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

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

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

Contents
1 Introduction 2
2 An overview of the review process 4
3 Mismatched reviewer expertise 4
   3.1 Computing similarity scores ............................................. 5
      3.1.1 Subject-area selection ............................................. 5
      3.1.2 Text matching ..................................................... 5
      3.1.3 Bidding ............................................................ 5
      3.1.4 Combining data sources ......................................... 7
   3.2 Computing the assignment ............................................. 7
4 Dishonest behavior 8
   4.1 Lone wolf ............................................................. 8
   4.2 Coalitions ........................................................... 9
   4.3 Other issues ........................................................ 12
5 Miscalibration 13
6 Subjectivity 15
   6.1 Commensuration bias ............................................... 15
   6.2 Confirmation bias and homophily .................................. 16
   6.3 Acceptance via obfuscation (“Dr. Fox effect”) .................... 16
   6.4 Surprisingness and hindsight bias .................................. 17
   6.5 Hindering novelty .................................................... 17
   6.6 Positive-outcome bias .............................................. 17
   6.7 Interdisciplinary research .......................................... 18

*This is an extended version of an article in the Communications of the ACM [Sha22a]. This is an evolving draft. The first version appeared on July 8, 2021. The current version is dated March 14, 2024.
## 1 Introduction

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

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

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

Problems in peer review have consequences much beyond the outcome for a specific paper or grant proposal, particularly due to the widespread prevalence of the Matthew effect (“rich get richer”) in academia [Mer68; TC14; SG12]. As noted in [TT07] “an incompetent review may lead to the rejection of the submitted paper, or of the grant application, and the ultimate failure of the career of the author.” This raises the important question [Lee15]: “In public, scientists and scientific institutions celebrate truth and innovation. In private, they perpetuate peer review biases that thwart these goals... what can be done about it?” Additionally, the large number of submissions in fields such as machine learning and artificial intelligence (Figure 1) has put a considerable strain on the peer-review process. The increase in the number of submissions is also large in many other fields: “Submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint” [McC06]. In medical research, publication of flawed research can cause direct harm to patients [Zar+13; Pou+22]. The annual financial value of the time that researchers dedicate to peer reviewing is estimated to be over two billion dollars [ASH21].

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

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

---

1For those unfamiliar 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.
academia understand what goes on behind the scenes in the peer-review process and the challenges that lie therein.

2 An overview of the review process

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

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

Each reviewer is required to provide reviews for their assigned papers before a pre-specified deadline. The reviews comprise an evaluation of the paper and suggestions to improve the paper. The authors may then provide a rebuttal to the review, which could clarify any inaccuracies or misunderstandings in the reviews. Reviewers are asked to read the authors’ rebuttal (as well as other reviews) and update their reviews accordingly. A discussion for each paper then takes place between its reviewers and meta reviewer. Based on all of this information, the meta reviewer then recommends to the program chairs a decision about whether or not to accept the paper to the conference. The program chairs eventually make the decisions on all papers.

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

3 Mismatched reviewer expertise

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

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

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

### 3.1 Computing similarity scores

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

#### 3.1.1 Subject-area selection

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

#### 3.1.2 Text matching

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

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

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

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

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

#### 3.1.3 Bidding

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

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

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

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

These issues motivate an algorithm [FSR20] that dynamically orders papers for every reviewer by trading off reviewer satisfaction (showing papers with higher similarity at the top, using metrics like the discounted cumulative gain or DCG) with balancing paper bids (showing papers with fewer bids at the top). A second paper [Mei+21] also looks to address the problem of imbalanced bids across papers, but via a different approach. Specifically, it proposes a market-style bidding scheme where it is more “expensive” for reviewer to bid on a paper which has already received many bids. These approaches are empirically evaluated in [Roz+23], which finds that both these approaches [FSR20; Mei+21] help in balancing the bids.
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 $s_{p,r} = \frac{2 bid-score_{p,r} - (subject-score_{p,r} + text-score_{p,r})}{4}$ used in the NeurIPS 2016 conference [Sha+18], or $(\frac{1}{2} subject-score_{p,r} + \frac{1}{2} text-score_{p,r})$ used in the AAAI 2021 conference [LBM21], or a convex combination of the data sources as done in the OpenReview.net platform. A second approach involves program chairs trying out various combinations, eyeballing the resulting assignments, and picking the combination that seems to work best. Finally and importantly, if any reviewer $r$ has a conflict with an author of any paper $p$ (that is, if the reviewer is an author of the paper or is a colleague or collaborator of any author of the paper), then the similarity $s_{p,r}$ is set as $-1$ to ensure that this reviewer is never assigned this paper.

### 3.2 Computing the assignment

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

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

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

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

![Table showing similarity scores for different papers and reviewers](image.png)
This assignment is an example of an alternative approach [Gar+10; SSS21b; KSM19; Lia+18] that optimizes for the paper which is worst-off in terms of the similarities of its assigned reviewers:

\[
\text{maximize } \sum_{\text{papers } p} \sum_{\text{reviewers } r} s_{p,r},
\]

minimize \( \text{assignment} \) \( \text{papers } p \) \( \text{assigned to paper } p \)

The approach then optimizes for the paper that is the next worst-off and so on. Evaluations [KSM19; SSS21b] of this approach on several conferences reveal that it significantly mitigates the problem of imbalanced assignments, with only a moderate reduction in the sum-similarity score value as compared to the approach of maximizing sum-similarity scores. Furthermore, the assignment algorithm [SSS21b] is found to also have desirable properties such as low “envy”, high “Nash social welfare”, and a high similarity on the bottom 10% and the bottom 25% papers [PZ21]. This approach is now adopted in conferences such as the International Conference on Machine Learning (ICML) 2020 [SSS21b].

Recent work also incorporates various other desiderata in the reviewer-paper assignments such as geographic diversity [LBM21], envy-freeness [PZ21], and addressing uncertainty in the similarities [CPZ23]. See the paper [JR22] for a survey of researchers on the importance they place on various desiderata in the assignments. An emerging concern when doing the assignment is that of dishonest behavior, as we discuss next.

4 Dishonest behavior

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

4.1 Lone wolf

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

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

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

This motivates the requirement of “strategyproofness”: no reviewer must be able to influence the outcome of their own submitted paper by manipulating the reviews they provide. A simple yet effective idea to ensure strategyproofness is called the partition-based method introduced in [Alo+11] and studied subsequently in many papers [HM13; BNV14; FK15; Kur+15; Kah+18; Xu+19; Azi+19; Dhu+22] (see [OW22] for an extensive survey on strategyproofing). The key idea of the partition-based method is illustrated in Figure 5. Consider the “authorship” graph in Figure 5a whose vertices comprise the submitted papers and reviewers, and an edge exists between a paper and reviewer if the reviewer is an author of that paper. The partition-based method first partitions the reviewers and papers into two (or more) groups such that all authors of any paper are in the same group as the paper (Figure 5b). Each paper is then assigned for review to reviewers in the other group(s) (Figure 5c). Finally, the decisions for the papers in any group are made independent of the other group(s) (Figure 5d). This method is strategyproof since any reviewer’s reviews influence only papers in other groups, whereas the reviewer’s own authored papers belong to the same group as the reviewer.
The partition-based method is largely studied in the context of peer-grading-like settings. In peer grading, one may assume each paper (homework) is authored by one reviewer (student) and each reviewer authors one paper, as is the case in Figure 5. Conference peer review is more complex: papers have multiple authors and authors submit multiple papers. Consequently, in conference peer review it is not clear if there even exists a partition. Secondly, peer grading is more homogeneous where any paper can be assigned to any reviewer, whereas papers and reviewers in peer review are much more specialized (Section 3). Hence, even if such a partition exists, the partition-based constraint on the assignment could lead to a considerable reduction in the assignment quality. Such questions about realizing the partition-based method in conference peer review are still open, with promising initial results [Xu+19; Dhu+22] showing that such partitions do exist in practice and the reduction in quality of assignment may not be too drastic.

4.2 Coalitions

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

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

A second line of defense addresses attacks where two or more reviewers (who have also submitted their own papers) aim to review each other’s papers. This has motivated the design of assignment algorithms [Guo+18; BBN21] with an additional constraint of disallowing any loops in the assignment, that is, ensuring to not assign two people each others’ papers. Such a condition of forbidding loops of size two was also used in the reviewer assignment for the Association for the Advancement of Artificial Intelligence (AAAI) 2021 conference [LBM21]. This defence prevents colluders engaging in a quid pro quo in the same venue. However,\(^2\)

\(^2\)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.
Figure 6: An vulnerability in the assignment system, exposed in [Mar+17; TJ19], via attacks that exploit font embedding in the PDF of the submitted paper. Suppose the colluding reviewer has the word “minion” as most frequently occurring in their previous papers, whereas the paper submitted by the colluding author has “review” as most commonly occurring. The author creates two new fonts that map the plain text to rendered text as shown. The author then chooses fonts for each letter in the submitted paper in such a manner that the word “minion” in plain text renders as “review” in the PDF. A human reader will now see “review” but an automated parser will read “minion”. The submitted paper will then be assigned to the target reviewer by the assignment system, whereas no human reader will see “minion” in the submitted paper.

this defense can be circumvented by colluders who avoid forming a loop, for example, where a reviewer helps an author in a certain conference and the author reciprocates elsewhere. Moreover, it has been uncovered that, in some cases, an author pressures a certain reviewer to get assigned and accept a paper [Lau20]. This line of defense does not guard against such situations where there is no quid pro quo within the conference.

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

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

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

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

The first line of attacks on automated text matching involves an author manipulating the PDF (portable document format) of their submitted paper so that a certain reviewer gets assigned [Mar+17; TJ19]. These attacks insert text in the PDF of the submitted paper in a manner that satisfies three properties: (1) the inserted text matches keywords from a target reviewers’ paper; (2) this text is not visible to the human reader; and (3) this text is read by the (automated) parser which computes the text-similarity-score between
the submitted paper and the reviewer’s past papers. These three properties guarantee a high similarity for the colluding reviewer-paper pair, while ensuring that no human reader detects it. These attacks are accomplished by targeting the font embedding in the PDF, as illustrated in Figure 6. Empirical evaluations on the reviewer-assignment system used at the International Conference on Computer Communications (INFOCOM) demonstrate the high efficacy of these attacks by being able to get papers matched to target reviewers.

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

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

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

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

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

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

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

This assignment procedure also mitigates potential “torpedo reviewing” [Lan12a] where a reviewer intentionally tries to get assigned a paper to reject it, possibly because it is a competing paper or if it is from an area the reviewer does not like. Also interestingly, in the SIGCOMM 2006 conference, the assignments were done randomly among the reviewers who were qualified in the paper topic area to “improve the confidence intervals” [And09] of the evaluation of any paper.
Figure 7: Trading off the quality of the assignment (sum similarity on y-axis) with the amount of randomness (value specified by program chairs on x-axis) to mitigate dishonest coalitions [Jec+20]. The similarity scores for the “ICLR” plot are reconstructed [Xu+19] via text-matching from the International Conference on Learning Representations (ICLR conference) 2018 which had 911 submissions. The “Preflib” plots are computed on bidding data from three small-sized conferences (with 54, 52 and 176 submissions), obtained from the Preflib database [MW13].

4.3 Other issues

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

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

- **Selling authorship** [Hvi13]
- **Bribing journal editors** [Joe24]
- **Faking reviewer identities** [FMO14; Coh+16]
- **Plagiarism of papers** [MADV05; HE15]
- **Plagiarism of reviews** [Pin+24]
- **Data fabrication** [Woo16; Fan09; AM+05; TVS14]
- **Fake paper mills** [EVN+21]
- **Multiple submissions** [Bow99]
- **Stealing confidential information from grant proposals submitted for review** [TW18; Mur20]
- **Breach of confidentiality** [RGFP08; Ras+24]
- **Other issues** [MADV05; ER17; FW17].
### Table 1: Distribution of review ratings in NeurIPS 2016 [Sha+18]. The column headings contain the guidelines provided to reviewers.

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

5 Miscalibration

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

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

Miscalibration may also occur if there is a mismatch between the conference’s overall expectations and reviewers’ individual expectations. As a concrete example, the NeurIPS 2016 conference asked reviewers to rate papers according to four criteria on a scale of 1 through 5 (where 5 is best), and specified an expectation regarding each value on the scale. However, as shown in Table 1, there was a significant difference between the expectations and the ratings given by reviewers [Sha+18]. For instance, the program chairs asked reviewers to give a rating of 3 or better if the reviewer considered the paper to lie in the top 30% of all submissions, but the actual number of reviews with the rating 3 or better was nearly 60%. Eventually the conference accepted approximately 22% of the submitted papers.

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

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

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

$$a_r \theta_p + b_r + \text{noise}.$$
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 [BGK05; GB08]; see Figure 8 for an illustration. Algorithms based on such assumptions were tried in some conferences, but based on manual inspection by the program chairs, were found to perform poorly. For instance: “We experimented with reviewer normalization and generally found it significantly harmful” in ICML 2012 [Lan12b].

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

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

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

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

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

---

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

\(^5\)Interestingly, randomized decisions are used in practice by certain funding agencies to allocate grants [Liu+20; Cha21; Hey+22]. Such randomized decision-making has found support among researchers [Phi21] as long as it is combined with the peer review process and is not pure randomness. Identified benefits of such randomization include overcoming ambiguous decisions for similarly-qualified proposals, decreasing reviewer effort, circumventing old-boys’ networks, and increasing chances for unconventional research [Phi21].
Figure 9: Mapping of individual criteria to overall ratings by reviewers in IJCAI 2017 [NSP21]. The conference used five criteria, and hence the mapping is five-dimensional. The figure plots representative two-dimensional cross sections of the mapping of the following pairs of criteria to overall ratings: (a) writing and relevance, (b) significance and technical quality, and (c) originality and relevance.

that allow conferences to share some reviewer-specific calibration data with one another in order to calibrate better.

6 Subjectivity

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

6.1 Commensuration bias

Program chairs of conferences often provide criteria to reviewers for judging papers. However, different reviewers have different, subjective opinions about the relative importance of various criteria in judging papers [Mah77; Chu05; Lam09; BMS87; HO21]. The overall evaluation of a paper then depends on the individual reviewer’s preference on how to aggregate the evaluations on the individual criteria. This dependence on factors exogenous to the paper’s content results in arbitrariness in the review process. On the other hand, in order to ensure fairness, all (comparable) papers should be judged by the same yardstick. This issue is known as “commensuration bias” [Lee15].

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

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

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

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

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

6.2 Confirmation bias and homophily

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

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

6.3 Acceptance via obfuscation (“Dr. Fox effect”)

A controlled study [Arm80] asked evaluators to each rate one passage, where the readability of these passages was varied across reviewers but the content remained the same. The study found that the passages which

---

6 According to Mahoney [PC82], for this study, “the emotional intensity and resistance of several participants were expressed in the form of charges of ethical misconduct and attempts to have me fired. Several editors later informed me that correspondence from my office was given special scrutiny for some time thereafter to ascertain whether I was secretly studying certain parameters of their operation.”
were harder to read were rated higher in research competence. The author cheekily concludes with the phrase “If you can’t convince them, confuse them.”

6.4 Surprisingness and hindsight bias

One criteria that reviewers often use in their judgment of a paper is the paper’s informativeness or surprisingness. Anecdotally, it is not uncommon to see reviews criticizing a paper as “the results are not surprising.” But are the results as unsurprising as the reviewers claim them to be? Slovic and Fischhoff [SF77] conducted a controlled study to investigate reviewers’ perceptions of surprisingness. They divided the participants in the study randomly into two groups: a “foresight” group and a “hindsight” group. Each participant in the foresight group was shown a fictitious manuscript which contained the description of an experiment but not the results. There results could take two possible values. Each participant in the hindsight group were shown the manuscript containing the description as well as the result. The result of the manuscript shown to any participant was chosen randomly as one of the two possible values. The foresight participants were then asked to assess how surprising each of the two possible results would seem if they obtained, whereas the hindsight 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 [Chu05]: “Reviewers love safe (boring) papers, ideally on a topic that has been discussed before (ad nauseam)... The process discourages growth”. Naughton makes a noteworthy point regarding one reason for this problem: “Today reviewing is like grading: When grading exams, zero credit goes for thinking of the question. When grading exams, zero credit goes for a novel approach to solution. (Good) reviewing: acknowledges that the question can be the major contribution. (Good) reviewing: acknowledges that a novel approach can be more important than the existence of the solution” [Nau10].

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

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

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

6.6 Positive-outcome bias

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

6.7 Interdisciplinary research

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

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

7 Biases pertaining to author identities

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

Such biases with respect to author identities are widely debated in computer science and elsewhere. These debates have led to two types of peer-review processes: single-blind reviewing where reviewers are
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.

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. We refer the reader to [Sha22b, Section 4] for a variety of arguments for and against anonymizing authors by researchers in theoretical computer science.

7.1 Studies in computer science

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

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

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

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

7.2 Studies outside computer science

These results augment a vast body of literature in various scientific fields outside of computer science investigating biases pertaining to author identities. The study [BMD07] finds gender bias, [WW01] finds biases with respect to gender and personal connections, the studies [Gin+11; Nak+21] find bias with respect to race, whereas the study [For+19] finds little evidence of gender or racial bias. Several studies [Oki+16; Bla91; Ros+06; Gar+94] find bias in favor of authors’ status. In particular, [Gar+94] observes a significant

7On the topic of analysis of review text, some recent works analyze arguments [Hua+19; Fro+20; Ken+21; Kuz+22; Gho+22] and sentiments [Gho+19; CGM20; Bul+20; Rib+21] 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.
bias for brief reports but not for major papers. This observation suggests that reviewers tend to use author characteristics more when less information about the research is available. The study [Nie+21] finds weak evidence of country and institution bias when scientists evaluate abstracts. Bias with respect to author fame is also investigated in the paper [Bla91], which finds that the top and bottom institutions’ papers unaffected, but those in the middle were affected. In a similar vein, the study [Fou+00] suggests that “evaluation of absolutely outstanding articles will not be biased, but articles of ambiguous merit may be judged based on the author’s gender.” A randomized controlled trial [FFS94] found that authors with more past papers were given better scores by blinded reviewers. The paper [Bud+08] finds an increased representation of women authors following a policy change from single to double blind. The study [GW99] finds that blinding reviewers to the author’s identity does not usefully improve the quality of reviews. The study [Hub+22] found a significant preference of reviewers towards a manuscript that had the name of a Nobel laureate as author than an identical manuscript with a junior researcher’s name. Surveys of researchers [War08; MHR13] reveal that double blind review is preferred and perceived as most effective.

7.3 Design of experiments

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

7.4 Use of author identities in non-anonymized review

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

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

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

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

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

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

8 Incentives

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

8.1 Author incentives

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

8.1.1 Open Review

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

Alongside these benefits, the open-review format may also result in some challenges such as threats to anonymity of reviewers (see Section 9.5 for more details). We discuss one other 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 [The13]. A large fraction of papers accepted at top conferences are previously rejected at least once [FK20b]. To reuse review effort, many conferences are adopting policies where authors of a paper must provide past rejection information along with the submission. For instance, the IJCAI 2020 conference required authors to prepend their submission with details of any previous rejections including prior reviews and the revisions made by authors. While these policies are well-intentioned towards ensuring that authors do not simply ignore reviewer feedback, the information of previous rejection could bias the reviewers.

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

8.1.3 Rolling deadlines

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

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

8.1.4 Authors’ perspectives

The paper [Su21] presents a novel idea of asking authors to provide a ranking their submitted papers, and using the authors’ ranking to “denoise” reviews. However, several challenges remain to make this interesting approach practical. For instance, the method cannot handle disagreements between co-authors [Ras+22a] about the rankings. It can incentivize authors to falsely report their ranking of their own papers, which can in turn lead to poorer quality papers being accepted. Furthermore, this method requires authors to predict the ranking in terms of what reviewers will think, which may add to the stress of submitting papers.

8.1.5 Citation-based incentives

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

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

8.2 Reviewer incentives

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

8.2.1 Materialistic and non-materialistic incentives

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

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

Many venues make it compulsory for qualified submitting authors to also review papers. We are not aware of any studies quantitatively comparing the quality of compulsory versus volunteer reviews of scientific papers. A related application is that of peer evaluation of lab reports in a classroom setting. The study [PBB23] conducted an experiment asking students to provide such evaluations on a compulsory basis in one year and...
Squazzoni et al. [SBT13] empirically evaluate the effects of various incentive mechanisms via “investment game” that mimics various characteristics of incentives in peer review. Within this game, they conduct a controlled trial that compares a setting with no payoffs for reviewers, an incentive comprising a fixed payoff for reviewers, and two incentive structures involving a variable payoff for reviewers. They quantitatively find that the no-payoff setting results in the most effective peer review. Surveys of participants point to the trust and cooperation in the no-payoff setting as the key to more effective peer review in this setting in the experiment.

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

### 8.2.2 Reviewing the reviews

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

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

### 8.2.3 Game-theoretic approaches

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

### 8.2.4 Signed reviews

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

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

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

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

9 Norms and policies

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

9.1 Peer-review objectives

In this section, we discuss research on the three key objectives of peer review (discussed in Section 1). We summarize them first, and detail them subsequently. (1) **Ensuring correctness:** Reviewers do reject flawed papers a large fraction of the time, although they find only a subset of flaws. (2) **Highlighting the “best” research:** There is at best weak evidence that peer review can separate the “top-tier” research from the next tier. (3) **Providing constructive feedback to authors:** A majority of authors across several surveys report that they find the reviews helpful, however, authors’ opinions are also biased by whether the reviews recommended acceptance. Overall, we think there should be more emphasis on ensuring correctness and providing good feedback, while de-emphasizing “highly selective” nature of venues and associated focus on counting papers published therein.

9.1.1 Ensuring correctness

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

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

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

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

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

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

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

These results suggest that while peer review does filter out bad science to some extent, more emphasis needs to be placed on evaluating correctness. It is possible that the frequently used evaluation methods that focus on subjective aspects (e.g., perceived novelty or impact) may distract from the goal of ensuring correctness of published work. To this end, journals such as PLOS ONE, JSys, and the Transactions on Machine Learning Research (TMLR) focus on rigor as the sole criterion for acceptance.

9.1.2 Selecting the ‘best’ research

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

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

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

Further, in Section 9.4 we discuss several experiments which have independent sets of reviewers reviewing the same set of submissions. These experiments find a significant amount of disagreement in the subset of papers accepted by the two independent sets of reviewers; some of the experiments in fact find a zero correlation between reviewers across the two sets.

An experiment at the NeurIPS2021 conference [Ras+22a] asked authors who submitted multiple papers to rank their papers in terms of their own perceived scientific contributions of these papers. Interestingly, they found that the amount of disagreement between co-authors about their jointly authored papers was as high as the disagreement with reviewers. Specifically, co-authors disagreed on the relative ranking of a pair of jointly authored papers roughly 1/3rd of the time. This result suggests existence of an intrinsic level of noise in identifying the ‘best’ research, that may be fundamentally hard to overcome.8

Finally, we discuss studies analyzing the correlation between the reviews and the future impact of the papers. These studies measure impact in terms of the number of citations received by the papers. The paper [Rag+13] finds that for the accepted papers, the reviewer scores are nearly uncorrelated with the number of citations received subsequently. Likewise, the study [Eys22] finds that reviewers’ ratings of perceived impact are uncorrelated with citations, but sometimes correlated better with altmetrics (such as social media impressions). Similarly, the study [CMF14] finds that reviewer ratings are uncorrelated with citations or downloads. An analysis [CL21] of the NeurIPS 2014 conference finds no significant correlation: “If we accept that final paper citation counts are some measure of paper quality, then we see that reviewers fail to capture this in their scores.” The study [CL21] also analyses the rejected papers and finds a weak

---

8The experiment [Ras+22a] also asked authors to predict the probability of acceptance of their papers. Authors significantly over-predicted: for an acceptance rate of 21% to 25%, the mean author prediction was 67%.
correlation between the reviewer scores and future citations. The result on rejected papers may however be taken with a pinch of salt, as authors of rejected papers may have taken different paths for these papers depending on the reviews. Finally, Schroter et al. [Sch+22] explicitly task evaluators to forecast future citations, and find that the evaluators fail to make such predictions accurately.

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

9.1.3 Constructive feedback to authors

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

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

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

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

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

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

9.2 Review quality

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

9.2.1 Reviewer training.

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

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

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

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

### 9.2.2 Following peer-review guidelines

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

### 9.2.3 Review timeliness

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

### 9.3 Author rebuttal

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

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

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

The paper [Gao+19] designs a model to predict post-rebuttal scores based on initial reviews and the authors’ rebuttals. They find that the rebuttal has a marginal (but statistically significant) influence on the
final scores, particularly for borderline papers. They also find that the final score given by a reviewer is largely dependent on their initial score and the scores given by other reviewers for that paper.\footnote{The paper \cite{Gao+19} concludes “Peer pressure” to be “the most important factor of score change”. This claim should be interpreted with caution as there is no evidence presented for this causal claim. The reader may instead refer to the controlled experiment \cite{Lan+22} on this topic, discussed in Section 9.4.}

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

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

In order to improve the rebuttal process, a suggestion was made long ago by Porter and Rossini \cite{PR85} 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 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 \cite{DI15} suggests absence of such a relation. This brings us to the topic of discussions among reviewers.

### 9.4 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 \cite{NIH20} between these modes. We discuss a few studies on this topic.

#### 9.4.1 Do panel discussions improve consistency?

Several studies \cite{OTD07; Fog+12; Pie+17} conduct controlled experiments in the peer review of grant proposals to quantify the reliability of the process. The peer-review process studied here involves discussions among reviewers in panels. In each panel, reviewers first submit independent reviews, following which the panel engages in a discussion about the proposal, and reviewers can update their opinions. These studies reveal the following three findings. First, reviewers have quite a high level of disagreement with each other in their independent reviews. Second, the inter-reviewer disagreement within a panel decreases considerably after the discussions (possibly due to implicit or explicit pressure on reviewers to arrive at a consensus). This observation seems to suggest that discussions actually improve the quality of peer review. After all, it appears that the wisdom of all reviewers is being aggregated to make a more “accurate” decision. To quantify this aspect, these studies form multiple panels to evaluate each proposal, where each panel independently conducts the entire review process including the discussion. The studies then measure the amount of disagreement in the outcomes of the different panels for the same proposal. Their third finding is that, surprisingly, the level of disagreement across panels does not decrease after discussions, and instead often increases. Please see Figure 13 for more details.
The paper [Hof+00] performed a similar study in the peer review of hospital quality, and reached similar conclusions: “discussion between reviewers does not improve reliability of peer review.”

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

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

9.4.2 Influence of other reviewers

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

9.4.3 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-
A scale controlled experiment conducted at the ICML 2020 conference studied the existence of a “herding” effect [Ste+20]. The study investigated the question: Does the final decision of the paper depend on the order in which reviewers join the discussion? They partitioned the papers at random into two groups. In one group, the most positive reviewer was asked to start the discussion, then later the most negative reviewer was asked to contribute to the discussion. In the second group, the most negative reviewer was asked to start the discussion, then later the most positive reviewer was asked to contribute. The study found no difference in the outcomes of the papers in the two groups. The absence of a “herding” effect in peer review discussions thus suggests that from this perspective, the current absence of any policy for choosing the discussion initiator does not hurt.

9.4.4 Anonymous versus non-anonymous discussions

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

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

9.4.5 A survey of reviewers.

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

9.5 Ensuring reviewer anonymity

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

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

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

9.6 Automated reviewing

There have been several uses of AI in reviewing of the paper itself, such as ensuring papers adhere to appropriate submission and reporting guidelines, statistical rigor, plagiarism checking, and to mitigate fraud by finding duplicated images or fake papers [NP20; FMG19; VN+22; CLM22; NW23].

There have been some attempts, albeit largely unsuccessful, to develop AI algorithms that can autonomously review academic papers in their entirety. The paper [Hua18] develops a system to make accept/reject papers based on the visual appearance of the paper. The authors did not have rejected papers, and so they trained a binary classifier to classify between papers accepted to a conference and those presented at a workshop. The paper [Wan+20] compares knowledge elements in various sections of the submitted paper with those in other sections as well as other papers in order to predict novelty, soundness, and other aspects of the paper. Neither of these works present any evidence that their systems can actually evaluate the soundness of the paper, which is a foremost requirement for publication. Furthermore, such systems may latch on to spurious features (e.g., low-quality paper with visual appeal or novel phrases) and there is no investigation of the robustness of these systems to this. The paper [Che+21] uses previously submitted papers and its decisions to train a supervised model. The paper [YLN21] evaluates the state-of-the-art natural language processing tools available as of 2021, to write reviews. They find while these tools were adept at summarizing papers and discussing diverse aspects of the work, the outputs struggled to maintain factual accuracy.

A study [LS23] explores the use of large language models (LLMs), as of May 2023, for the task of reviewing scientific papers. They find that directly asking the LLM to review the paper is not useful, however, assigning the LLM specific tasks is quite promising. An initial comparison conducted within the study revealed that GPT-4 significantly outperforms Google Bard and several open-source models. Then they found that when tasked with identifying errors in scientific papers, GPT-4 exhibited commendable performance. The researchers compiled 13 papers (plus a pilot paper), each intentionally containing an error. These errors ranged from conceptual to mathematical mistakes. GPT-4 successfully detected the mistakes in seven out of the 13 papers (and also in the pilot paper). Additionally, when GPT-4 was tasked to verify author checklists for papers from the NeurIPS conference, it achieved an impressive success rate of 86.6%.

The study [LS23] also found that, on the downside, GPT-4 struggled when asked to select the superior paper from a pair. The researchers designed ten pairs of abstracts, each pair containing one abstract with superior results. Additionally, some pairs were manipulated to contain elements such as overblown language or prompt injection attacks, to test the LLM’s robustness. Unfortunately, GPT-4 did not perform well, failing to accomplish the task correctly in six out of the ten instances. Therefore, while the LLM shows promise in some aspects of scientific paper review, it still struggles with others.

Recent efforts to build “AI reviewers” rely on using large language models such as GPT-4 in various ways [Lia+23; Ope; D’A+24]. A key challenge therein is to evaluate the quality of the AI generated reviews, given the biases in review evaluation discussed previously in Section 8.2.2.

A “chimera” test: With the motivation of objectively evaluating AI-generated reviews (as in [LS23]), the author of the present survey performed a chimera test on recent AI reviewer systems [Lia+23; Ope; D’A+24]. Here, the author put together parts of three of his own papers – all three on different problems in theoretical statistics – to create a nonsensical paper. When reviewed by these AI systems, none of them pointed out this obvious major problem with it.

9.7 Other aspects

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

10 Conclusions

There are many challenges of biases and unfairness in peer review. Improving peer review is sometimes characterized as a “fundamentally difficult problem” [LG15]: “Every program chair who cares tries to tweak the reviewing process to be better, and there have been many smart program chairs that tried hard. Why isn’t it better? There are strong nonvisible constraints on the reviewers time and attention.” 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” [Rag+13; JWD02]. 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 [AR09; And09; DOR12; ALW19; Cha20; Rez+20; DLCRGTS14].

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

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 ACM 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


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


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


A. Checco et al. “AI-assisted peer review”. In: Humanities and Social Sciences Communications (2021).


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


Dhu+22

Dhu+22


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


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


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


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


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


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


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


M. Obrecht, K. Tibelius, and G. D’Aloisio. “Examining the value added by committee discussion in the review of applications for research awards”. In: Research Evaluation (2007).


E. Pier et al. “Your comments are meaner than your score: score calibration talk influences intra-and inter-panel variability during scientific grant peer review”. In: Research Evaluation (2017).


Y. M. Smulders. “A two-step manuscript submission process can reduce publication bias”. In: Journal of clinical epidemiology (2013).


I. Stelmakh et al. “A Novice-Reviewer Experiment to Address Scarcity of Qualified Reviewers in Large Conferences”. In: AAAI. 2021.


I. Stelmakh et al. “Cite-seeing and reviewing: A study on citation bias in peer review”. In: Plos one (2023).


