DOI:10.1145/3528086

Improving the peer review process in a scientific manner shows promise.

BY NIHAR B. SHAH

Challenges, Experiments, and Computational Solutions in Peer Review

PEER REVIEW IS a cornerstone of scientific research. Although quite ubiquitous today, peer review in its current form became popular only in the middle of the 20th century. Peer review looks to assess research in terms of its competence, significance, and originality.[6] It aims to ensure quality control to reduce misinformation and confusion[4] thereby upholding the integrity of science and the public trust in science.[49] It also helps in improving the quality of the published research.[17] In the presence of an overwhelming number of papers written, peer review also has another role:[40] "Readers seem to fear the firehose of the Internet: they want somebody to select, filter, and purify research material."

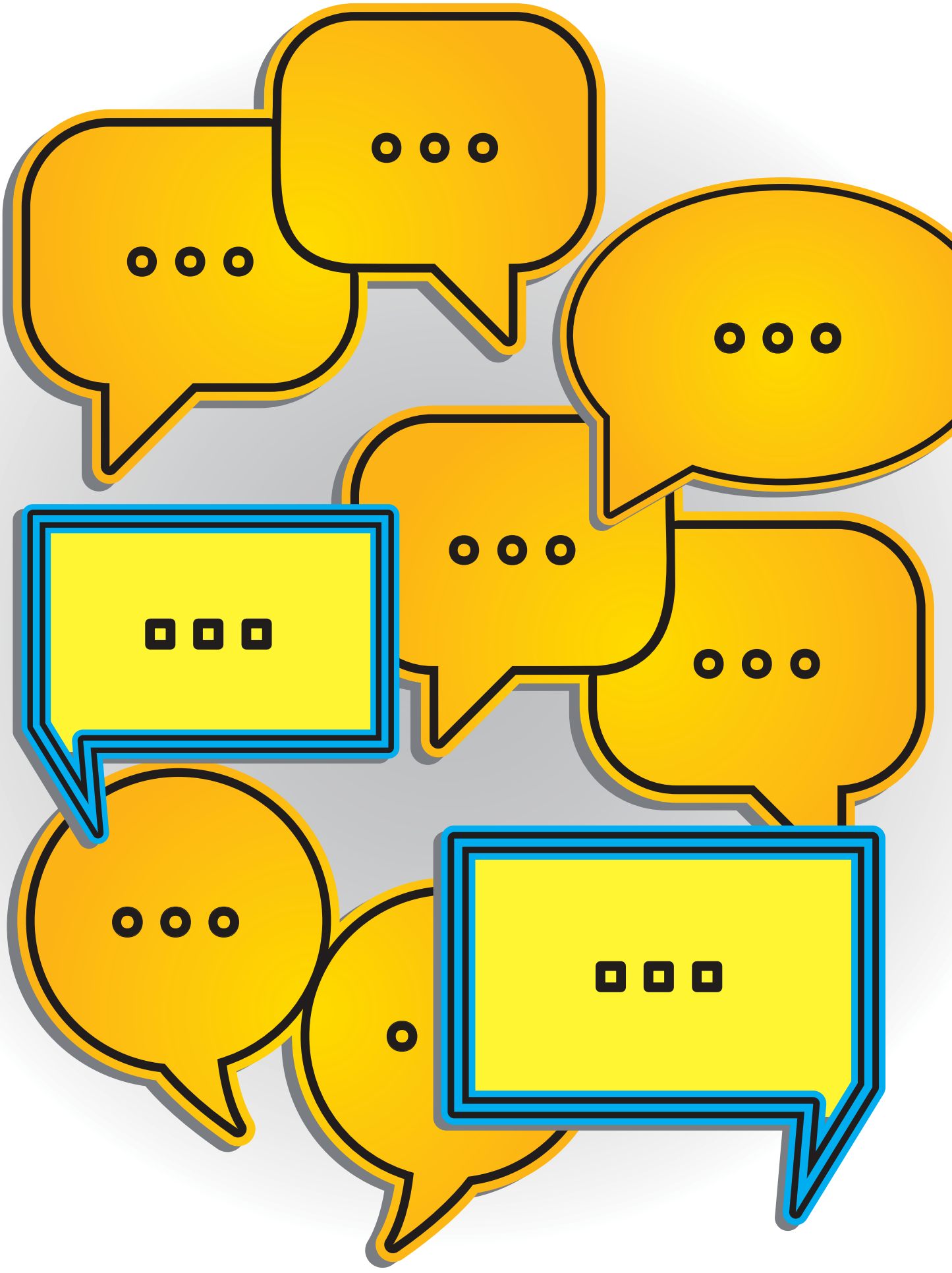
Surveys[48] of researchers in several scientific fields find that peer review is highly regarded by most researchers. Indeed, most researchers believe peer review gives confidence in the academic rigor of published articles and that it improves the quality of the published papers. These surveys also find 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."[17] For this, one must ensure evaluations are "independent of the author's and reviewer's social identities and independent of the reviewer's theoretical biases and tolerance for risk."[22] There are, however, key challenges toward these goals. The following quote from Rennie[35] 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.

In this article, we discuss several manifestations of the aforementioned challenges, experiments that help understand these issues and the trade-offs involved, and various (computational) solutions in the literature. For concrete-

» key insights

- **In computer science, the design of computational tools to assist peer review and experiments to understand various trade-offs form a fast-growing area of research. This line of research has already made significant impact with various computational tools widely deployed and experiments informing review policies.**
- **Further progress on these open problems can have substantial impact on peer review as well as on many other applications involving distributed human evaluations.**



ness, 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 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 outcomes of peer review can have a considerable influence on the career trajectories of authors. While we believe most participants in peer review are honest, the stakes can unfortunately incentivize dishonest behavior.

An Overview of the Review Process

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

Authors must submit their papers by a predetermined 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 of (typically three to six) reviewers. The number of papers per reviewer varies across conferences and can range from a handful (three to eight 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 set 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 this information, the meta reviewer then recommends to the program chairs a decision about whether 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. For example, many smaller-sized conferences do not have meta reviewers, and the final decisions are made via an in-person or online discussion between the entire pool of reviewers and program chairs. That said, most of the content to follow in this article is applicable broadly.

Mismatched Reviewer Expertise

The assignment of the reviewers to papers determines whether reviewers have the necessary expertise to review a paper. Time and again, a top reason for authors to be dissatisfied with reviews is the mismatch of the reviewers' expertise with the paper. For small conferences, the program chairs may assign reviewers themselves. However, this approach does not scale to conferences with hundreds or thousands of papers. 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). There are two stages in the automated assignment procedure.

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

Subject-area selection. When sub-

mitting 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.

Text matching. The text of the reviewer's previous papers is matched with the text of the submitted papers using natural language processing techniques. We summarize a couple of approaches here.[9,29] 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 reviewer's past papers.

The design of algorithms to compute similarities more accurately through advances in natural language processing is an active area of research.[32]

Bidding. Many conferences employ a "bidding" procedure where reviewers are shown the list of submitted papers and asked to indicate which papers they are willing or unwilling to review. A sample bidding interface is shown in Figure 1.

Cabanac and Preuss[7] analyze the bids made by reviewers in several conferences. Here, 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.

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.[36] The Neural Information Processing Systems (NeurIPS) 2016 conference in the field of machine learning aimed to assign six reviewers and one meta-reviewer to each of the 2,425 papers, but 278 papers received at most two positive bids and 816 papers received at most five positive bids from reviewers, and 1,019 papers received zero positive bids from meta reviewers.[38]

Cabanac and Preuss[7] 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[31] that humans are more likely to interact with an item if shown at the top of a list rather than down the list. Hence, this choice of interface results in a systematic bias against papers with greater values of assigned paper IDs.

Cabanac and Preuss suggest exploiting serial-position effects to ensure a better distribution of bids across papers by ordering the papers shown to any reviewer in increasing order of bids already received. However, this approach can lead to a high reviewer dissatisfaction since papers of the reviewer's interest and expertise may end up significantly down the list, whereas papers unrelated to the reviewer may show up at the top. An alternative ordering strategy used commonly in conference management systems today is to first compute a similarity between all reviewer-paper pairs using other data sources, and then order the papers in decreasing order of similarities with the reviewer. Although this approach addresses reviewer satisfaction, it does not exploit serial-position effects like the idea of Cabanac and Preuss. Moreover, papers with only moderate similarity with all reviewers (for example, if the paper is interdisciplinary) will not be shown at the top of the list to anyone. These issues motivate an algorithm[10] that dynamically orders papers for every reviewer by trading off reviewer satisfaction (showing papers with higher similarity at the top) with balancing paper bids (showing papers with fewer bids at the top).

Combining data sources. The data sources discussed above are then merged into a single similarity score. One approach is to use a specific formula for merging, such as $S_{p,r} = 2^{\text{bid-score}_{p,r}}(\text{subject-score}_{p,r} + \text{text-score}_{p,r})/4$ used in the NeurIPS 2016 conference.[38] 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 this reviewer is never assigned this paper.

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

Figure 1. A sample interface for bidding.

Papers:	Not Willing To Review	Indifferent	Eager To Review
Toward More Accurate NLP Models	○	○	○
Interpreting AI Decision Making	○	○	○
Multiagent Cooperative Board Games	○	○	○

the total sum of the similarity scores of all assigned reviewer-paper pairs:[9]

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

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

This approach of maximizing the sum of similarity scores can lead to unfairness to certain papers.[42] As a toy example illustrating this issue, consider a conference with three papers and six reviewers, where each paper is assigned one reviewer and each reviewer is assigned two papers. Suppose the similarities are given by the table on the left-hand side of Figure 2. 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 2. Observe that the assignment for paper C is quite poor: all assigned reviewers have a zero similarity. 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[18] analyzes data from the Computer Vision and Pattern Recognition (CVPR) 2017 and 2018 conferences, which have several thousand papers. The analysis reveals there is at least one paper each to which this method assigns all reviewers with a similarity score of zero, whereas other assignments can ensure that every paper has at least some reasonable reviewers.

The right-hand side of Figure 2 depicts the same similarity matrix. The cells shaded light/blue depict an alternative assignment. This assignment is more balanced: it assigns papers A and B reviewers of lower similarity as compared to earlier, but paper C now has reviewers with a total similarity of 1 rather than 0. This assignment is an example of an alternative approach[13,18,42] that optimizes for the paper which is worst-off in terms of the similarities of its assigned reviewers:

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

The approach then optimizes for the paper that is the next worst-off and so on. Evaluations[18,42] of this approach on several conferences reveal 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.

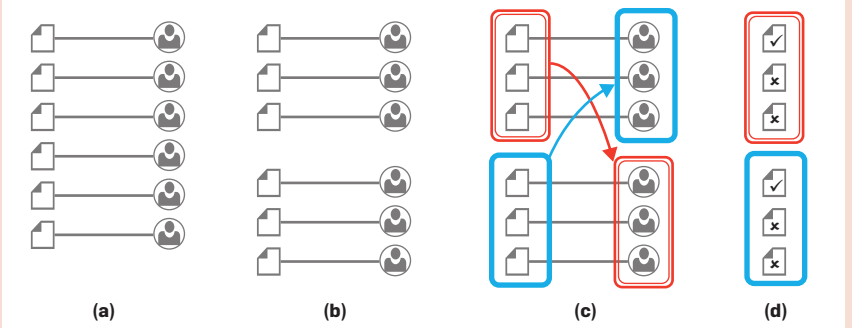
Recent work also incorporates various other desiderata in the reviewer-paper assignments.[23] An emerging concern when doing the assignment is that of dishonest behavior.

Figure 2. Assignment in a fictitious example conference using the popular sum-similarity optimization method (left) and a more balanced approach (right).

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

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

Figure 3. Partition-based method for strategyproofness.



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. We discuss two such issues.

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 (for example, providing lower ratings) for other papers.

A controlled study by Balietti et al.[3] 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.

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.[1] The key idea of the partition-based method is illustrated in Figure 3. Consider the "authorship" graph in Figure 3a 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 3b). Each paper is then assigned for review to reviewers in the other group(s) (Figure 3c). Finally, the decisions for the papers in any group are made independent of the other group(s) (Figure 3d). This method is strategy-proof 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 3. 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. 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[51] showing that such partitions do exist in practice and the reduction in quality of assignment may not be too drastic.

Coalitions. Several recent investigations have uncovered dishonest coalitions in peer review.[24,46] Here a reviewer and an author come to an understanding: the reviewer manipulates the system to try to be assigned the author's paper, then accepts the paper if assigned, and the author offers quid pro quo either in the same conference or elsewhere. There may be coalitions between more than two people, where a group of reviewers (who are also authors) illegitimately push for each other's papers.

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 coauthor-ship 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:[46] *"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)."*

Biases with respect to author identities are widely debated in computer science.

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[14] with an additional constraint of disallowing any loops in the assignment, that is, ensuring to not assign two people each other's' papers. This defense prevents colluders engaging in a quid pro quo in the same venue. However, this defense can be circumvented by colluders who avoid forming a loop, for example, where a reviewer helps an author in a certain conference and the author reciprocates elsewhere. Moreover, it has been uncovered that, in some cases, an author pressures a certain reviewer to get assigned and accept a paper.[19] 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.[16,50] This suggests curtailing or auditing bids, and this approach is followed in the paper.[50] This work uses the bids from all reviewers as labels to train a machine learning model that 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.

Dishonest collusions may also be executed without bidding manipula-

tions. For example, the reviewer/paper subject areas and reviewer profiles may be strategically selected to increase the chances of getting assigned the target papers.

Security researchers have demonstrated the vulnerability of paper assignment systems to attacks where an author could manipulate the PDF (portable document format) of their submitted paper so that a certain reviewer gets assigned.[27] These attacks insert text in the PDF of the submitted paper in a manner that satisfies three properties: the inserted text matches keywords from a target reviewers' paper; this text is not visible to the human reader; and 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 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 4. Empirical evaluations on the reviewer-assignment system used at the International Conference on Computer Communications (INFOCOM) demonstrated the high efficacy of these attacks by being able to get papers matched to target reviewers. In practice, there may be other attacks used by malicious participants beyond what program chairs and security researchers have detected to date.

In some cases, the colluding reviewers may naturally be assigned to the target papers without any manipulation of the assignment process:[46] *"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 final defense[16] discussed here 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 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 trade-off between the amount of randomization and the quality of the assignment (Figure 5). This defense was used in the Advancement of Artificial Intelligence (AAAI) 2022 conference.

The recent discoveries of dishonest behavior also pose important questions of law, policy, and ethics for dealing with such behavior: How should program chairs deal with suspicious behavior, and what constitutes appropriate penalties? A case that led to widespread debate is an ACM investigation that 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, and that guilty parties may still participate and possibly continue dishonest behavior in other conferences or grant reviews.

Miscalibration

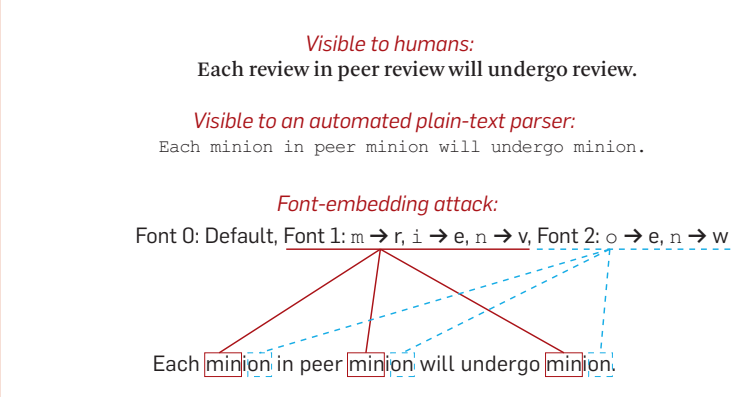
Reviewers are often asked to provide assessments of papers in terms of ratings, and these ratings form an integral part of the final decisions. However, it is well known[12,30,39] that the same rating may have different meanings for different individuals: *"A raw rating of 7 out of 10 in the absence of any other information is potentially useless."*[30] 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, and so on.

Miscalibration causes arbitrariness and unfairness in the peer review process:[39] *"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

Figure 4. An attack on the assignment system via font embedding in the PDF of the submitted paper.[27]

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.



reviewers' individual expectations. As a concrete example, the NeurIPS 2016 conference asked reviewers to rate papers on a scale of 1 through 5 (where 5 is best) and specified an expectation regarding each value on the scale. However, there was a significant difference between the expectations and the ratings given by reviewers.[38] 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%.

There are two popular approaches toward addressing the problem of miscalibration of individual reviewers. The first approach[11,37] is to make simplifying assumptions on the nature of the miscalibration, for instance, assuming that miscalibration is linear or affine. Most works taking this approach assume that each paper p has some "true" underlying rating θ_p, that each reviewer r has two "miscalibration parameters" $a_r > 0$ and b_r, and that the rating given by any reviewer r to any paper p is given by $a_r\theta_p + b_r +$ noise. These algorithms then use the ratings to estimate the "true" paper ratings θ, and possibly also reviewer parameters

The simplistic assumptions described here are frequently violated in the real world.[5] Algorithms based on such assumptions were tried in some conferences, but based on manual inspection by the program chairs, were found to perform poorly.

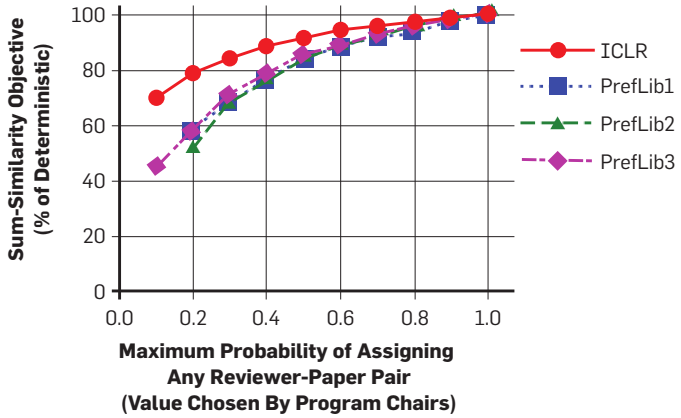
A second popular approach[12,30] toward handling miscalibrations is via rankings: either ask reviewers to give a ranking of the papers they are reviewing (instead of providing ratings), or alternatively, use the rankings obtained by converting any reviewer's ratings into a ranking of their reviewed papers. 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."*[12]

Ratings can provide some information even in isolation. It was shown recently[47] that even if the miscalibration is arbitrary or adversarially chosen, unquantized ratings can yield better results than rankings alone. 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.[38] In such situations, rankings can help break ties among these papers, and this approach was followed in the International Conference on Machine Learning (ICML) 2021. 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 2,425) reviewers had rated some paper as strictly better than another on all four criteria but reversed the pair in the overall ranking.[38]

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 difficult to infer from this data as to whether the reviewer is generally strict. This impediment calls for designing protocols or privacy-preserving algorithms that allow conferences to share some reviewer-specific calibration data with one another to calibrate better.

Figure 5. 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.[16]

The similarity scores for the "ICLR" plot are reconstructed[51] 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.[28]



Subjectivity

We discuss two challenges in peer review pertaining to reviewers' subjective preferences that hamper the objectivity of peer review.

Commensuration bias. Conference program chairs 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. 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, to ensure fairness, all (comparable) papers should be judged by the same yardstick. This issue is known as "commensuration bias."[21]

For 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. The paper's fate thus depends on the subjective preference of the assigned reviewers.

The program chairs of 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 eight criteria into an overall rating. The goal was commendable, but unfortunately, the proposed rules had shortcomings. For example,[33] 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. 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.[33] propose an algorithmic solution to this problem. They consider an often-suggested 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 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 papers. This method was used in the AAAI 2022 conference to identify reviews with significant commensuration bias.

Confirmation bias. A controlled study by Mahoney[25] asked each reviewer to assess a fictitious manuscript. The contents of the manuscripts sent to different reviewers were identical in their reported experimental procedures but differed in their reported results. The study found that reviewers were strongly biased against papers with results that contradicted the reviewers' own prior views. The difference in the results section also manifested in other aspects: a manuscript whose results agreed with the reviewer's views was more likely to be rated as methodologically better, as having a better data presentation, and the reviewer was less likely to catch mistakes in the paper, even though these components were identical across the manuscripts.

Biases Regarding Author Identity

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*)." This is an example of how a review can take into consideration the authors' identities even when we expect it to focus exclusively on the scientific contribution.

Such biases with respect to author identities are widely debated in computer science and elsewhere. These debates have led to two types of peer-review processes: single-blind reviewing where reviewers are shown authors' identities, and double-blind reviewing where author identities are hidden from reviewers. 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 reviewing include: efforts to make a manuscript double blind, efficacy of double blinding (since many manuscripts are posted with author identities on preprint servers and social media), hindrance in checking (self-)plagiarism and conflicts of interest, and the use of author identities as a guarantee of trust for the details that reviewers have not been able to check carefully. In addition, the debate over single- vs-double-blind reviewing rests on the frequently asked question: "Where is the evidence of bias in single-blind reviewing in my field of research?"

A remarkable experiment was conducted at the Web Search and Data Mining (WSDM) 2017 conference,[45] which had 500 submitted papers and 1,987 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. 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 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 U.S., 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[26] 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 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.

Such studies have also prompted a focus on careful design of experimental methods and measurement algorithms to evaluate biases in peer review, while mitigating confounding factors that may arise due to the complexity of the peer-review process.

Making reviewing double blind can mitigate these biases but may not fully eliminate them. Reviewers in three double-blind conferences were asked to guess the authors of the papers they were reviewing.[20] 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 to preprint servers such as arXiv (arxiv.org) 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. These preprints contain information about the authors, thereby potentially revealing the identities of the authors to reviewers. 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 the content of the submitted paper can give clues about the identity of the authors.[20] 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.

These studies provide valuable quantitative information toward policy choices and trade-offs on blinded reviewing. That brings us to our next discussion on norms and policies.

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.

Author 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.[2] We discuss some policies and associated effects pertaining to such author incentives, that are motivated by the rapid increase in the number of submissions in many conferences.

The current research on improving peer review, particularly using computational methods, has only scratched the surface.

Open review. It is said that authors submitting a below-par paper have little to lose but lots to gain: hardly anyone will see the below-par version if it gets rejected, whereas the arbitrariness in the peer-review process gives it some chance of acceptance.

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. A survey[41] of participants at the ICLR 2013 conference, which 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.[2]

The open-review format can also result in some drawbacks; here is one such issue related to public visibility of rejected papers.

Resubmission policies. Many conferences are adopting policies where authors of a paper must provide past rejection information along with the submission. For instance, the 2020 International Joint Conference on Artificial Intelligence (IJCAI) 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 toward ensuring that authors do not simply ignore reviewer feedback, the information of previous rejection could bias the reviewers.

A controlled experiment[43] tested for such a bias. Each reviewer was randomly shown one of two versions of a paper to review: 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.

Rolling deadlines. In conferences with a fixed deadline, a large fraction of submissions are made on or very near the deadline. 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.

The U.S. National Science Foundation experimented with this idea in certain programs.[15] 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.

Introduction to reviewing. While researchers are trained to do research, there is little training for peer review. Several initiatives and experiments have looked to address this challenge. Recently, the ICML 2020 conference adopted a method to select and then mentor junior reviewers, who would not have been asked to review otherwise, with a motivation of expanding the reviewer pool to address the large volume of submissions.[43] 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[8] 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.

While researchers are trained to do research, there is little training for peer review ... Training reviewers at the beginning of their careers is a good start but may not be enough.

Discussions and group dynamics. After submitting the initial reviews, reviewers of a paper are often allowed to see each other's' reviews. The reviewers and the meta reviewer then engage in a discussion to arrive at a final decision.

Several studies[34] 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 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.

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.

Teplitskiy et al.[44] conducted a controlled study to understand influence of other reviewers. They 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 fre-

quently 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.

Stelmakh et al.[43] investigated "herding" effects: Do discussions in peer review lead to the decisions getting biased toward the opinion of the reviewer who initiates the discussion? They found no evidence of such a bias.

Conclusion

The current research on improving peer review, particularly using computational methods, has only scratched the surface of this important application domain. There is much more to be done, with numerous open problems that 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. 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.

A second challenge is the unavailability of data. 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. Designing policies and privacy-preserving computational tools to enable research on such data is an important open problem.

Nevertheless, there is increasing interest among research communities and conferences in improving peer review in a scientific manner. Researchers are conducting several 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. C

References

1. Alon, N., Fischer, F., Procaccia, A., and Tennenholtz, M. Sum of us: Strategy proof selection from the selectors. In *Proceedings of Conf. on Theoretical Aspects of Rationality and Knowledge*, (2011).
2. Anderson, T. Conference reviewing considered harmful. *ACM SIGOPS Operating Systems Rev.*, (2009).
3. Balietti, S., Goldstone, R., and Helbing, D. Peer review and competition in the art exhibition game. In *Proceedings of the National Academy of Sciences*, (2016).
4. Benos, D., et al. The ups and downs of peer review. *Advances in Physiology Education*, (2007).
5. Brenner, L., Griffin, D., and Koehler, D. Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes*, (2005).
6. Brown, T. Peer review and the acceptance of new scientific ideas: Discussion paper from a working party on equipping the public with an understanding of peer peview: November 2002–May 2004. *Sense About Science*, (2004).
7. Cabanac, G. and Preuss, T. Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees. *J. Assoc. Information Science and Tech.* (2013).
8. Callaham, M. and McCulloch, C. Longitudinal trends in the performance of scientific peer reviewers. *Annals of Emergency Medicine*, (2011).
9. Charlin, L. and Zemel, R. The Toronto Paper Matching System: An automated paper-reviewer assignment system. In *Proceedings of ICML Workshop on Peer Reviewing and Publishing Models*, (2013).
10. Fiez, T., Shah, N., and Ratliff, L. A SUPER* algorithm to optimize paper bidding in peer review. In *Proceedings of Conf. Uncertainty in Artificial Intelligence*, (2020).
11. Flach, P., Spiegler, S., Golénia, B., Price, S., Guiver, J., Herbrich, R., Graepel, T., and Zaki, M. Novel tools to streamline the conference review process: Experiences from SIGKDD'09. *SIGKDD Explor. Newsl*, (2010).
12. Freund, Y., Iyer, R., Schapire, R., and Singer, Y. An efficient boosting algorithm for combining preferences. *J. Machine Learning Research*, (2003).
13. Garg, N., Kavitha, T., Kumar, A., Mehlhorn, K., and Mestre, J. Assigning papers to referees. *Algorithmica*, (2010).
14. Guo, L., Wu, J., Chang, W., Wu, J., and Li, J. K-loop free assignment in conference review systems. In *Proceedings of ICNC*, (2018).
15. Hand, E. No pressure: NSF test finds eliminating deadlines halves number of grant proposals. *Science*, (2016).
16. Jecmen, S., Zhang, H., Liu, R., Shah, N., Conitzer, V., and Fang, F. Mitigating manipulation in peer review via randomized reviewer assignments. *NeurIPS*, (2020).
17. Jefferson, T., Alderson, P., Wager, E., and Davidoff, F. Effects of editorial peer review: a systematic review. *JAMA*, (2002).
18. Kobren, A., Saha, B., and McCallum, A. Paper matching with local fairness constraints. In *Proceedings of ACM KDD*, (2019).
19. Lauer, M. Case study in review integrity: Asking for favorable treatment. *NIH Extramural Nexus*, (2020).
20. Le Goues, C., Brun, Y., Apel, S., Berger, E., Khurshid, S., and Smaragdakis, Y. Effectiveness of anonymization in double-blind review. *Commun. ACM*, (2018).
21. Lee, C. Commensuration bias in peer review. *Philosophy of Science*, (2015).
22. Lee, C., Sugimoto, C., Zhang, G., and Cronin, B. Bias in peer review. *J. Assoc. Information Science and Technology*, (2013).
23. Leyton-Brown, K. and Mausam. AAAI 2021—Introduction; https://bit.ly/3r2L3Rr; (min. 8 onward).
24. Littman, M. Collusion rings threaten the integrity of computer science research. *Commun. ACM*, (2021).
25. Mahoney, M. Publication prejudices: an experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research* (1977).
26. Manzoor, E. and Shah, N. Uncovering latent biases in text: Method and application to peer review. In *Proceedings of AAAI*, (2021).
27. Markwood, I., Shen, D., Liu, Y., and Lu, Z. Mirage: Content masking attack against information-based online services. In *Proceedings of USENIX Security Symp.*, (2017).
28. Mattei, N. and Walsh, T. Preflib: A library for preferences. In *Proceedings of Intern. Conf. Algorithmic Decision Theory*. Springer, 2013; http://www.preflib.org
29. Mimno, D. and McCallum, A. Expertise modeling for matching papers with reviewers. In *Proceedings of KDD*, (2007).
30. Mitliagkas, I., Gopalan, A., Caramanis, C., and Vishwanath, S. User rankings from comparisons: Learning permutations in high dimensions. In *Proceedings of Allerton Conf.*, (2011).
31. Murphy, J., Hofacker, C., and Mizerski, R. Primacy and recency effects on clicking behavior. *J. Computer-Mediated Commun.*, (2006).
32. Neubig, G., Wieting, J., McCarthy, A., Stent, A., Schluter, N., and Cohn, T. ACL reviewer matching code; https://github.com/acl-org/reviewerpaper-matching.
33. Noothigattu, R., Shah, N., and Procaccia, A. Loss functions, axioms, and peer review. *J. Artificial Intelligence Research*, (2021).
34. Pier, E., Raclaw, J., Kaatz, A., Brauer, M., Carnes, M., Nathan, M., and Ford, C. Your comments are meaner than your score: Score calibration talk influences intra-and inter-panel variability during scientific grant peer review. *Research Evaluation* (2017).
35. Rennie, D. Let's make peer review scientific. *Nature*, (2016).
36. Rodriguez, M., Bollen, J., and Van de Sompel, H. Mapping the bid behavior of conference referees. *J. Informetrics* (2007).
37. Roos, M., Rothe, J., and Scheuermann, B. How to calibrate the scores of biased reviewers by quadratic programming. In *Proceedings of AAAI*, (2011).
38. Shah, N., Tabibian, B., Muandet, K., Guyon, I., and Von Luxburg, U. Design and analysis of the NIPS 2016 review process. *JMLR*, (2018).
39. Siegelman, S. Assassins and zealots: Variations in peer review. *Radiology*, (1991).
40. Smith, R. Peer review: Reform or revolution? Time to open up the black box of peer review. (1997).
41. Soergel, D., Saunders, A., and McCallum, A. Open scholarship and peer review: A time for experimentation, (2013).
42. Stelmakh, I., Shah, N., and Singh, A. PeerReview4All: Fair and accurate reviewer assignment in peer review. *JMLR*, (2021).
43. Stelmakh, I., Shah, N., Singh, A., Daumé III, H., and Rastogi, C. Experiments with the ICML 2020 peer-review process, (2020); https://blog.ml.cmu.edu/2020/12/01/icml2020exp/
44. Teplitskiy, M., Ranub, H., Grayb, G., Meniettid, M., Guinan, E., and Lakhani, K. Social influence among experts: Field experimental evidence from peer review, (2019).
45. Tomkins, A., Zhang, M., and Heavlin, W. Reviewer bias in single-versus double-blind peer review. In *Proceedings of the National Academy of Sciences*, (2017).
46. Vijaykumar, T. Potential organized fraud in ACM/IEEE computer architecture conferences, (2020); https://bit.ly/3o2Zjb3.
47. Wang, J. and Shah, N. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *Proceedings of AAMAS*, (2019).
48. Ware, M. Publishing research consortium peer review survey 2015. *Publishing Research Consortium*, (2016).
49. Wing, J. and Chi, E. Reviewing peer review. *Communications of the ACM* (2011).
50. Wu, R., Guo, C., Wu, F., Kidambi, R., van der Maaten, L., and Weinberger, K. Making paper reviewing robust to bid manipulation attacks, (2021); *arXiv:2102.06020*.
51. Xu, Y., Zhao, H., Shi, X., and Shah, N. On strategyproof conference review. In *Proceedings of IJCAI*, (2019).

An extended version of this article discussing additional challenges, experiments, and solutions is available at http://bit.ly/PeerReviewOverview.

Nihar B. Shah is an assistant professor in the Machine Learning and Computer Science departments of Carnegie Mellon University, Pittsburgh, PA, USA.