview_2021.pdf.

[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.

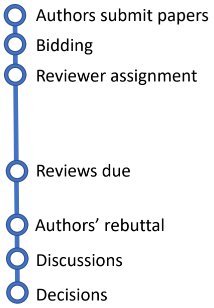


**Figure 2: Typical timeline of the review process in computer science conferences.**

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.

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 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 entire pool of 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: mismatched reviewer expertise (Section 3), dishonest behavior (Section 4), miscalibration (Section 5), subjectivity (Section 6), biases pertaining to author identities (Section 7), incentives (Section 8), and norms and policies (Section 9).

## 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"* [23]. Time and again, a top reason for authors to be dissatisfied with reviews is the mismatch of the reviewers' expertise with the paper [24].

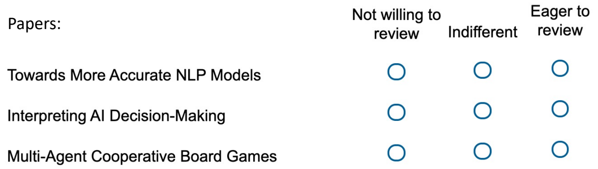
For small conferences, the program chairs may assign reviewers themselves. However, this approach does not scale to conferences with hundreds or thousands of papers. One may aim to have meta reviewers assign reviewers, but this approach has two problems. First, papers handled by meta reviewers who do 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



| 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.**

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 [25–35]. We overview a couple of approaches here [27, 28]. 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.

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) [28] additionally use reviewer-provided confidence scores for each review to improve the similarity computation via supervised learning. The paper [35] builds language models using citations as a form of supervision.

The design of algorithms to compute similarities more accurately through advances in natural language processing is an active area of research [36].

*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.

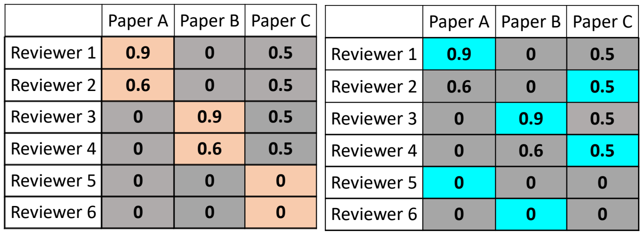
Cabanac and Preuss [37] analyze the bids made by reviewers in several conferences. In these conferences, along with each review, the reviewer is also asked to report their confidence in their evaluation. They find that assigning papers for which reviewers have made positive (willing) bids is associated 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 [2], 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 [23]. In IMC 2010, 68% of the papers had no positive bids [38]. 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 [39]. 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 [37] 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 [40] 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 [41] 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). The paper [42] 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.



| | 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).**

*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 [39]. 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 [28, 43–48]:

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

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 [49]. As a toy example illustrating this issue, consider a conference with three papers and six reviewers, where each paper is assigned one reviewer and each reviewer is assigned two papers. 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 [50] analyzes data from the Computer Vision and Pattern Recognition (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 [49–52] that optimizes for the paper which is worst-off in terms of the similarities of its assigned reviewers:

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

The approach then optimizes for the paper that is the next worst-off and so on. Evaluations [49, 50] 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 [49] 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 [53]. This approach is now adopted in conferences such as the International Conference on Machine Learning (ICML) 2020 [49].

Recent work also incorporates various other desiderata in the reviewer-paper assignments such as geographic diversity [54] and envy-freeness [53]. See the paper [55] for a survey of researhers 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. A number of dishonest behaviors are well documented in various fields of research, including selling authorship [56], faking reviewer identities [57, 58], plagiarism [59], data fabrication [60–63], fake paper mills [64], multiple submissions [65], stealing confidential information from grant proposals submitted for review [66, 67], breach of confidentiality [68] and others [69–71]. In this article we focus on some issues that are more closely tied to conference peer review.

## 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.

A controlled study by Balietti et al. [72] 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

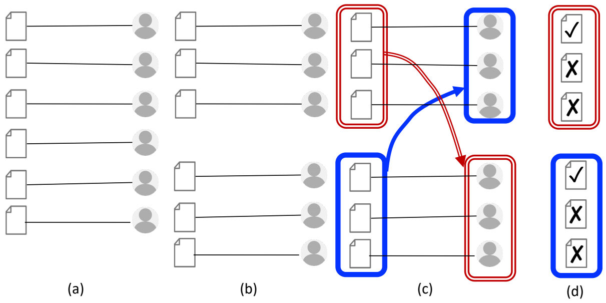


**Figure 5: Partition-based method for strategyproofness.**

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 [73]. See [74–76] for more anecdotes and [77] for a dataset comprising such strategies. The paper [78] 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 [79] and studied subsequently in many papers [80–87]. 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 [85, 87] showing that such partitions do exist in practice and the reduction in quality of assignment may not be too drastic.

## 4.2 Coalitions

Several recent investigations have uncovered dishonest coalitions in peer review [88–90]. Here a reviewer and an author come to an understanding: the reviewer manipulates the system to try to be assigned the author's paper, then accepts the paper if assigned, and the author offers quid pro quo either in the same conference or elsewhere. There may be coalitions between more than two people, where a group of reviewers (who are also authors) illegitimately push for each others' papers. Problems of this nature are also reported in grant peer review [91, 92].[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 [88]: *"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)."*

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 [93, 94] 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 [54]. 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 [91]. 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 [95, 96]. This suggests curtailing or auditing bids, and this approach is followed in the paper [96]. 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

[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.

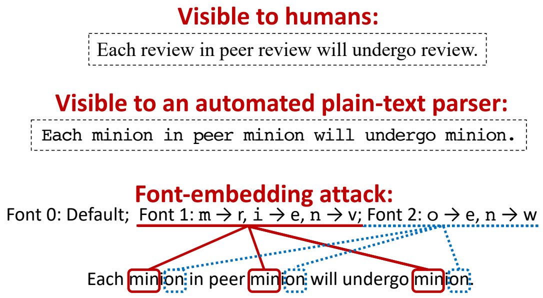

Visible to humans:
Each review in peer review will undergo review.
Visible to an automated plain-text parser:
Each minion in peer minion will undergo minion.
Font-embedding attack:
Font 0: Default; Font 1: m → r, i → e, n → v; Font 2: o → e, n → w
Each minion in peer minion will undergo minion.

**Figure 6: An attack on the assignment system via font embedding in the PDF of the submitted paper [99, 100]. 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.**

by de-emphasizing bids that are significantly different from the remaining data.

A challenge with the aforementioned method [96], however, is that there remains only little influence of the bids (of honest reviewers) on the choice of papers assigned to them [97]. 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 [98].

Security researchers have demonstrated the vulnerability of paper assignment systems to attacks where an author could manipulate the PDF (portable document format) of their submitted paper so that a certain reviewer gets assigned [99, 100]. 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 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. In practice, there may be other attacks used by malicious participants beyond what program chairs and security researchers have detected to date.

In some cases, the colluding reviewers may naturally be assigned to the target papers without any manipulation of the assignment process [88]: *"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 [95] 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 [54].

The final defense we discuss [95] 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 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). This defence was used in the AAAI 2022 conference.

There are various tradeoffs between the aforementioned approaches, discussed in [97]. 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 [104].

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 [105] 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

[3] This assignment procedure also mitigates potential "torpedo reviewing" [101] 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" [102] of the evaluation of any paper.

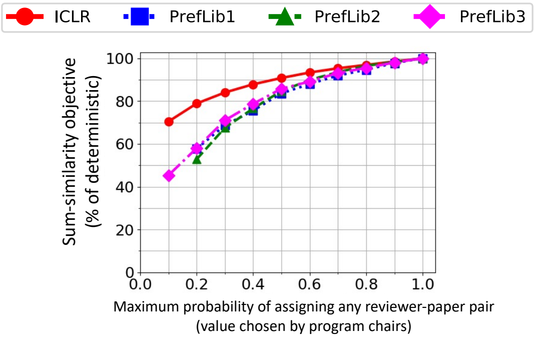


**Figure 7: Trading off the quality of the assignment (sum similarity on y-axis) with the amount of randomness (value specified by program chairs on x-axis) to mitigate dishonest coalitions [95]. The similarity scores for the "ICLR" plot are reconstructed [85] via text-matching from the International 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 [103].**

author is found to indulge in dishonest behavior including blatant plagiarism. This raises concerns of lack of transparency [106], 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 [107, 108].

### 4.3 Temporary plagiarism

Issues of plagiarism [69]—where an author copies another paper without appropriate attribution—are well known and have existed for many years. Here we discuss an incident in computer science that involved an author taking plagiarism to a new level.

The author in contention wrote a paper. Then 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.

## 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 [109–115] 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"* [109]. 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 [111]: *"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 expectations and the ratings given by reviewers [39]. 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" [116, 117]. 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 [118] 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 [112] 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 [119–125] 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 $\theta_p$, 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

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

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| | (low or very low) | (sub-standard) | (poster level: top 30%) | (oral level: top 3%) | (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 [39]. The column headings contain the guidelines provided to reviewers.**

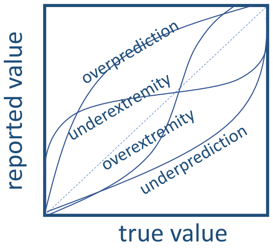


**Figure 8: A caricature of a few types of miscalibration [127]. The diagonal line represents perfect calibration. An affine (or linear) miscalibration would result in a straight line.**

These algorithms then use the ratings to estimate the "true" paper ratings $\theta$, and possibly also reviewer parameters.[4]

The simplistic assumptions described above are frequently violated in the real world [114, 127]; 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 [128].

One exception to the simplistic-modeling approach is the paper [129] 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 $\theta_p$ and $f_r$ which best fit the review data.

A second popular approach [109, 110, 113, 115, 130, 131] 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 combine the rankings of many viewers who often use completely different ranges of scores to express identical preferences"* [110].

Ratings can provide some information even in isolation. It was shown recently [132] 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 [132] 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 sxame reviewer, the reviewer gave an identical rating to both papers for 40% of the pairs [39]. 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 [39]. Given the tradeoffs between rankings and ratings, the papers [136, 137] 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 [138] that allow conferences to share some reviewer-specific calibration data with one another in order to calibrate better.

## 6 SUBJECTIVITY

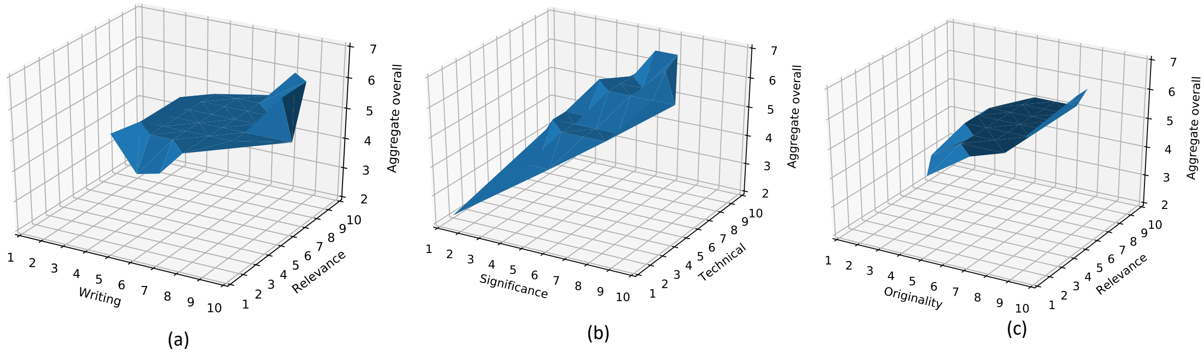
A number of issues pertaining to reviewers' personal subjective preferences exist in peer review. We begin with a discussion on commensuration bias towards which several approaches have been proposed, including a computational mitigating technique. We then discuss other issues pertaining to subjectivity 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

[4] The paper [126] considers this model but assumes $a_r = 1$, treats the noise term as the reviewer's subjective opinion, and estimates $\theta_p$ + noise as a calibrated review score.

[5] Interestingly, randomized decisions are used in practice by certain funding agencies to allocate grants [133, 134]. Such randomized decision-making has found support among researchers [135] 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 [135].



(a)

(b)

(c)

**Figure 9: Mapping of individual criteria to overall ratings by reviewers in IJCAI 2017 [139]. 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.**

judging papers [140–144]. 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" [21].

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 [145], 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 [139], 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 [146]. 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. [139] propose an algorithmic solution to this problem. They consider an often-recommended [126, 147–149] 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 [139] 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 controlled study by Mahoney [140] 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 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, as having a better data presentation, and the reviewer was less likely to catch mistakes in the paper, even though these components were identical across the manuscripts.[6] Confirmation biases have also been found in other studies [151, 152].

A related challenge is that of "homophily," that is, reviewers often favor topics which are familiar to them [153–155]. For instance, a study [153] 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 [156], reviewers "simply do not fight so hard for subjects that are not close to their hearts". In contrast, the paper [157] 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 [158] asked reviewers 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. No wonder researchers often make tongue-in-cheek remarks about "acceptance via obfuscation"!

## 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 [159] 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 foresight subjects were asked to assess the surprisingness of the result obtained.

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 [141]: *"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"* [118]. The bias against papers that are novel but imperfect can incentivize researchers to work on only mediocre ideas [102].

The paper [157] presents an evaluation the effects of novelty of submissions 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.

## 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 [160] 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

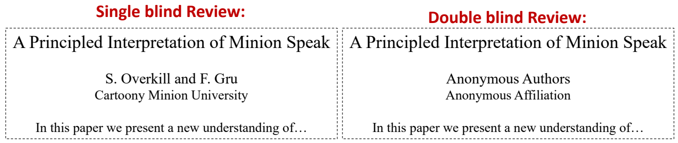
[6] According to Mahoney [150], for this study, "the emotional intensity and resistance of several participants were expressed in the form of charges of ethical misconduct and attempts to have me fired. Several editors later informed me that correspondence from my office was given special scrutiny for some time thereafter to ascertain whether I was secretly studying certain parameters of their operation."

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 [161].

## 6.7 Interdisciplinary research

Interdisciplinary research is considered a bigger evaluation challenge, and at a disadvantage, as compared to disciplinary research [2, 142, 153, 154, 156, 162–165]. 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 [153, 165]. Second, if there are such reviewers, there may be only a few in that interdisciplinary area, thereby “leading to dangers of inbreeding” [153]. 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 [162]. Moreover, a multidisciplinary review team can have difficulties reconciling different perspectives [162]. 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” [163]. Consequently, “interdisciplinary work needs to simultaneously satisfy expert criteria in its disciplines as well as generalist criteria” [142].

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 [153]. 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) [166]. The aforementioned discussion suggests this approach will disadvantage interdisciplinary papers [153]. 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.



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.**

# 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)”* [167]. 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 interest, and the use of author identities as a guarantee of trust for the details that reviewers have not been able to check carefully. In addition, the debate over single-vs-double blind reviewing rests on the frequently-asked question: “Where is the evidence of bias in single-blind reviewing in my field of research?”

*Experiments in computer science.* In the conference-review setting, a remarkable experiment was conducted at the Web Search and Data Mining (WSDM) 2017 conference [168] 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.

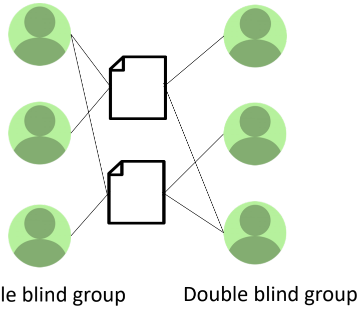


**Figure 11: The WSDM 2017 experiment [168] comparing single and double blind reviewing.**

Another study [169] 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]

*Design of experiments.* Such studies 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 [178] 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 [179]. The paper [180] discusses some challenges in the methods employed in the aforementioned WSDM experiment and provides a framework for design of such experiments. The paper [181] 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.

*De-anonymization of authors in double blind.* Making reviewing double blind can mitigate these biases, but may not fully eliminate them. Reviewers in three double-blind conferences were asked to guess the authors of the papers they were reviewing [182]. 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 [183]. 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 [184]. Furthermore, the study [184] 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 [185–187] design algorithms that can identify author identity or affiliations to a moderate degree based on the content of the paper. The aforementioned survey [182] 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.

*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 [188] finds gender bias, [189] finds biases with respect to gender and personal connections, the study [190] finds bias with respect to race, whereas the study [191] finds little evidence of gender or racial bias. Several studies [192–195] find bias in favor of authors' status. In particular, [195] 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 [196] finds weak evidence of country and institution bias when scientists evaluate abstracts. Bias with respect to author fame is also investigated in the paper [193], which finds that the top and bottom institutions' papers unaffected, but those in the middle were affected. In a similar vein, the study [197] 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 [198] found that authors with more past papers were given better scores by blinded reviewers. The paper [199] finds an increased representation of women authors following a policy change from single to double blind. The study [200] finds that blinding reviewers to the author's identity does not usefully improve the quality of reviews. Surveys of researchers [12, 13] reveal that double blind review is preferred and perceived as most effective.

Finally, studies [201–203] have found a significant gender skew in terms of representation in computer science conferences. These studies provide valuable quantitative information towards policy choices and tradeoffs on blinded reviewing.

[7] On the topic of analysis of review text, some recent works analyze arguments [170–173] and sentiments [174–177] in the text of reviews and discussions. With tremendous progress in natural language processing in recent times, there is a wide scope for much more research on evaluating various aspects of the review process via deeper studies of the *text* of the reviews.

# 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

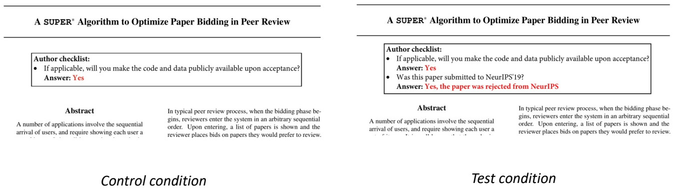
It is said that authors submitting a below-par paper have little to lose but lots to gain: very few people will see the below-par version if it gets rejected, whereas the arbitrariness in the peer-review process gives it some chance of acceptance. The rapid increase in the number of submissions in various conferences has prompted policies that incentivize authors to submit papers only when they are of suitably high quality [102].

*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 and f1000research.com, where the latter is one of the few venues that also publishes reviewer identities. A survey [204] 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 [102]. Additionally, the public nature of the reviews has yielded useful datasets for research on peer review [169, 205–210].

Alongside these benefits, the open-review format can also result in some drawbacks. We discuss one such issue next, related to public visibility of rejected papers.

*8.1.2 Resubmission policies.* It is estimated that every year, 15 million hours of researchers' time is spent in reviewing papers that are eventually rejected [211]. A large fraction of papers accepted at top conferences are previously rejected at least once [212]. 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 [213] 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 resbumission and when (from the beginning or after they submit their initial review).



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 [213]: paper shown to reviewers in the control and test conditions.**

*8.1.3 Rolling deadlines.* In conferences with a fixed deadline, a large fraction of submissions are made on or very near the deadline [204]. 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 [214]. 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 Using authors' opinions to make decisions.* The paper [215] 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 proposed method can incentivize authors to falsely report their ranking of their own papers, which can in turn lead to poorer quality papers being accepted.

## 8.2 Reviewer incentives

*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 [216] and lead to other problems.

PeerJ journals [217] award contribution points to reviewers for each review. Some research communities also use a commercial system called Publons where reviewers receive points to review a paper. However, there is mixed evidence of its usefulness, and moreover, there is evidence of reviewers trying to get points by providing superficial or poor reviews [218].

Squazzoni et al. [219] 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 the 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.

Among non-materialistic incentives, a survey [216] 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 [220] in order to allocate points for high-quality reviews. A key concern though is that if the points will be applied in any high-stakes setting, then the biases and problems in reviewing papers are likely to arise in reviewing of reviewers as well. The Transactions on Machine Learning Research (TMLR) journal has formed a lower-stakes carrot-based policy as a middle ground: meta-reviewers will evaluate reviews, and reviewers with the highest quality reviews will have their own-authored papers highlighted as a reward for good reviewing.

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 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. *"Satisfaction [of the author with the review] had a strong, positive association with acceptance of the manuscript for publication... Quality of the review of the manuscript was not associated with author satisfaction" [221].* See also [222–225] for more evidence of this bias, [226] for a case where no such bias was found, and [227] for some initial work on debiasing this bias.

*8.2.3 Game-theoretic approaches.* The papers [228, 229] 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 [230] 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 [231] randomized controlled trial also did not find a significant difference in the review quality. The study [232] 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 [233] 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 [22]. 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 [234] and IEEE Symposium on Security and Privacy (S&P) 2017 [235]. Recently, 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 [236]. 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 [237, 238] 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. Another study [239] – a randomized controlled trial – 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 [240] find that there are no easily identifiable types of formal training or experience that could predict reviewers' review quality.

*9.1.2 Evaluating 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.

Baxt et al. [241] 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 the submission of a revised version of the manuscript. The number of errors identified differed significantly across recommended disposition. The reviewers identified one-third of the major errors on average, but failed to identify two-thirds of the major errors. Moreover, 68% of the reviewers did not realize that the conclusions of the manuscript were not supported by the results.

Godlee et al [231] modified a manuscript to deliberately introduce 8 weaknesses. This modified manuscript was reviewed by over 200 reviewers, who on average identified 2 weaknesses. There was no difference in terms of single versus double blind reviewing and in terms of whether reviewer names were revealed publicly.

These results suggest that while peer review does filter out bad science to some extent, perhaps more emphasis may need to be placed on evaluating correctness rather than interestingness.

*9.1.3 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 [242] 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.

*9.1.4 Review timeliness.* Review timeliness is a major issue in journals due to the (perceived) flexibility of the review-submission timeline [243, 244], and there are also concerns about reviewers working on a competing idea unethically delaying their review [6, 68, 74]. 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 [126] 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 [226]. 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 [39]. 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 [245].

The paper [246] 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 [212] 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 [235] 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 [153] in the context of evaluating interdisciplinary papers. They suggested that reviewers should not be asked to 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

[8] The paper [246] 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 by Teplitsky et al. [247] on this topic, discussed in Section 9.3.

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 [226] 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 [248] between these modes. We discuss a few studies on this topic.

*9.3.1 Do panel discussions improve consistency?* Several studies [147, 249, 250] 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.[9]

The paper [253] 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."*

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.

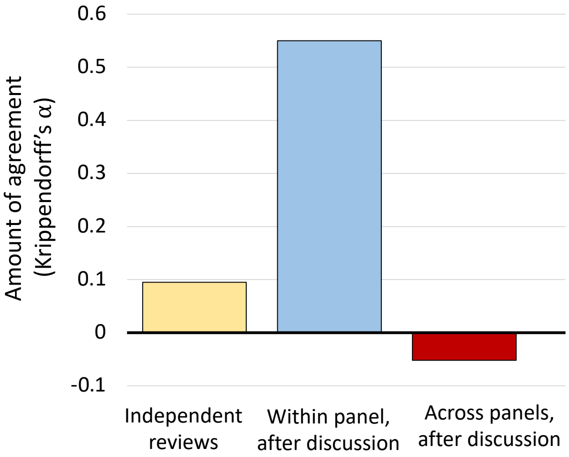


**Figure 13: Amount of agreement before and after discussions [250] in terms of the Krippendorff's alpha coefficient $\alpha = 1 - \frac{\textbf{amount of observed disagreement}}{\textbf{amount of disagreement expected by chance}}$. (Left bar) independent reviews submitted by reviewers have a low agreement, (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. [247] 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.

*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 [254]. 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

[9] In computer science, an experiment was carried out at the NeurIPS 2014 conference [126, 251] 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 [252] found that the levels of inconsistency were consistent with 2014 despite an order of magnitude increase in the number of submissions.

current absence of any policy for choosing the discussion initiator does not hurt.

*9.3.4 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 [235]. 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 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 [102, 141], various problems surrounding the reproducibility crisis [255, 256] (including HARKing [257] and data withholding [258]), desk rejection [183], socio-political issues [259], post-publication review [260], reviewer forms [39, 261], two-stage reviewing [54, 181, 262–264], alternative modes of reviewing [7, 265–267], and others [146], including calls to abolish peer review altogether [268].

Finally, there are a few attempts [208, 269, 270] to create AI algorithms that can review the entire paper. Such a goal seems quite far at this point, with current attempts not being successful. However, AI can still be used to evaluate specific aspects of the paper such as ensuring it adheres to appropriate submission and reporting guidelines [271, 272] and to mitigate fraud by finding duplicated images [273].

## 10 CONCLUSIONS

There are many challenges of biases and unfairness in peer review. Improving peer review is sometimes characterized as a "fundamentally difficult problem" [274]: *"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."* Current research on peer review aims to understand the challenges in rigorous manner and make fundamental systematic improvements to the process (via design of computational tools or other approaches). The current research on improving peer 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.

Research on peer review faces at least two overarching challenges. First, there is no "ground truth" regarding which papers should have been accepted to the conference under an "ideal" peer-review process. 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"* [112, 275]. One can 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. Proxies such as subsequent citations (of accepted versus rejected papers) are sometimes employed, but they face a slew of other biases and problems [102, 276–280].

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"* [72]. 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 [168] 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 [95, 281].

Nevertheless, there is increasing interest among research communities and conferences in improving peer review in a scientific manner. Researchers are conducting a number of experiments to understand issues and implications in peer review, designing methods and policies to address the various challenges, and translating research on this topic into practice. This bodes well for peer review, the cornerstone of scientific 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

[1] N. B. Shah. 2022. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM.*

[2] S. Price and P. A. Flach. 2017. Computational support for academic peer review: a perspective from artificial intelligence. *Communications of the ACM.*

[3] R. Spier. 2002. The history of the peer-review process. *TRENDS in Biotechnology.*

[4] M. Baldwin. 2018. Scientific autonomy, public accountability, and the rise of "peer review" in the cold war united states. *Isis.*

[5] T. Brown. 2004. *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.

[6] D. J. Benos, E. Bashari, J. M. Chaves, A. Gaggar, N. Kapoor, M. LaFrance, R. Mans, D. Mayhew, S. McGowan, A. Polter, et al. 2007. The ups and downs of peer review. *Advances in physiology education.*

[7] J. M. Wing and E. H. Chi. 2011. Reviewing peer review. *Communications of the ACM.*

[8] T. Jefferson, P. Alderson, E. Wager, and F. Davidoff. 2002. Effects of editorial peer review: a systematic review. *Jama.*

[9] R. Smith. 1997. Peer review: reform or revolution?: time to open up the black box of peer review. (1997).

[10] M. Ware. 2016. Publishing research consortium peer review survey 2015. *Publishing Research Consortium.*

[11] Taylor and Francis group. 2015. Peer review in 2015 a global view. https://authorservices.taylorandfrancis.com/publishing-your-research/peer-review/peer-review-global-view/. (2015).

[12] M. Ware. 2008. Peer review: benefits, perceptions and alternatives. *Publishing Research Consortium.*

[13] A. Mulligan, L. Hall, and E. Raphael. 2013. Peer review in a changing world: an international study measuring the attitudes of researchers. *Journal of the Association for Information Science and Technology.*

[14] D. Nicholas, A. Watkinson, H. R. Jamali, E. Herman, C. Tenopir, R. Volentine, S. Allard, and K. Levine. 2015. Peer review: still king in the digital age. *Learned Publishing.*

[15] C. J. Lee, C. R. Sugimoto, G. Zhang, and B. Cronin. 2013. Bias in peer review. *Journal of the Association for Information Science and Technology.*

[16] D. Rennie. 2016. Let's make peer review scientific. *Nature.*

[17] R. K. Merton. 1968. The Matthew effect in science. *Science.*

[18] W. Thorngate and W. Chowdhury. 2014. By the numbers: track record, flawed reviews, journal space, and the fate of talented authors. In *Advances in Social Simulation.* Springer.

[19] F. Squazzoni and C. Gandelli. 2012. Saint Matthew strikes again: an agent-based model of peer review and the scientific community structure. *Journal of Informetrics.*

[20] C. R. Triggle and D. J. Triggle. 2007. 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?" *Vascular health and risk management.*

[21] C. J. Lee. 2015. Commensuration bias in peer review. *Philosophy of Science.*

[22] A. McCook. 2006. 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? *The scientist.*

[23] M. A. Rodriguez, J. Bollen, and H. Van de Sompel. 2007. Mapping the bid behavior of conference referees. *Journal of Informetrics.*

[24] J. McCullough. 1989. First comprehensive survey of NSF applicants focuses on their concerns about proposal review. *Science, Technology, & Human Values.*

[25] S. T. Dumais and J. Nielsen. 1992. 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.*

[26] S. Ferilli, N. Di Mauro, T. M. A. Basile, F. Esposito, and M. Biba. 2006. Automatic topics identification for reviewer assignment. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems.* Springer.

[27] D. Mimno and A. McCallum. 2007. Expertise modeling for matching papers with reviewers. In *KDD.*

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

[29] X. Liu, T. Suel, and N. Memon. 2014. A robust model for paper reviewer assignment. In *ACM Conference on Recommender Systems.*

[30] M. A. Rodriguez and J. Bollen. 2008. An algorithm to determine peer-reviewers. In *ACM Conference on Information and Knowledge Management.*

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

[32] O. Anjum, H. Gong, S. Bhat, W.-M. Hwu, and J. Xiong. 2019. Pare: a paper-reviewer matching approach using a common topic space. In *EMNLP-IJCNLP.*

[33] W. E. Kerzendorf. 2019. Knowledge discovery through text-based similarity searches for astronomy literature. *Journal of Astrophysics and Astronomy.*

[34] J. Wieting, K. Gimpel, G. Neubig, and T. Berg-Kirkpatrick. 2019. Simple and effective paraphrastic similarity from parallel translations. In *ACL.* Florence, Italy.

[35] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld. 2020. Specter: document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*

[36] G. Neubig, J. Wieting, A. McCarthy, A. Stent, N. Schluter, and T. Cohn. 2019–2021 and ongoing. ACL reviewer matching code. https://github.com/acl-org/reviewer-paper-matching. (2019–2021 and ongoing).

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

[38] R. Beverly and M. Allman. 2012. Findings and implications from data mining the IMC review process. *ACM SIGCOMM Computer Communication Review.*

[39] N. Shah, B. Tabibian, K. Muandet, I. Guyon, and U. Von Luxburg. 2018. Design and analysis of the NIPS 2016 review process. *JMLR.*

[40] J. Murphy, C. Hofacker, and R. Mizerski. 2006. Primacy and recency effects on clicking behavior. *Journal of Computer-Mediated Communication.*

[41] T Fiez, N Shah, and L Ratliff. 2020. A SUPER* algorithm to optimize paper bidding in peer review. In *Conference on Uncertainty in Artificial Intelligence.*

[42] R. Meir, J. Lang, J. Lesca, N. Kaminsky, and N. Mattei. 2020. A market-inspired bidding scheme for peer review paper assignment. In *Games, Agents, and Incentives Workshop at AAMAS.*

[43] J. Goldsmith and R. Sloan. 2007. The AI conference paper assignment problem.

[44] C. J. Taylor. 2008. On the optimal assignment of conference papers to reviewers.

[45] W. Tang, J. Tang, and C. Tan. 2010. Expertise matching via constraint-based optimization. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.*

[46] L. Charlin, R. S. Zemel, and C. Boutilier. 2012. A framework for optimizing paper matching. *CoRR.*

[47] C. Long, R. Wong, Y. Peng, and L. Ye. 2013. On good and fair paper-reviewer assignment. In *ICDM.*

[48] B. Li and Y. T. Hou. 2016. The new automated ieee infocom review assignment system. *IEEE Network.*

[49] I. Stelmakh, N. Shah, and A. Singh. 2021. PeerReview4All: fair and accurate reviewer assignment in peer review. *JMLR.*

[50] A. Kobren, B. Saha, and A. McCallum. 2019. Paper matching with local fairness constraints. In *ACM KDD.*

[51] N. Garg, T. Kavitha, A. Kumar, K. Mehlhorn, and J. Mestre. 2010. Assigning papers to referees. *Algorithmica.*

[52] J. W. Lian, N. Mattei, R. Noble, and T. Walsh. 2018. The conference paper assignment problem: using order weighted averages to assign indivisible goods. In *Thirty-Second AAAI Conference on Artificial Intelligence.*

[53] J. Payan and Y. Zick. 2021. I will have order! optimizing orders for fair reviewer assignment. *arXiv preprint arXiv:2108.02126.*

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

[55] T. S. T. Jakobsen and A. Rogers. 2022. What factors should paper-reviewer assignments rely on? community perspectives on issues and ideals in conference peer-review. *arXiv preprint arXiv:2205.01005.*

[56] M. Hvistendahl. 2013. China's publication bazaar. (2013).

[57] C. Ferguson, A. Marcus, and I. Oransky. 2014. Publishing: the peer-review scam. *Nature News.*

[58] A. Cohen, S. Pattanaik, P. Kumar, R. R. Bies, A. De Boer, A. Ferro, A. Gilchrist, G. K. Isbister, S. Ross, and A. J. Webb. 2016. Organised crime against the academic peer review system. *British Journal of Clinical Pharmacology.*

[59] G. Helgesson and S. Eriksson. 2015. Plagiarism in research. *Medicine, Health Care and Philosophy.*

[60] M. Woodhead. 2016. 80% of china's clinical trial data are fraudulent, investigation finds. (2016).

[61] D. Fanelli. 2009. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PloS one.*

[62] S. Al-Marzouki, S. Evans, T. Marshall, and I. Roberts. 2005. Are these data real? statistical methods for the detection of data fabrication in clinical trials. *BMJ.*

[63] J. K. Tijdink, R. Verbeke, and Y. M. Smulders. 2014. Publication pressure and scientific misconduct in medical scientists. *Journal of Empirical Research on Human Research Ethics.*

[64] H. Else, R. Van Noorden, et al. 2021. The fight against fake-paper factories that churn out sham science. *Nature.*

[65] K. Bowyer. 1999. Multiple submission: professionalism, ethical issues, and copyright legalities. *IEEE Trans. PAMI.*

[66] L. Tabak and M. R. Wilson. 2018. Foreign influences on research integrity.

[67] S. Murrin. 2020. Nih has acted to protect confidential information handled by peer reviewers, but it could do more.

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

[69] B. C. Martinson, M. S. Anderson, and R. De Vries. 2005. Scientists behaving badly. *Nature.*

[70] M. A. Edwards and S. Roy. 2017. Academic research in the 21st century: maintaining scientific integrity in a climate of perverse incentives and hyper-competition. *Environmental engineering science.*

[71] E. A. Fong and A. W. Wilhite. 2017. Authorship and citation manipulation in academic research. *PloS one.*

[72] S. Balietti, R. Goldstone, and D. Helbing. 2016. Peer review and competition in the art exhibition game. *Proceedings of the National Academy of Sciences.*

[73] P. Naghizadeh and M. Liu. 2013. Incentives, quality, and risks: a look into the NSF proposal review pilot. *arXiv preprint arXiv:1307.6528.*

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

[75] M. S. Anderson, E. A. Ronning, R. De Vries, and B. C. Martinson. 2007. The perverse effects of competition on scientists' work and relationships. *Science and engineering ethics*.

[76] J. Langford. 2008. Adversarial academia. (2008).

[77] I. Stelmakh, N. Shah, and A. Singh. 2021. Catch me if i can: detecting strategic behaviour in peer assessment. In *AAAI*.

[78] S. Thurner and R. Hanel. 2011. Peer-review in a world with rational scientists: toward selection of the average. *The European Physical Journal B*.

[79] N. Alon, F. Fischer, A. Procaccia, and M. Tennenholtz. 2011. Sum of us: strategyproof selection from the selectors. In *Conf. on Theoretical Aspects of Rationality and Knowledge*.

[80] R. Holzman and H. Moulin. 2013. Impartial nominations for a prize. *Econometrica*.

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

[82] F. Fischer and M. Klimm. 2015. Optimal impartial selection. *SIAM Journal on Computing*.

[83] D. Kurokawa, O. Lev, J. Morgenstern, and A. D. Procaccia. 2015. Impartial peer review. In *IJCAI*.

[84] A. Kahng, Y. Kotturi, C. Kulkarni, D. Kurokawa, and A. Procaccia. 2018. Ranking wily people who rank each other. In *AAAI*.

[85] Y. Xu, H. Zhao, X. Shi, and N. Shah. 2019. On strategyproof conference review. In *IJCAI*.

[86] H. Aziz, O. Lev, N. Mattei, J. S. Rosenschein, and T. Walsh. 2019. Strategyproof peer selection using randomization, partitioning, and apportionment. *Artificial Intelligence*.

[87] K. Dhull, S. Jecmen, P. Kothari, and N. B. Shah. 2022. Strategyproofing peer assessment via partitioning: the price in terms of evaluators' expertise. In *HCOMP*.

[88] T. N. Vijaykumar. 2020. Potential organized fraud in ACM/IEEE computer architecture conferences. https://medium.com/@tnvijayk/potential-organized-fraud-in-acm-ieee-computer-architecture-conferences-ccd61169370d. (2020).

[89] M. L. Littman. 2021. Collusion rings threaten the integrity of computer science research. *Communications of the ACM*.

[90] T. N. Vijaykumar. 2020. Potential organized fraud in on-going asplos reviews. (2020).

[91] M. Lauer. 2020. Case study in review integrity: asking for favorable treatment. *NIH Extramural Nexus*.

[92] M. Lauer. 2019. Case study in review integrity: undisclosed conflict of interest. *NIH Extramural Nexus*.

[93] L. Guo, J. Wu, W. Chang, J. Wu, and J. Li. 2018. K-loop free assignment in conference review systems. In *ICNC*.

[94] N. Boehmer, R. Bredereck, and A. Nichterlein. 2021. Combating collusion rings is hard but possible. *arXiv preprint arXiv:2112.08444*.

[95] S. Jecmen, H. Zhang, R. Liu, N. B. Shah, V. Conitzer, and F. Fang. 2020. Mitigating manipulation in peer review via randomized reviewer assignments. In *NeurIPS*.

[96] R. Wu, C. Guo, F. Wu, R. Kidambi, L. van der Maaten, and K. Weinberger. 2021. Making paper reviewing robust to bid manipulation attacks. *arXiv:2102.06020*.

[97] S. Jecmen, N. B. Shah, F. Fang, and V. Conitzer. 2022. Tradeoffs in preventing manipulation in paper bidding for reviewer assignment. In *ICLR workshop on ML Evaluation Standards*.

[98] A. Ailamaki, P. Chrysogelos, A. Deshpande, and T. Kraska. 2019. The sigmod 2019 research track reviewing system. *ACM SIGMOD Record*.

[99] I. Markwood, D. Shen, Y. Liu, and Z. Lu. 2017. Mirage: content masking attack against information-based online services. In *USENIX Security Symposium*.

[100] D. Tran and C. Jaiswal. 2019. 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.

[101] J. Langford. 2012. Bidding problems. https://hunch.net/?p=407 [Online; accessed 6-Jan-2019]. (2012).

[102] T. Anderson. 2009. Conference reviewing considered harmful. *ACM SIGOPS Operating Systems Review*.

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

[104] S. Jecmen, M. Yoon, V. Conitzer, N. B. Shah, and F. Fang. 2022. A dataset on malicious paper bidding in peer review. *arXiv preprint 2207.02303*.

[105] ACM. 2021. Public announcement of the results of the joint investigative committee (JIC) investigation into significant allegations of professional and publications related misconduct. https://www.sigarch.org/wp-content/uploads/2021/02/JIC-Public-Announcement-Feb-8-2021.pdf.

[106] B. Falsafi, N. Enright Jerger, K. Strauss, S. Adve, J. Emer, B. Grot, M. Kim, and J. F. Martinez. 2021. 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).

[107] J. Hilgard. 2021. Crystal prison zone: i tried to report scientific misconduct. how did it go? (2021).

[108] H. Else. 2021. Scientific image sleuth faces legal action for criticizing research papers. *Nature*.

[109] I. Mitliagkas, A. Gopalan, C. Caramanis, and S. Vishwanath. 2011. User rankings from comparisons: learning permutations in high dimensions. In *Allerton Conference*.

[110] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. 2003. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*.

[111] S. Siegelman. 1991. Assassins and zealots: variations in peer review. *Radiology*.

[112] A. Ragone, K. Mirylenka, F. Casati, and M. Marchese. 2013. On peer review in computer science: analysis of its effectiveness and suggestions for improvement. *Scientometrics*.

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

[114] 2008. *Perspectives on probability judgment calibration. Blackwell Handbook of Judgment and Decision Making*. Wiley-Blackwell. Chapter 9.

[115] A.-W. Harzing, J. Baldueza, W. Barner-Rasmussen, C. Barzantny, A. Canabal, A. Davila, A. Espejo, R. Ferreira, A. Giroud, K. Koester, Y.-K. Liang, A. Mockaitis, M. J. Morley, B. Myloni, J. O. Odusanya, S. L. O'Sullivan, A. K. Palaniappan, P. Prochno, S. R. Choudhury, A. Saka-Helmhout, S. Siengthai, L. Viswat, A. U. Soydas, and L. Zander. 2009. Rating versus ranking: what is the best way to reduce response and language bias in cross-national research? *International Business Review*.

[116] M. Y. Vardi. 2010. Hypercriticality. *Communications of the ACM*.

[117] J. Wing. 2011. Yes, computer scientists are hypercritical. *Communications of the ACM*.

[118] J. Naughton. 2010. DBMS research: first 50 years, next 50 years. *Keynote at ICDE*. http://pages.cs.wisc.edu/~naughton/naughtonicde.pptx.

[119] P. Flach, S. Spiegler, B. Golénia, S. Price, J. Guiver, R. Herbrich, T. Graepel, and M. Zaki. 2010. Novel tools to streamline the conference review process: experiences from SIGKDD'09. *SIGKDD Explor. Newsl.*

[120] M. Roos, J. Rothe, and B. Scheuermann. 2011. How to calibrate the scores of biased reviewers by quadratic programming. In *AAAI*.

[121] H. Ge, M. Welling, and Z. Ghahramani. 2013. 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).

[122] S. R. Paul. 1981. Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology*.

[123] Y. Baba and H. Kashima. 2013. Statistical quality estimation for general crowdsourcing tasks. In *KDD*.

[124] A. Spalvieri, S. Mandelli, M. Magarini, and G. Bianchi. 2014. Weighting peer reviewers. In *2014 Twelfth Annual International Conference on Privacy, Security and Trust*. IEEE.

[125] R. S. MacKay, R. Kenna, R. J. Low, and S. Parker. 2017. Calibration with confidence: a principled method for panel assessment. *Royal Society Open Science*.

[126] C. Cortes and N. D. Lawrence. 2021. Inconsistency in conference peer review: revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*.

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

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

[129] S. Tan, J. Wu, X. Bei, and H. Xu. 2021. Least square calibration for peer reviews. *Advances in Neural Information Processing Systems*.

[130] M. Rokeach. 1968. The role of values in public opinion research. *Public Opinion Quarterly*.

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

[132] J. Wang and N. B. Shah. 2019. Your 2 is my 1, your 3 is my 9: handling arbitrary miscalibrations in ratings. In *AAMAS*.

[133] M. Liu, V. Choy, P. Clarke, A. Barnett, T. Blakely, and L. Pomeroy. 2020. The acceptability of using a lottery to allocate research funding: a survey of applicants. *Research integrity and peer review*.

[134] D. S. Chawla. 2021. Swiss funder draws lots to make grant decisions. *Nature*.

[135] A. Philipps. 2021. Research funding randomly allocated? a survey of scientists' views on peer review and lottery. *Science and Public Policy*.

[136] M. Pearce and E. A. Erosheva. 2022. A unified statistical learning model for rankings and scores with application to grant panel review. *arXiv preprint arXiv:2201.02539*.

[137] Y. Liu, Y. Xu, N. B. Shah, and A. Singh. 2022. Integrating rankings into quantized scores in peer review. *arXiv preprint arXiv:2204.03505*.

[138] W. Ding, G. Kamath, W. Wang, and N. B. Shah. 2022. Calibration with privacy in peer review. In *ISIT*.

[139] R. Noothigattu, N. Shah, and A. Procaccia. 2021. Loss functions, axioms, and peer review. *Journal of Artificial Intelligence Research.*

[140] M. J. Mahoney. 1977. Publication prejudices: an experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research.*

[141] K. Church. 2005. Reviewing the reviewers. *Computational Linguistics.*

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

[143] V. Bakanic, C. McPhail, and R. J. Simon. 1987. The manuscript review and decision-making process. *American Sociological Review.*

[144] S. E. Hug and M. Ochsner. 2021. Do peers share the same criteria for assessing grant applications? *arXiv preprint arXiv:2106.07386.*

[145] S. Kerr, J. Tolliver, and D. Petree. 1977. Manuscript characteristics which influence acceptance for management and social science journals. *Academy of Management Journal.*

[146] A. Rogers and I. Augenstein. 2020. What can we do to improve peer review in nlp? *arXiv preprint arXiv:2010.03863.*

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

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

[149] A. Marcoci, A. Vercammen, M. Bush, D. G. Hamilton, A. Hanea, V. Hemming, B. C. Wintle, M. Burgman, and F. Fidler. 2022. Reimagining peer review as an expert elicitation process. *BMC Research Notes.*

[150] D. P. Peters and S. J. Ceci. 1982. Peer-review practices of psychological journals: the fate of published articles, submitted again. *Behavioral and Brain Sciences.*

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

[152] J. J. Koehler. 1993. The influence of prior beliefs on scientific judgments of evidence quality. *Organizational behavior and human decision processes.*

[153] A. L. Porter and F. A. Rossini. 1985. Peer review of interdisciplinary research proposals. *Science, technology, & human values.*

[154] P. Dondio, N. Casnici, F. Grimaldo, N. Gilbert, and F. Squazzoni. 2019. The "invisible hand" of peer review: the implications of author-referee networks on peer review in a scholarly journal. *Journal of Informetrics.*

[155] D. Li. 2017. Expertise versus bias in evaluation: evidence from the nih. *American Economic Journal: Applied Economics.*

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

[157] K. J. Boudreau, E. C. Guinan, K. R. Lakhani, and C. Riedl. 2016. Looking across and looking beyond the knowledge frontier: intellectual distance, novelty, and resource allocation in science. *Management science.*

[158] J. S. Armstrong. 1980. Unintelligible management research and academic prestige. *Interfaces.*

[159] P. Slovic and B. Fischhoff. 1977. On the psychology of experimental surprises. *Journal of Experimental Psychology: Human Perception and Performance.*

[160] G. B. Emerson, W. J. Warme, F. M. Wolf, J. D. Heckman, R. A. Brand, and S. S. Leopold. 2010. Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial. *Archives of internal medicine.*

[161] Y. M. Smulders. 2013. A two-step manuscript submission process can reduce publication bias. *Journal of clinical epidemiology.*

[162] G. Laudel. 2006. Conclave in the Tower of Babel: how peers review interdisciplinary research proposals. *Research Evaluation.*

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

[164] Anonymous. 2013. Is your phd a monster? https://thesiswhisperer.com/2013/09/11/help-i-think-i-have-created-a-monster/. (2013).

[165] S. Frolov. 2021. Quantum computing's reproducibility crisis: majorana fermions. (2021).

[166] O. Nierstrasz. 2000. Identify the champion. *Pattern Languages of Program Design.*

[167] R. Bernstein. 2015. PLOS ONE ousts reviewer, editor after sexist peer-review storm. *Science.*

[168] A. Tomkins, M. Zhang, and W. D. Heavlin. 2017. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences.*

[169] E. Manzoor and N. B. Shah. 2021. Uncovering latent biases in text: method and application to peer review. In *AAAI.*

[170] X. Hua, M. Nikolov, N. Badugu, and L. Wang. 2019. Argument mining for understanding peer reviews. *arXiv preprint arXiv:1903.10104.*

[171] M. Fromm, E. Faerman, M. Berrendorf, S. Bhargava, R. Qi, Y. Zhang, L. Dennert, S. Selle, Y. Mao, and T. Seidl. 2020. Argument mining driven analysis of peer-reviews. *arXiv preprint arXiv:2012.07743.*

[172] N. N. Kennard, T. O'Gorman, A. Sharma, C. Bagchi, M. Clinton, P. K. Yelugam, R. Das, H. Zamani, and A. McCallum. 2021. A dataset for discourse structure in peer review discussions. *arXiv preprint arXiv:2110.08520.*

[173] I. Kuznetsov, J. Buchmann, M. Eichler, and I. Gurevych. 2022. Revise and resubmit: an intertextual model of text-based collaboration in peer review. *arXiv preprint arXiv:2204.10805.*

[174] T. Ghosal, R. Verma, A. Ekbal, and P. Bhattacharyya. 2019. DeepSentiPeer: harnessing sentiment in review texts to recommend peer review decisions. In *ACL.*

[175] S. Chakraborty, P. Goyal, and A. Mukherjee. 2020. Aspect-based sentiment analysis of scientific reviews. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020.*

[176] I. Buljan, D. Garcia-Costa, F. Grimaldo, F. Squazzoni, and A. Marušić. 2020. Meta-research: large-scale language analysis of peer review reports. *eLife.*

[177] A. C. Ribeiro, A. Sizo, H. Lopes Cardoso, and L. P. Reis. 2021. Acceptance decision prediction in peer-review through sentiment analysis. In *EPIA Conference on Artificial Intelligence.* Springer.

[178] S. Madden and D. DeWitt. 2006. Impact of double-blind reviewing on sigmod publication rates. *ACM SIGMOD Record.*

[179] A. K. Tung. 2006. Impact of double blind reviewing on sigmod publication: a more detail analysis. *ACM SIGMOD Record.*

[180] I. Stelmakh, N. Shah, and A. Singh. 2019. On testing for biases in peer review. In *NeurIPS.*

[181] S. Jecmen, H. Zhang, R. Liu, F. Fang, V. Conitzer, and N. B. Shah. 2022. Near-optimal reviewer splitting in two-phase paper reviewing and conference experiment design. In *HCOMP.*

[182] C. Le Goues, Y. Brun, S. Apel, E. Berger, S. Khurshid, and Y. Smaragdakis. 2018. Effectiveness of anonymization in double-blind review. *CACM.*

[183] A. Beygelzimer, E. Fox, F. d'Alché Buc, and H. Larochelle. 2019. What we learned from NeurIPS 2019 data. (2019).

[184] C. Rastogi, I. Stelmakh, X. Shen, M. Meila, F. Echenique, S. Chawla, and N. Shah. 2022. To arxiv or not to arxiv: a study quantifying pros and cons of posting preprints online. *arXiv preprint arXiv:2203.17259.*

[185] S. Hill and F. J. Provost. 2003. The myth of the double-blind review? author identification using only citations. *SIGKDD Explorations.*

[186] C. Caragea, A. Uban, and L. P. Dinu. 2019. 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).*

[187] Y. Matsubara and S. Singh. 2020. Citations beyond self citations: identifying authors, affiliations, and nationalities in scientific papers. In *International Workshop on Mining Scientific Publications.*

[188] L. Bornmann, R. Mutz, and H.-D. Daniel. 2007. Gender differences in grant peer review: a meta-analysis. *Journal of Informetrics.*

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

[190] D. K. Ginther, W. T. Schaffer, J. Schnell, B. Masimore, F. Liu, L. L. Haak, and R. Kington. 2011. Race, ethnicity, and nih research awards. *Science.*

[191] P. S. Forscher, W. T. Cox, M. Brauer, and P. G. Devine. 2019. Little race or gender bias in an experiment of initial review of nih r01 grant proposals. *Nature human behaviour.*

[192] K. Okike, K. T. Hug, M. S. Kocher, and S. S. Leopold. 2016. Single-blind vs double-blind peer review in the setting of author prestige. *Jama.*

[193] R. M. Blank. 1991. The effects of double-blind versus single-blind reviewing: experimental evidence from the american economic review. *The American Economic Review.*

[194] J. S. Ross, C. P. Gross, M. M. Desai, Y. Hong, A. O. Grant, S. R. Daniels, V. C. Hachinski, R. J. Gibbons, T. J. Gardner, and H. M. Krumholz. 2006. Effect of blinded peer review on abstract acceptance. *Jama.*

[195] J. M. Garfunkel, M. H. Ulshen, H. J. Hamrick, and E. E. Lawson. 1994. Effect of institutional prestige on reviewers' recommendations and editorial decisions. *JAMA.*

[196] M. W. Nielsen, C. F. Baker, E. Brady, M. B. Petersen, and J. P. Andersen. 2021. Meta-research: weak evidence of country-and institution-related status bias in the peer review of abstracts. *Elife.*

[197] N. Fouad, S. Brehm, C. I. Hall, M. E. Kite, J. S. Hyde, and N. F. Russo. 2000. Women in academe: two steps forward, one step back. *Report of the Task Force on Women in Academe, American Psychological Association.*

[198] M. Fisher, S. B. Friedman, and B. Strauss. 1994. The effects of blinding on acceptance of research papers by peer review. *Jama.*

[199] A. E. Budden, T. Tregenza, L. W. Aarssen, J. Koricheva, R. Leimu, and C. J. Lortie. 2008. Double-blind review favours increased representation of female authors. *Trends in ecology & evolution.*

[200] S. Goldbeck-Wood. 1999. Evidence on peer review—scientific quality control or smokescreen? *BMJ: British Medical Journal.*

[201] J. Wang and N. Shah. 2019. Gender distributions of paper awards. Research on Research blog. https://researchonresearch.blog/2019/06/18/gender-distributions-of-paper-awards/. (2019).

[202] S. Mattauch, K. Lohmann, F. Hannig, D. Lohmann, and J. Teich. 2020. A bibliometric approach for detecting the gender gap in computer science. *Communications of the ACM.*

[203] E. Frachtenberg and R. Kaner. 2020. Representation of women in high-performance computing conferences. Technical report. EasyChair.

[204] D. Soergel, A. Saunders, and A. McCallum. 2013. Open scholarship and peer review: a time for experimentation.

[205] D. Kang, W. Ammar, B. Dalvi, M. van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz. 2018. A dataset of peer reviews (peerread): collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*.

[206] D. Tran, A. Valtchanov, K. Ganapathy, R. Feng, E. Slud, M. Goldblum, and T. Goldstein. 2020. An open review of openreview: a critical analysis of the machine learning conference review process. *arXiv preprint arXiv:2010.05137*.

[207] H. Bharadhwaj, D. Turpin, A. Garg, and A. Anderson. 2020. De-anonymization of authors through arxiv submissions during double-blind review. *arXiv preprint arXiv:2007.00177*.

[208] W. Yuan, P. Liu, and G. Neubig. 2021. Can we automate scientific reviewing? *arXiv preprint arXiv:2102.00176*.

[209] E. Mohammadi and M. Thelwall. 2013. Assessing non-standard article impact using f1000 labels. *Scientometrics*.

[210] D. A. Wardle. 2010. Do'faculty of 1000'(f1000) ratings of ecological publications serve as reasonable predictors of their future impact? *Ideas in Ecology and Evolution*.

[211] The AJE Team. 2013. Peer review: how we found 15 million hours of lost time. (2013).

[212] E. Frachtenberg and N. Koster. 2020. A survey of accepted authors in computer systems conferences. *PeerJ Computer Science*.

[213] I. Stelmakh, N. Shah, A. Singh, and H. Daumé III. 2021. Prior and prejudice: the novice reviewers' bias against resubmissions in conference peer review. In *CSCW*.

[214] E. Hand. 2016. No pressure: NSF test finds eliminating deadlines halves number of grant proposals. *Science*.

[215] W. Su. 2021. You are the best reviewer of your own papers: an owner-assisted scoring mechanism. *Advances in Neural Information Processing Systems*.

[216] S. Nobarany, K. S. Booth, and G. Hsieh. 2016. What motivates people to review articles? the case of the human-computer interaction community. *Journal of the Association for Information Science and Technology*.

[217] 2022. Peerj - user contribution. Last retrieved: March 30, 2022. (2022).

[218] J. Teixeira da Silva and A. Al-Khatib. 2017. The Clarivate Analytics acquisition of Publons–an evolution or commodification of peer review? *Research Ethics*.

[219] F. Squazzoni, G. Bravo, and K. Takács. 2013. Does incentive provision increase the quality of peer review? an experimental study. *Research Policy*.

[220] I. Arous, J. Yang, M. Khayati, and P. Cudré-Mauroux. 2021. Peer grading the peer reviews: a dual-role approach for lightening the scholarly paper review process.

[221] E. J. Weber, P. P. Katz, J. F. Waeckerle, and M. L. Callaham. 2002. Author perception of peer review: impact of review quality and acceptance on satisfaction. *JAMA*.

[222] S. Van Rooyen, F. Godlee, S. Evans, R. Smith, and N. Black. 1999. Effect of blinding and unmasking on the quality of peer review. *Journal of general internal medicine*.

[223] W. E. Kerzendorf, F. Patat, D. Bordelon, G. van de Ven, and T. A. Pritchard. 2020. Distributed peer review enhanced with natural language processing and machine learning. *Nature Astronomy*.

[224] K. Papagiannaki. 2007. Author feedback experiment at PAM 2007. *SIGCOMM Comput. Commun. Rev.*

[225] A. Khosla, D. Hoiem, and S. Belongie. 2013. Analysis of reviews for CVPR 2012.

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

[227] J. Wang, I. Stelmakh, Y. Wei, and N. Shah. 2021. Debiasing evaluations that are biased by evaluations. In *AAAI*.

[228] S. Srinivasan and J. Morgenstern. 2021. Auctions and prediction markets for scientific peer review. *arXiv preprint arXiv:2109.00923*.

[229] A. Ugarov. 2021. Peer prediction for peer review: designing a marketplace for ideas.

[230] S. Van Rooyen, F. Godlee, S. Evans, N. Black, and R. Smith. 1999. Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *Bmj*.

[231] F. Godlee, C. R. Gale, and C. N. Martyn. 1998. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial. *Jama*.

[232] S. Van Rooyen, T. Delamothe, and S. J. Evans. 2010. Effect on peer review of telling reviewers that their signed reviews might be posted on the web: randomised controlled trial. *Bmj*.

[233] E. Walsh, M. Rooney, L. Appleby, and G. Wilkinson. 2000. Open peer review: a randomised controlled trial. *The British Journal of Psychiatry*.

[234] A. Feldmann. 2005. Experiences from the SIGCOMM 2005 european shadow pc experiment. *ACM SIGCOMM Computer Communication Review*.

[235] B Parno, U Erlingsson, and W Enck. 2017. 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).

[236] I. Stelmakh, N. Shah, A. Singh, and H. Daumé III. 2021. A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In *AAAI*.

[237] M. Callaham and C. McCulloch. 2011. Longitudinal trends in the performance of scientific peer reviewers. *Annals of emergency medicine*.

[238] D. Joyner and A. Duncan. 2020. 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).

[239] S. Schroter, N. Black, S. Evans, J. Carpenter, F. Godlee, and R. Smith. 2004. Effects of training on quality of peer review: randomised controlled trial. *Bmj*.

[240] M. L. Callaham and J. Tercier. 2007. The relationship of previous training and experience of journal peer reviewers to subsequent review quality. *PLoS medicine*.

[241] W. G. Baxt, J. F. Waeckerle, J. A. Berlin, and M. L. Callaham. 1998. Who reviews the reviewers? feasibility of using a fictitious manuscript to evaluate peer reviewer performance. *Annals of emergency medicine*.

[242] A. Chauvin, P. Ravaud, G. Baron, C. Barnes, and I. Boutron. 2015. The most important tasks for peer reviewers evaluating a randomized controlled trial are not congruent with the tasks most often requested by journal editors. *BMC medicine*.

[243] J. L. Cornelius. 2012. Reviewing the review process: identifying sources of delay. *The Australasian medical journal*.

[244] B.-C. Björk and D. Solomon. 2013. The publishing delay in scholarly peer-reviewed journals. *Journal of informetrics*.

[245] M.-Y. Kan. 2017. Author response: does it help? ACL 2017 PC Chairs Blog https://acl2017.wordpress.com/2017/03/27/author-response-does-it-help/. (2017).

[246] Y. Gao, S. Eger, I. Kuznetsov, I. Gurevych, and Y. Miyao. 2019. Does my rebuttal matter? insights from a major nlp conference. In *Proceedings of NAACL-HLT*.

[247] M. Teplitskiy, H. Ranub, G. S. Grayb, M. Meniettid, E. C. Guinan, and K. R. Lakhani. 2019. Social influence among experts: field experimental evidence from peer review.

[248] NIH center for scientific review. 2020. 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).

[249] M. Fogelholm, S. Leppinen, A. Auvinen, J. Raitanen, A. Nuutinen, and K. Väänänen. 2012. Panel discussion does not improve reliability of peer review for medical research grant proposals. *Journal of clinical epidemiology*.

[250] E. Pier, J. Raclaw, A. Kaatz, M. Brauer, M. Carnes, M. Nathan, and C. Ford. 2017. Your comments are meaner than your score: score calibration talk influences intra-and inter-panel variability during scientific grant peer review. *Research Evaluation*.

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

[252] A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan. 2021. The neurips 2021 consistency experiment. https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment/. (2021).

[253] T. P. Hofer, S. J. Bernstein, S. DeMonner, and R. A. Hayward. 2000. Discussion between reviewers does not improve reliability of peer review of hospital quality. *Medical care*.

[254] I. Stelmakh, C. Rastogi, N. B. Shah, A. Singh, and H. Daumé III. 2020. A large scale randomized controlled trial on herding in peer-review discussions. *arXiv preprint arXiv:2011.15083*.

[255] A. Cockburn, P. Dragicevic, L. Besançon, and C. Gutwin. 2020. Threats of a replication crisis in empirical computer science. *Communications of the ACM*.

[256] M. Baker. 2016. Reproducibility crisis. *Nature*.

[257] O. Gencoglu, M. van Gils, E. Guldogan, C. Morikawa, M. Süzen, M. Gruber, J. Leinonen, and H. Huttunen. 2019. Hark side of deep learning–from grad student descent to automated machine learning. *arXiv preprint arXiv:1904.07633*.

[258] E. G. Campbell, B. R. Clarridge, M. Gokhale, L. Birenbaum, S. Hilgartner, N. A. Holtzman, and D. Blumenthal. 2002. Data withholding in academic genetics: evidence from a national survey. *jama*.

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

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

[261] A. Stent. 2017. Our new review form. NAACL 2018 Chairs blog https://naacl2018.wordpress.com/2017/12/14/our-new-review-form. (2017).

[262] J. C. Mogul. 2013. Towards more constructive reviewing of sigcomm papers. (2013).

[263] PLDI. 2015. PLDI 2015 author and reviewer surveys. https://conf.researchr.org/track/pldi2015/pldi2015-papers#Surveys. (2015).

[264] K. Leyton-Brown, Mausam, Y. Nandwani, H. Zarkoob, C. Cameron, N. Newman, D. Raghu, et al. 2022. Matching papers and reviewers at large conferences. *arXiv preprint arXiv:2202.12273*.

[265] B. Barak. 2016. Computer science should stay young. *Communications of the ACM*.

[266] T. K. Mackey, N. Shah, K. Miyachi, J. Short, and K. Clauson. 2019. A framework proposal for blockchain-based scientific publishing using shared governance. *Frontiers in Blockchain*.

[267] S. H. Emile, H. K. Hamid, S. D. Atici, D. N. Kosker, M. V. Papa, H. Elfeki, C. Y. Tan, A. El-Hussuna, and S. D. Wexner. 2022. Types, limitations, and possible alternatives of peer review based on the literature and surgeons' opinions via twitter: a narrative.

[268] L. Wasserman. 2012. A world without referees. *ISBA Bulletin*.

[269] J.-B. Huang. 2018. Deep paper gestalt. *arXiv preprint arXiv:1812.08775*.

[270] Q. Wang, Q. Zeng, L. Huang, K. Knight, H. Ji, and N. F. Rajani. 2020. Reviewrobot: explainable paper review generation based on knowledge synthesis. *arXiv preprint arXiv:2010.06119*.

[271] T. Houle. 2016. An introduction to StatReviewer. EMUG, available online https://www.ariessys.com/wp-content/uploads/EMUG2016_PPT_StatReviewer.pdf. (2016).

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

[273] R. Van Noorden et al. 2022. Journals adopt ai to spot duplicated images in manuscripts. *Nature*.

[274] J. Langford and M. Guzdial. 2015. The arbitrariness of reviews, and advice for school administrators. *Communications of the ACM*.

[275] T. Jefferson, E. Wager, and F. Davidoff. 2002. Measuring the quality of editorial peer review. *Jama*.

[276] D. W. Aksnes and A. Rip. 2009. Researchers' perceptions of citations. *Research Policy*.

[277] DORA. 2012. San Francisco declaration on research assessment. https://sfdora.org/read/. (2012).

[278] D. W. Aksnes, L. Langfeldt, and P. Wouters. 2019. Citations, citation indicators, and research quality: an overview of basic concepts and theories. *Sage Open*.

[279] D. S. Chawla. 2020. 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).

[280] R. Rezapour, J. Bopp, N. Fiedler, D. Steffen, A. Witt, and J. Diesner. 2020. Beyond citations: corpus-based methods for detecting the impact of research outcomes on society. In *Proceedings of The 12th Language Resources and Evaluation Conference*.

[281] W. Ding, N. B. Shah, and W. Wang. 2020. On the privacy-utility tradeoff in peer-review data analysis. In *AAAI Privacy-Preserving Artificial Intelligence (PPAI-21) workshop*.