

## Reviewer No. 2: Old and New Problems in Peer Review

Iryna Gurevych<sup>\*1</sup>, Anna Rogers<sup>\*2</sup>, Nihar B. Shah<sup>\*3</sup>, and  
Jingyan Wang<sup>†4</sup>

1 Department of Computer Science, TU Darmstadt, DE.  
[iryna.gurevych@tu-darmstadt.de](mailto:iryna.gurevych@tu-darmstadt.de)

2 Department of Computer Science, IT University of Copenhagen, DK.  
[aarog@itu.dk](mailto:aarog@itu.dk)

3 Machine Learning and Computer Science Departments, Carnegie Mellon  
University – Pittsburgh, US. [nihars@cs.cmu.edu](mailto:nihars@cs.cmu.edu)

4 Georgia Institute of Technology – Atlanta, US. [jingyanw@gatech.edu](mailto:jingyanw@gatech.edu)

---

### Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 24052 “Reviewer No. 2: Old and New Problems in Peer Review”. This seminar provided a point of reflection on decades of personal experience of the participants in organizing different kinds of peer-reviewed venues in Natural Language Processing (NLP) and beyond, enabling an in-depth discussion of what has been tried, what seems to work and what doesn’t. The outcomes of the seminar include a white paper co-authored by most of the seminar participants, which outlines the research program, methodological and empirical challenges for NLP for peer review. The discussions at the seminar also resulted in several concrete policy proposals and initiatives, some of which are already in motion at the Association for Computational Linguistics and elsewhere.

**Seminar** January 28 – February 2, 2024 – <https://www.dagstuhl.de/24052>

**2012 ACM Subject Classification** Computing methodologies → Natural language processing;  
Information systems → Expert systems; Information systems → Web applications

**Keywords and phrases** Peer Review, Natural Language Processing

**Digital Object Identifier** 10.4230/DagRep.14.1.130

## 1 Executive Summary

*Anna Rogers (IT University of Copenhagen, Denmark, [aarog@itu.dk](mailto:aarog@itu.dk))*

*Nihar Shah (Carnegie Mellon University, USA, [nihars@cs.cmu.edu](mailto:nihars@cs.cmu.edu))*

*Iryna Gurevych (TU Darmstadt, Germany, [iryna.gurevych@tu-darmstadt.de](mailto:iryna.gurevych@tu-darmstadt.de))*

**License**  Creative Commons BY 4.0 International license  
© Anna Rogers, Nihar Shah, and Iryna Gurevych

## Background

Peer review is the best mechanism for assessing scientific validity of new research that we have so far. But this mechanism has many well-known issues, such as the different incentives of the authors and reviewers, difficulties with preserving reviewer and author anonymity to avoid social biases [22, 58, 68, 50, 39], confirmation and other cognitive biases [71, 16, 1, 32, 64], that even researchers fall prey to. These intrinsic problems are exacerbated in interdisciplinary fields like Natural Language Processing (NLP), where groups of researchers may vary so much in their methodology, terminology, and research agendas, that sometimes they have trouble even recognizing each other’s contributions as “research” [53].

---

\* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed  
under a Creative Commons BY 4.0 International license

Reviewer No. 2: Old and New Problems in Peer Review, *Dagstuhl Reports*, Vol. 14, Issue 1, pp. 130–161

Editors: Iryna Gurevych, Anna Rogers, and Nihar Shah



DAGSTUHL  
REPORTS

Dagstuhl Reports  
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Our Dagstuhl Seminar covered a range of topics related to organization of peer review in NLP, Machine Learning (ML), and venues more broadly in Artificial Intelligence for intelligent support of peer-reviewing, including the following:

- Improving the paper-reviewer matching by processes/algorithms that take into account both topic matches and reviewer interest in a given research question.
- Peer review vs methodological and demographic diversity in the field.
- Better practices for designing review forms and peer review policies.
- Improving the structural incentives for reviewers.
- Use of NLP and ML for intelligent peer reviewing support: increasing the quality and efficiency of peer review, opportunities and challenges.
- Peer-reviewing and research integrity.

## Goals

We intended for the seminar to serve as a point of reflection on decades of personal experience of the participants in organizing different kinds of peer-reviewed venues in NLP and beyond, enabling an in-depth discussion of what has been tried, what seems to work and what doesn't. The objectives of the seminar included collaborative research on the methodological challenges of peer review, NLP and ML for intelligent support of peer-reviewing and actionable proposals, for example for paper-reviewer assignment policies and peer reviewing guidelines and workflows, informed by the experience of participants as chairs, editors, conference organizers, and reviewers.

## Outcomes

The seminar was attended by researchers at different levels of seniority and from a variety of research backgrounds. While a large number of the attendees represented the Natural Language Processing community, about a third represented other communities within the broader sphere of Machine Learning. Most discussions focused on the peer review in the world of ultra-large conferences with thousands of submissions, but we also had a senior representative from fields where journals are most prominent, and hence an opportunity to learn from their experience.

## Knowledge Sharing

The seminar started by contributed talks by a diverse group of participants (see section 3), which allowed us to share relevant experience and research findings pertinent to the topics of the seminar, across communities. Peer review issues are at most discussed in the business meetings of specific conferences, and there are hardly any opportunities to share this knowledge across communities. Hence, this knowledge-sharing section of the seminar by itself has been unique, and it proved to be useful to establish a common ground and points of reference for subsequent work during the seminar.

## Problem elucidation

After the contributed talks, all the subsequent work was organized into breakout sessions (two running in parallel) on the following topics:

- Integrity issues in peer review (2 sessions)
- Diversity issues in peer review (3 sessions)

- Assisting peer review with NLP (3 sessions)
- Peer review policies (2 sessions)
- Incentives in peer review (3 sessions)
- Paper-reviewer matching (3 sessions)

The work in all these sessions combined brainstorming, establishing common ground and terms, discussing practical solutions for specific problems that were tried in various communities represented by the participants, and ideas for the future. Summaries of work in all the above topics are provided in section 4.

There were also two slots reserved for unstructured breakouts, and every day concluded with an overall summary session in which the leads for various topics summarized the discussions in that day.

### **Research program and community formation**

The key outcome of the seminar is a white paper with the working title “*What Can NLP do for Peer Review?*”, co-authored by the majority of the participants of the seminar. It formulates the goals and research agenda of assisting peer review with NLP techniques, and we hope that it would play a key role in shaping this research field. This paper is available at [80]. It is accompanied by a repository for tracking research papers in this area, available at <https://github.com/OAfzal/nlp-for-peer-review>.

### **Concrete policies**

The work in various breakout sessions culminated in the proposal of a new peer review committee for the Association of Computational Linguistics (ACL), that would oversee the systematic research and data-driven peer review policy development in the NLP community. This proposal has already been formally submitted to the ACL board, and generally approved. The work on formally establishing and announcing the committee will be finished in 2024.

### **Research problems and collaborations**

This Dagstuhl Seminar also helped surface and crystallize a number of open problems, and alongside, helped establish inter-disciplinary collaborations for working on them, which may not have happened if not for this seminar.

### **Next steps**

This Dagstuhl Seminar brought together an international, community of NLP and ML researchers from academia and industry to discuss the problems with peer review in large-scale conferences. This is a topic for which various subcommunities have different practices, expectations, and strong opinions, and the seminar brought much discussion throughout all days of the seminar (and also long into the night). This was also a unique opportunity to share the lessons learned the hard way, on issues which are often misconstrued as merely organizational issues. In fact, this is something to be seriously discussed as a research problem, for which much conceptual and empirical work is needed.

We hope that this seminar was the first in a series of events devoted to this topic, and that this inaugural event proves pivotal in the formation of a cohesive research community. The white paper prepared as the main outcome of this seminar aims to galvanize the NLP

and ML communities by offering them a wide selection of realistic research problems with peer review as an application area.

## 2 Table of Contents

### Executive Summary

<i>Anna Rogers, Nihar Shah, and Iryna Gurevych</i> . . . . .	130
--	-----

### Overview of Talks

Natural Language Processing Meets Scientific Argumentation: The Case of Peer Reviewing <i>Anne Lauscher</i> . . . . .	136
Peer review as text-based collaboration <i>Iliia Kuznetsov</i> . . . . .	136
Peer review at ACL'23 <i>Anna Rogers</i> . . . . .	136
Mitigating Biases in Peer Review <i>Jingyan Wang</i> . . . . .	137
Technical Pitfalls and Possibilities in a [Rolling] Review System <i>Jonathan Kummerfeld</i> . . . . .	137
Better Peer Review via AI <i>Kevin Leyton-Brown</i> . . . . .	137
HotCRP for Dagstuhl <i>Eddie Kohler</i> . . . . .	138
Natural Language Processing for Peer Review Assistance <i>Nils Dycke</i> . . . . .	138
Studies on Citation Influence and Prediction <i>Xiaodan Zhu</i> . . . . .	138
Semantic Scholar & Peer Review <i>Tom Hope</i> . . . . .	139
Some experiments in reviewing <i>Nihar B. Shah</i> . . . . .	139
Using ARR to Tackle Climate Change <i>Roy Schwartz</i> . . . . .	139
Evaluating the peer review process <i>Alexander Goldberg</i> . . . . .	140
Optimization of Scoring Rules <i>Jason Hartline</i> . . . . .	140
ARR Reflexions on 1.5 years <i>Thamar Solorio</i> . . . . .	140
Working conditions and satisfaction of early career NLP researchers in the era of LLMs <i>Sheng Lu</i> . . . . .	141
Ethics Reviewing in NLP <i>Margot Mieskes</i> . . . . .	141

**Working Groups**

Working Group on Policies for Peer Review  
*Anna Rogers* . . . . . 142

Working Group on Diversity in Peer Review  
*Anna Rogers* . . . . . 144

Working Group on Paper-Reviewer Matching  
*Anna Rogers* . . . . . 147

Working Group on Incentives in Peer Review  
*Nihar Shah* . . . . . 149

Working Group on Natural Language Processing (NLP) for Peer Review  
*Nihar Shah* . . . . . 151

Working Group on Integrity in Peer Review  
*Nihar Shah, Iryna Gurevych* . . . . . 153

**Participants** . . . . . 161

### 3 Overview of Talks

The ordering of the talks is randomized (as opposed to ordering alphabetically).

#### 3.1 Natural Language Processing Meets Scientific Argumentation: The Case of Peer Reviewing


*Anne Lauscher (Universität Hamburg, DE)*

License  Creative Commons BY 4.0 International license  
© Anne Lauscher

Peer reviewing is a prime example of scientific argumentation: the authors present their work, a scientific claim in the form of a scientific publication, to the reviewers. The reviewers then engage in a debate, arguing why the claim should or should not be accepted to the body of tentatively accepted knowledge in a field. In this talk, I argue that theories and approaches rooted in argumentation theory and NLP/ computational argumentation can thus be leveraged to effectively support the different actors within this process. For instance, I discuss the case of automatic rebuttal template generation based on Jui-Jitsu argumentation [47] – a theory that has been proposed for the case of anti-science argumentation.

#### 3.2 Peer review as text-based collaboration

*Iliia Kuznetsov (TU Darmstadt, DE)*

License  Creative Commons BY 4.0 International license  
© Iliia Kuznetsov

Text-based collaboration is at the core of modern knowledge work. Peer review is a prime example of text-based collaboration, where authors, reviewers and area chairs work together to improve the initial paper draft via review and feedback. I introduce InterText [30] – a major project at UKP lab dedicated to the modeling of text as a living object in context, which we instantiate in the domain of academic peer review. I will describe our graph-based document representation and three novel intertextual modeling tasks – pragmatic tagging, linking and versioning. I will present F1000RD – the first corpus for intertextual NLP, and will discuss the results of our annotation studies, analysis, as well as existing challenges and ways forward towards NLP for living texts in context.

#### 3.3 Peer review at ACL'23

*Anna Rogers (IT University of Copenhagen, DK)*

License  Creative Commons BY 4.0 International license  
© Anna Rogers

ACL'23 implemented several changes to the standard ACL peer review process, followed up with survey-based evaluations from reviewers and area chairs. This talk describes the most successful innovations, including (a) paper-reviewer matching based on area + contribution + language 3-dimensional criteria, (b) soundness+excitement scores replacing the single

“overall recommendation” score, (c) structured reviewer complaints and issue flagging, (d) new format for reporting on peer review data as part of conference proceedings, (d) updated reviewer guidelines & first ACL policy on generative AI.

### 3.4 Mitigating Biases in Peer Review

*Jingyan Wang (Georgia Institute of Technology – Atlanta, US)*

License © Creative Commons BY 4.0 International license  
© Jingyan Wang

I describe a particular type of bias in peer review where authors provide ratings to the reviewers’ review quality [74]. In this setting, a particular type of bias is induced by the author’s outcome (i.e., whether the author’s paper is accepted or not). In this work, we propose mild ordering assumptions to model the bias, and design a debiasing algorithm to correct student ratings adaptively to the amount of bias vs noise in the data. I also briefly describe other types of human bias in ratings, including miscalibration [73] and different causes of sequential effects [72].

### 3.5 Technical Pitfalls and Possibilities in a [Rolling] Review System

*Jonathan Kummerfeld (The University of Sydney, AU)*

License © Creative Commons BY 4.0 International license  
© Jonathan Kummerfeld

A rolling review system differs from conferences because it has memory across time. That introduces a range of challenges (i.e., “pitfalls”) as well as opportunities (i.e., “possibilities”). I provide an overview of the technical infrastructure that is used by the ACL Rolling Review team, and discuss our experiences in the past and hopes for the future. In particular, I describe how tools can help with many reviewing tasks, but there is typically still a social dimension that is not solved. This context can help inform discussions in the seminar in terms of what exists today and what is possible tomorrow.

### 3.6 Better Peer Review via AI

*Kevin Leyton-Brown (University of British Columbia – Vancouver, CA)*

License © Creative Commons BY 4.0 International license  
© Kevin Leyton-Brown

The talk summarized the reviewer-paper matching system used by Kevin and Mausam at AAAI 2021. The system decomposed into (1) collecting and processing input data; (2) formulating an optimization problem; (3) solving that problem; (4) two-phase reviewing. It also summarized an empirical analysis of data from the conference. The slides, but not the oral presentation, also briefly summarize a similar system for peer grading in large classes.



### 3.7 HotCRP for Dagstuhl

*Eddie Kohler (Harvard University – Allston, US)*

License  Creative Commons BY 4.0 International license  
© Eddie Kohler

HotCRP is an online submission and review system used broadly in the CS theory, systems, networking, architecture, and security communities. Every reviewing community has values, as does every reviewing system; HotCRP’s values include speed, smoothness, ease of use, and openness to PC members. The talk also highlights some thoughts from earlier peer review discussions in other communities, and issues experienced by review system designers.

### 3.8 Natural Language Processing for Peer Review Assistance

*Nils Dycke (TU Darmstadt, DE)*

License  Creative Commons BY 4.0 International license  
© Nils Dycke

Research on natural language processing (NLP) to support reviewers, authors and editors in the academic peer review process is still in its early stages. NLP for peer reviewing assistance holds promise for improving the quality of reviewing and increasing the efficiency of the process. In this talk I give an overview of the hurdles faced in the early stages of this young research field including the scarcity of open data, the lack of practical tasks, and the need for tools to disseminate NLP models to support peer review. I explain the rationale behind why and how NLP can help reviewers improve their work. Finally, I highlight our past work addressing these challenges, encompassing our data collection at ARR [11], the peer reviewing data corpus NLPEER [12], and the reading assistance tool CARE [79].

### 3.9 Studies on Citation Influence and Prediction


*Xiaodan Zhu (Queen’s University – Kingston, CA)*

License  Creative Commons BY 4.0 International license  
© Xiaodan Zhu

We believe that most papers are based on a set of essential references. By an essential reference, we mean a reference that was highly influential or inspirational for the core ideas of the citing paper. In this talk, I first describe our previous research on predicting influential references from non-influential ones. Then, I move on to present our recent work on citation prediction in the legal domain, where citations are a foundation for many legal decision-making processes. I specifically present a prototype-based model that has some built-in interpretability for legal citation prediction.

### 3.10 Semantic Scholar & Peer Review


*Tom Hope (The Hebrew University of Jerusalem, IL)*

License  Creative Commons BY 4.0 International license  
© Tom Hope

I presented some work done by AI2’s Semantic Scholar group which can be useful for building NLP-powered peer review systems. This includes the Semantic Scholar API which allows developers to access rich publication and author information from the Semantic Scholar Academic Graph, the Bridger tool for author matching based on shared methods and tasks authors work on, the Aspire scientific document embedding model for finding related papers, and the Semantic Reader platform which can support enhanced interactive reading experiences. Finally, I presented some of our recent work directly focused on peer review, led by Mike D’Arcy and Doug Downey: ARIES, which focuses on matching review comments to edits and generating edits in response to comments [9], and MARG, which introduces a multi-agent system for automatically generating reviews [8].

### 3.11 Some experiments in reviewing

*Nihar B. Shah (Carnegie Mellon University – Pittsburgh, US)*

License  Creative Commons BY 4.0 International license  
© Nihar B. Shah

We discuss a set of experiments in scientific reviewing:

1. **Preprinting:** An anonymous survey of reviewers in (dual anonymous) ICML and EC conferences on whether they searched for their assigned papers online, finding that over a third of the reviewers do so [50].
2. **Author perceptions:** An experiment in NeurIPS 2021 finding that authors significantly overestimate the chances of their own papers getting accepted, and that co-authors significantly disagree on the relative merits of their co-authored papers [49].
3. **Discussions:** A randomized controlled trial at UAI 2022 on whether to show reviewers each others’ identities or not [48].
4. **Rebuttals:** A randomized controlled trial which finds no evidence of reviewers anchoring to their original opinions [35].
5. **AI reviewing:** A “chimera” test which finds that AI reviewers are unable to call out a nonsensical paper formed by pasting together parts of multiple papers (unpublished); and an evaluation of LLMs on their (in)capability to perform certain tasks in the review process [57].

### 3.12 Using ARR to Tackle Climate Change

*Roy Schwartz (The Hebrew University of Jerusalem, IL)*

License  Creative Commons BY 4.0 International license  
© Roy Schwartz

The environmental effect of AI has been mostly studied in the context of the carbon footprint of models, while far less attention has been devoted to the cost of conference air travel. In this talk we present experiments showing that this factor also has a substantial environmental

cost. To partly mitigate this cost, we propose to allow authors to choose where to present their accepted papers, by allowing ARR to make accept/reject decisions. Our assumption is that many of them would prefer shorter travel if given the option. Our experiments show that this proposal could lead to substantial reductions in carbon emissions.

### 3.13 Evaluating the peer review process

*Alexander Goldberg (Carnegie Mellon University – Pittsburgh, US)*

License  Creative Commons BY 4.0 International license  
© Alexander Goldberg

Evaluating outcomes and risks in the peer review process often proves quite challenging. This talk covers research on two aspects of evaluating peer review – (1) assessing review quality and (2) understanding privacy risks associated with open and transparent peer review. In evaluation of review quality, we highlight two biases that can arise in particular a positive bias towards (uselessly) longer reviews and bias by authors towards positive reviews [19]. On privacy risks, we describe deanonymization risk arising from revealing public comments on papers [18].

### 3.14 Optimization of Scoring Rules

*Jason Hartline (Northwestern University – Evanston, US)*

License  Creative Commons BY 4.0 International license  
© Jason Hartline

This paper introduces an objective for optimizing proper scoring rules. The objective is to maximize the increase in payoff of a forecaster who exerts a binary level of effort to refine a posterior belief from a prior belief. In this framework we characterize optimal scoring rules in simple settings, give efficient algorithms for computing optimal scoring rules in complex settings, and identify simple scoring rules that are approximately optimal. In comparison, standard scoring rules in theory and practice – for example the quadratic rule, scoring rules for the expectation, and scoring rules for multiple tasks that are averages of single-task scoring rules – can be very far from optimal.

These scoring rules are applied to the task of grading peer reviews against TA reviews of homework in advanced undergraduate courses. Here the classical scoring rules give little incentive for effort and this incentive is improved by optimal scoring rules.

### 3.15 ARR Reflexions on 1.5 years

*Thamar Solorio (MBZUAI – Abu Dhabi, AE)*

License  Creative Commons BY 4.0 International license  
© Thamar Solorio

The ACL Rolling Review (ARR) initiative launched in 2021 as a centralized review system for our \*CL conferences. I've been serving as a co-editor in chief for ARR for 1.5 years now. In this talk I will present an overview of ARR, a snapshot of a typical two months reviewing

cycle. Then, I'll highlight some of the major challenges we face when trying to fulfill the reviewing needs of the community, while simultaneously being responsive to the requests for changes in our already tight reviewing process.

### 3.16 Working conditions and satisfaction of early career NLP researchers in the era of LLMs

*Sheng Lu (TU Darmstadt, DE)*

License  Creative Commons BY 4.0 International license  
© Sheng Lu

The rapid development of large language models (LLMs) has led to challenges such as a lack of rigor in evaluation, an overwhelming amount of literature, and potentially negative impact on researchers' well-being due to the fast pace and a growing publication pressure. In June 2023, the Ubiquitous Knowledge Processing (UKP) Lab at the Technical University of Darmstadt and the Chair for Statistics and Data Science in Social Sciences and the Humanities (SODA) at Ludwig-Maximilians-Universität in Munich conducted an online survey with over 700 early career NLP researchers worldwide to gain insights into their working conditions and satisfaction in the era of LLMs. Even though these survey data do not allow us to make generalizable inferences, they provide important hints about the current working conditions and satisfaction of early career researchers in the NLP community. It would be beneficial for further research to include a more diverse range of respondents holding different types of positions and residing in different regions.

### 3.17 Ethics Reviewing in NLP

*Margot Mieskes (Hochschule Darmstadt, DE)*



License  Creative Commons BY 4.0 International license  
© Margot Mieskes

For a couple of years now Ethics reviewing has been established as a part of the regular reviewing process in the NLP domain. But experience shows that this is a very time-consuming and far from structured process. In my talk I will present experience and lessons learned from being Ethics Co-Chair for three major NLP conferences (EMNLP 2021, EMNLP 2022 and LREC-COLING 2024) which used different reviewing platforms and the general chairs and program chairs had different levels of experience. I will also present some suggestions on how to improve the process and how to ensure that authors support the Ethics reviewing as part of the scientific process, rather than opposing it.

## 4 Working Groups

### 4.1 Working Group on Policies for Peer Review

*Anna Rogers (IT University of Copenhagen, DK, arog@itu.dk)*

License  Creative Commons BY 4.0 International license  
 Anna Rogers

This working group focused on defining the scope of the peer review policies, and considering concrete problems in the communities with which the participants were familiar, and which could be addressed via various policies.

The discussion opened by considering what even counts as a policy. The following definitions were proposed: (a) a shared set of values and beliefs that evolves over time, (b) a long-term commitment, (c) a way to ensure consistent behavior by chairs etc.

This working group had 3 meetings in total, covering numerous topics. The topics that provoked the most discussion and suggested action items are summarized in this section.

#### 4.1.1 Award policies

Currently most conferences do little to encourage good reviewer behavior. The ACL conferences recently created a policy to increase the number of reviewers and chairs nominated for awards to 1-1.5%<sup>1</sup>, but this is still not enough: the chance to get such an award is still relatively small and not worth extra effort on the part of the reviewer. One suggestion was that if the awards are given to as much as 50% reviewers, this would reverse the incentive structure: *not* getting the award would create a negative social signal.

The current award offered to the outstanding reviewers at \*ACL conferences is either free conference attendance as a virtual participant, or a discount on the in-person attendance. This does not necessarily reflect the needs of the reviewers, some of whom come from wealthy industry labs and do not need the monetary incentive. A survey could be organized to establish what other kinds of incentives could be useful. Some other reward ideas proposed in the discussion included:

- Sharing reviewer history to ORCID, given that it's possible to create generic service records there (e.g. conference names, number of papers reviewed, reviewer awards);
- 10% conference discount to 50% reviewers? (PeerJ case)
- Sharing generic reviewer history with potential employers (e.g., lab leaders looking for PhD students/postdocs), as it provides a useful signal about reliability and commitment;
- Likewise, area chairs could find it useful in grant applications if the area and statistics of their work could be shared to showcase community leadership in their research area (e.g. “outstanding area chair for question answering track, ACL 2024”);
- Various certificates showing different levels of achievement (reviewer certificate, outstanding reviewer certificate);
- A “star” system for reviewers where people can gain a star for good reviewer behavior (e.g., number of reviews, on time, high-quality, detailed reviews, emergency reviews), and feedback can be given to reviewer for moving up in the star scoring.

<sup>1</sup> [https://2023.aclweb.org/program/best\\_reviewers/](https://2023.aclweb.org/program/best_reviewers/)

### 4.1.2 Review form design

Most conferences, with which the participants were familiar, use unstructured or semi-structured review forms, with questions like “summary”, “strengths” and “weaknesses”. An alternative is to have structured review forms where reviewers are asked to evaluate various aspects separately. The group discussed evidence from Elsevier studies on the use of structured review forms [38], which increase agreement between reviewers, and decrease the cognitive load (allowing them to comment on the aspects of the paper for which they have expertise). From the perspective of program chairs, structured review forms can allow the chairs to better understand which parts of the paper have been reviewed more reliably, and mitigate issues like commensuration bias [32, 43]. Such forms could include concrete questions relevant to a specific paper type: e.g. “Is the statistical analysis sound?”, with the answer option “I’m not an expert on this”.

Structured review forms allow for better control to remove subjective categories in the review where biases from author identities are easier to creep in. If reviewers focus on the technical correctness of the work, these aspects might alleviate the problem of non double blind reviewing.

### 4.1.3 Institutional memory

At present, most conferences are organized on one-off basis, with most program chairs not having access even to reviewer history from the previous editions of the same conference, much less across venues. However, each cycle generates much useful information (late reviews, low-quality reviews, outstanding reviews etc.), and simply being able to track and use this information would probably help a lot in increasing the review quality. But there is no structure in place to maintain and organize this information. ACL Rolling Review could theoretically perform this for NLP community, as it already performs some long-term data collection [11], but there needs to be a broad mandate of storing and using this information for the organizational purposes.

It is not clear what should happen to the “bad” reviewers: simply de-prioritizing them in assignment process is not a negative incentive, since in practice it just means less work for them. Conversely, the “good” reviewers should not be rewarded by simply having more work assigned to them.

The group also discussed the privacy vs equity issue: peer review data is highly confidential and should not be disclosed without reviewer consent, especially since authors are often tempted to publicly bash their reviewers. At the same time, in some cases the privacy considerations protect the wrongdoers. One possibility to introduce reviewer consent to this process is to have reviewer agreements include an optional checkbox for granting the authors the right to use reviews however they choose, including for public discussion.

Finally, the group discussed the possibility of having a single reviewer profile within a system with an institutional memory, such as ARR, that would list various items from reviewer history, so as to visualize their impact on the community and highlight the fact that their reviewing record *is* tracked. This profile could include the following information: how many reviews they did, for what tracks, how many were late, how many were emergency reviews, what were the outcomes for the reviewed papers, how many best paper nominations (and the outcomes for those papers), ratings distribution for this reviewer vs conference mean, length of reviews vs mean, number of discussions vs mean, any feedback notes from area chairs, number and issue types flagged by the authors.

#### 4.1.4 Findings Policy

Most of the current \*ACL conferences have the more prestigious main track publication (e.g. proceedings of ACL), and a less-prestigious *Findings* <https://2020.emnlp.org/blog/2020-04-19-findings-of-emnlp> publication with a higher acceptance rate (usually 30-40%). The current workflow for peer review through ACL Rolling Review is that acceptance decisions are decoupled from review process, and are done independently by senior area chairs, sometimes months after the reviews were finished. This is unsatisfactory for the authors, who of course want faster decisions and not just reviews.

The group discussed the possibility of having ACL Rolling Review provide at least *Findings* decisions, which can be made purely on the ground of the judgement of technical soundness of the paper. Once such a decision was made, a *Findings* publication is guaranteed. Then the authors can still commit the paper to be considered for the main track publication. If the paper was rejected from *Findings*, it needs to go through the revise-and-resubmit process.

#### 4.1.5 Next Steps: Peer Review Committee

Based on the above discussion, the seminar participants developed a proposal to the ACL executive board to establish the ACL Peer Review Committee: a working group dedicated to the development of peer review policies across ACL venues. This group would consider and help to develop proposals relevant to the peer review process, which originate either from ACL venues or the community. It would also monitor the implementation of any proposals, and ensure that the venues implementing them would consistently report on the results of any changes.

The proposal was approved, and the committee will be formally established in 2024. This committee will have a broad mandate to analyze internal peer review data from ACL venues for the purposes of developing evidence-based policies and improving the peer review process (but not for independent research by the committee members). For the sake of transparency, any results of such analysis will be made available as public reports.

While this committee will serve only ACL community, it is a good test case, since it has a lot of major conferences and already possesses the infrastructure for shared organization and institutional memory between them (ACL Rolling Review). The successful practices from this initiative could be shared with other communities.

## 4.2 Working Group on Diversity in Peer Review

*Anna Rogers (IT University of Copenhagen, DK, arog@itu.dk)*

License  Creative Commons BY 4.0 International license  
© Anna Rogers

This working group had two meetings, focusing on a wide range of topics, some of which overlapped with the discussions in the Policy and Incentives groups. We started by noting the issues with even defining “diversity”: in peer review it is often discussed in terms of geographic diversity and levels of seniority, but it can have many other facets, such as representation of various topics, subfields and languages.

### 4.2.1 Reviewer Pool Representativeness

The conferences often start recruiting from the reviewer lists from past conferences, which means that a biased sample could keep being reused. Extra reviewers could be brought in through the networks of the chairs, which could also contribute to the bias. In ACL'23, there was a vast imbalance between the number of submissions and reviews contributed by Chinese researchers [54].

Perhaps the process should be more often, with the reviewers recruited through an open call and self-nomination via a sign-up form. Community groups such as Widening NLP could also help. To better estimate the extent of the problem, a conference registration form could include a question about reviewing (e.g. *Have you published here in the past 10 years? Have you reviewed here? If not, why?*)

For the specific China under-representation issue, the recommendation is to make sure that there are enough area chairs who are from China and based in China, and to ask them to help recruit widely from their networks. It would also help to connect with the National Science Foundation, there is an identifier system that is widely used within China.

### 4.2.2 Preprinting and Double-Blind Review

ACL used to have an embargo on posting papers on ArXiv prior to submitting to ARR or ACL conferences; but the policy recently changed to lift all restrictions.<sup>2</sup> It is still considered an integrity violation to search for papers one is reviewing, although some reviewers might use it a check for plagiarism. This change of policy necessitates discussion of how we can protect double-blind review, given that single-blind review is known to be influenced by demographic features associated with authors, such as country of affiliation, lab affiliation, fame, seniority etc. [46, 68, 39, 22, 59]. Among the possible ways to remedy this situation the seminar participants discussed the following:

- Peer review platforms could try to automate the checks for plagiarism, undisclosed preprints, previous publications etc. Then the reviewers could be told that they do not need to look for this. However, it is likely that many would still deliberately deanonymize the submissions [50].
- During paper-reviewer matching, the reviewers who disclosed the knowledge of a submission could be de-prioritized. However, the reviewers would have to input this information for all submissions that they are qualified to review. This also provides an extra opportunity to collusion rings [70, 34, 26], whose members could pretend to recognize work outside of the ring to increase the chances of assignments to each others' papers.
- The scores from reviewers who recognized the submission could be visually distinguished in the review reports presented to chairs, to help them downweight such scores (as they are potentially unreliable).
- "Confidence" score is too ambiguous, it could be interpreted as confidence about impact, novelty etc. It should be worded as confidence in the assessment of the technical aspects of the paper. Maybe this could help the reviewers to calibrate their assessments better even in a single-blind situation.

---

<sup>2</sup> <https://www.aclweb.org/portal/content/report-acl-committee-anonymity-policy>



### 4.2.3 Equity of Workload

A related issue is equity: peer-review work needs to be shared equally between recruited people, but the current conferences have the problem of too many qualified people leaving the reviewer pool after going into industry jobs. At the same time, we always have a lot of junior people who do not have the incentives to go through reviewer training. The following ideas were discussed to try to counter this trend and to have more equitable review loads:

- As an incentive, first-time reviewers could have a reduced load to help them spend more time on individual papers.
- First-time reviewers could have dedicated recognition/awards, and priority in bidding/assignments.
- First-time reviewers could be offered mentoring [63], and given the option to nominate their own PI as the mentor.
- Mentoring option in reviewer invitations: usually the invitations only have accept/decline options. There could be an option to nominate someone you would mentor.
- Venues could mandate a review load for the authors of papers submitted to a venue, with the possibility for authors who are not qualified or are contributing in other ways to be excused. For each author on submission, a form could be provided to indicate their availability as reviewers, with some common pre-set options to excuse some authors (e.g. “too junior”, “on leave”, “collaborator from a different field”, etc.).
- For papers with more than ten authors, require more authors to review.

### 4.2.4 Equity of Dissemination

#### Social Media

A phenomenon that was recently discussed is the “science influencers”: the mentions of research papers by certain Twitter accounts result in much increased popularity and citation counts for these papers [75]. Only few papers get promoted this way, and they could be more diverse in terms of geographical and gender representation of the promoted authors. Is it possible for the venues to do more to promote their accepted papers?

A conference could systematically collect content for social media from the authors for promoting the paper after it is accepted. Some journals already do that. On one hand, authors are more likely to better present their own work (more details, more engaging), but on the other hand if they are choosing to not self promote on social media –would they be willing to create social media content for the conference/editor?

ARR experimented with a bot for posting their anonymous preprints, but it wasn’t very popular. Perhaps one pitfall was that it was one bot for all tracks. Perhaps more specialized bots per track, or hashtags could help to filter the automatically posted content, and then it would be more useful?

#### Conferences

Some interventions for increasing the popularity of work from, for instance, underrepresented communities could happen during conference:

- Oral sessions could have “spillover effects”: having even one famous author in the panel could make it more likely that more people would attend other talks in the same oral session.
- Bidding data could be used to estimate which papers are more likely to be very popular, and to ensure that they are not all crowded together in the program.

- Random promotion of poster talks: selecting a certain fraction of posters to give a spotlight presentation of their work during a social event, and studying the impact of that additional exposure.

The attendance to conference themselves is of course far from equitable, due to the unequal distribution of funding and visa restrictions. Hybrid conferences have not been successful [52], and many venues are going back to mostly-onsite, with virtual participants as second-class citizens. One option to incentivize virtual attendance by decoupling the virtual event from the on-site event (e.g., as done at ACM EC 2024), and by offering free registration to the former [21]. The participants who could not attend on-site due to visa denials should automatically be given the option to present in the next conference where they could attend in person.

The current big conferences with thousands of attendees are very intimidating by themselves. Hence another option could be to subdivide conferences in 500 max attendees groups, and live stream tutorials for cross-location interactions. This should help bring both the cost and CO<sub>2</sub> impact<sup>3</sup> down, but probably ups the organizational burden. Perhaps this could be organized as a distributed event at an international hotel chain, that has locations in many cities.

#### 4.2.5 Next steps

For the deanonymization issue, the group concluded with the recommendation for the new ACL peer review committee (see subsection 4.1) to start tracking the information about preprints and intended preprints at ARR, and to monitor them over time to see what effect the new policy has, as compared to the impact of preprints reported at ACL'23 [54].

For the issue with reviewer diversity, ARR has already made an effort to broaden reviewer pool via recruitment of the authors of papers published at ACL. The effect on reviewer pool diversity needs to be estimated, and further measures for recruiting reviewers from China need to be taken if necessary. Equity of workload will be addressed at ARR by introducing a compulsory review load for authors of submitted papers.

### 4.3 Working Group on Paper-Reviewer Matching

*Anna Rogers (IT University of Copenhagen, DK, arog@itu.dk)*

License  Creative Commons BY 4.0 International license  
© Anna Rogers

Paper-reviewer matching is a key element that has a lot of impact on the overall quality of peer review at large-scale conferences, and it is extremely important to get it right [51]. This working group had three meetings during the seminar. The discussion focused on the strategies of paper-reviewer matching and the lessons learned from the experience of the participants (as chairs and reviewers).

#### 4.3.1 Assignment Strategies

The assignment is usually formalized as a discrete optimization problem, given the considerations of maximum load, the data about the quality of the match between submission and candidate reviewers, constraints such as conflicts of interests, reviewer seniority, experience

<sup>3</sup> <https://gist.github.com/jacobeisenstein/ae0e13e270f3b00c9c2046b52297d018>

etc., and sometimes also prevention of strategic or dishonest behavior [20, 66, 6, 62, 28, 26, 10]. The discussion focused on the experience of AAAI [33] and ACL [54], as well as the theory conferences organized on HotCRP platform [29]. It became clear that in practice each platform relied on a unique set of complex constraints, and tuning these algorithms takes a lot of effort and expertise – and the result is difficult to evaluate.

The AAAI approach [33] was particularly complex and required a lot of variables to be set by hand. The group discussed the possibility of trying to learn the optimal parameters for such a system, but there is not enough data for this.

#### 4.3.2 Assignment Criteria

Determining what constraints should be used for the assignments is a crucial step. The most salient component is the affinity score between (also called ‘similarity score’) the candidate reviewer and the submission, which is typically computed on the basis of reviewer publication history [6, 76, 7, 44, 41]. A major challenge is that these techniques may fail to pick up the aspects of similarity that are actually relevant for the match (e.g., abstracts can be stylistically similar but dissimilar in research topics), and NLP techniques have much room for improvement here [65]. Accordingly, the reviewer, author and chair trust in such scores is currently low [67]. Assignments based on past research may also no longer be interesting for the reviewers.

Another issue is the lack of clarity about the goal of review. Should the reviewers be optimized so that they would be most qualified to evaluate the technical correctness/soundness of the submission, or its novelty, or clarity of writing, or reproducibility, or excitement/interest to the community? It is not necessarily the case that these criteria coincide and would produce the same best match. Still, conference reviewers are implicitly expected to perform all these different roles, even though they may not be equally qualified for all this. In journals editors can craft per-paper committees; how can we do that on scale in conferences?

Another major challenge in paper-reviewer matching is noise in reviewer data (e.g., due to name disambiguation issues, unmaintained scholar profiles).

#### 4.3.3 Bidding

Bidding is a process commonly used to directly elicit reviewer preferences. While that allows the reviewers to pick the papers they would be the most interested in, it has integrity risks (see subsection 4.6), and is quite laborious, which is why usually people only bid on a few papers shown at the top of the list [5, 13, 40]. Furthermore, nobody would like to bid on papers that look badly written or overall unpromising, but they still need to be reviewed. Finally, people often bid on what they want to learn about, not necessarily what they have expertise on.

There is potential for improving the bidding process by imputing bids: showing the reviewers not the full set of submissions, but an imputed subset to bid on. This could be done based on affinity scores, and so as to remove various potential conflicts of interest. (A similar approach was taken in [77] for the problem of collusion rings; see [24] for an evaluation and some pros and cons of it.) To help with deanonymization, the reviewers could be asked if they follow social media a lot, and if they do – they could be only shown non-preprinted submissions.

#### 4.3.4 Standard for Paper-Reviewer Scoring

A big problem for research on paper-reviewer matching is that each conference operates on its own format for all the data that is used as constraints in the optimization problem. Any candidate solution needs to be integrated into this specific system, and once that is done – there is no “ground truth”, so it is hard to tell which alternative is actually better [56].

It would stimulate research in this area if there was at least a unified interface for paper-reviewer matching algorithms, used by all major conference platforms. Then any new solution could be tested more easily.

### 4.3.5 Matching in Iterative Assignment Setting

In HotCRP conferences as well as more recently other conferences in AI [33], it is common to have a variable number of reviews, due to multiple rounds of review, or other reasons. There is an initial pass, with a smaller number of reviews, meant to quickly reject obvious rejects, or accept papers that are good with high confidence. Then the remaining papers are assigned more reviewers, and their goal is to consider what was not covered by the first two reviewers. Perhaps some automated analysis of initial reviews could be used to facilitate picking the new reviewers, and explaining to them why they were assigned.

### 4.3.6 Next steps

- *Standard interface*: for the research on paper-reviewer matching to gain more traction in the community and become a research problem rather than just a conference organization problem, we need to develop a standard interface for interacting with confidential paper-reviewer matching data, that would be supported by the major conference platforms (OpenReview, Softconf, HotCRP).
- *Imputing bids*: The integrity, quality, and overall user experience of the bidding process can be enhanced by assigning a specific set of papers for each reviewer to evaluate and input their bidding information, and imputing the rest from it.

## 4.4 Working Group on Incentives in Peer Review

Nihar Shah (Carnegie Mellon University, Pittsburgh, US, nihars@cs.cmu.edu)

License © Creative Commons BY 4.0 International license  
© Nihar Shah

Participants in this working group discussed two types of incentives: incentives for reviews for providing (high quality) reviews, and incentives for authors to submit only high-quality work. We discuss these two types of incentives in the following two subsections.

Although we present them separately for clarity of exposition, the discussions also captured some relations between the two. For instance, higher quality of reviews may reduce the number of submissions, if authors realize that low-quality submissions have little chance of getting in. Conversely, if authors were to curtail the number of submissions, the load on reviewers would reduce, and the quality of review may go up.

### 4.4.1 Incentives for Reviewers

Participants in this working group first discussed the various current incentives for reviewing:

- Prestige service roles, e.g., being area chairs or senior area chairs.
- People can mention it on their CVs.
- Building an informal social credit with colleagues that will lead to invitations (e.g., seminars) and positive response to future service requests.

- Listing and acknowledging of reviewers in the proceedings.
- Reviewer awards.
- Conference policies forcing eligible authors to also sign up to review.
- Additional motivations include keeping up with the literature, gatekeeping or influencing directions of their field, and improving scientific report quality [42].

Participants also discussed reasons why reviewers may not want to review [42]:

- Time may be better spent elsewhere.
- Too stressful.
- Assigned papers are not interesting.
- No recognition for the work.

Based on these observations, number of potential directions towards addressing this problem were then proposed and discussed:

- Assign reviewers in a manner that area chairs know the reviewers professionally, so that there is more accountability from reviewers.
- Collect (and possibly make public) reviewer performance over time.
- Overcome challenges in measuring review quality [19].
- If measuring review quality is hard, at least incentivize other measurable desiderata like completing the reviews in time or signing up for reviewing.
- Nudge people appropriately, e.g., via personalized reminders with actual names and the link to the paper they agreed to review, or personalized thank you emails for their service.
- Develop incentive structures where people who provide good reviews will have to review less.

#### 4.4.2 Incentives for Authors

As for the incentives for authors, a key challenge discussed is the high prestige often attributed solely to the act of publishing a paper. Additionally, various organizations and governments provide monetary and other forms of incentives for publishing an increased number of papers. It has been observed that *“introduction of incentives by a country is associated with an increase in submissions by the country; the relation is particularly strong between cash bonuses and submissions”* [14]. Therefore, there are both implicit and explicit pressures on authors to publish more frequently.

In response, some academic conferences have implemented measures to discourage excessive submissions. For example, the International Conference on Learning Representations (ICLR) publicly posts all submitted papers, including those that are rejected, along with their peer reviews. This transparency is intended to deter submissions of lower quality, as the public availability of reviews can dissuade authors from submitting subpar work. Furthermore, several conferences now require authors to include reviews from any prior rejections when resubmitting papers, which are then passed to the new reviewers. This practice aims to prevent repeated submissions of low-quality papers, as a previous rejection could negatively influence subsequent reviews [64]. Despite these measures, the effectiveness of such policies in reducing the number of submissions or improving submission quality has not been comprehensively documented or measured. Thus, developing reliable methods to assess the causal impact of these policies on submission behaviors remains a critical open issue.

### 4.4.3 Next steps

Concrete next steps comprise:

- Developing more fair and accurate ways of measuring review quality, for instance, overcoming the problems discovered by the experiments in [19].
- Developing economic models capturing differences between venues where the peer-review quality is perceived to be good versus venues where it is perceived to be poor.
- Designing protocols for better longitudinal compilation of reviewer performance.

### 4.4.4 Open problems

There are two primary types of incentives that need thoughtful design to support the peer-review process. First, reviewing is often a voluntary task, and creating incentives that promote high-quality reviews is a challenge. While some strategies focus on increasing the volume of reviews, such as mandating that eligible authors must participate in peer reviewing, these do not necessarily guarantee quality. More targeted approaches aim to enhance review quality, like awards for outstanding reviewers such as those at the NeurIPS conference, as well as more theoretical approaches using game theory [78, 60, 69]. However, these approaches face several hurdles, including the gap between theoretical assumptions and real-world scenarios, unclear effects of these policies on reviewer motivation, and difficulties in accurately assessing the quality of reviews [19].

The second type of incentive concerns the authors. With the high numbers of submissions to conferences, ensuring thorough and high-quality peer reviews is becoming increasingly difficult. For authors, the cost of submitting papers is relatively low: acceptance means inclusion in a prestigious conference; rejection has minimal consequences. This low-risk environment coupled with noise in the process encourages the submission of papers that may even be of unsuitable quality, which might still be accepted. Additionally, some institutions and governments reward researchers for having papers accepted at these conferences, further incentivizing high submission rates and compounding the challenges in the review process. It is thus crucial to explore incentive systems that motivate authors to submit only high-quality work. Initiatives like ICLR policy of making all submissions public can deter low-quality submissions by adding repercussions for rejection. Meanwhile, platforms like TMLR accept all competent and relevant submissions, which could diminish the prestige of mere acceptance. The effectiveness of these initiatives in maintaining high standards, however, remains to be evaluated.

## 4.5 Working Group on Natural Language Processing (NLP) for Peer Review

*Nihar Shah (Carnegie Mellon University, Pittsburgh, US, nihars@cs.cmu.edu)*

License © Creative Commons BY 4.0 International license  
© Nihar Shah

### 4.5.1 Overview

Natural Language Processing (NLP) has significant potential to improve the peer review process. There are various problems in peer review for which current works rely primarily on numerical data such as ratings provided by reviewers. These include challenges like miscalibration [17, 55, 73], subjectivity [43], elicitation [45, 37], and author-identity bias [68, 3].

Although these works focus on the numeric assessments, some of the main components of peer review—submissions and feedback—are text-based, making NLP a fitting tool for analysis and improvement.

Experiments in peer review also focus on quantifiable outcomes such as ratings and acceptance decisions [31, 2, 61, 63, 50, 64, 35]. Despite the numeric focus, the textual nature of the data suggests that NLP can offer substantial contributions to these areas. While some efforts have been made [39, 15, 48], much potential remains for NLP to further address these challenges.

Moreover, NLP can help tackle issues that are currently unaddressed, such as identifying unsubstantiated criticisms in reviews or pinpointing deficiencies in papers. However, developing these NLP methods must also consider the safety and fairness of their applications, and how these methods are evaluated and measured. This area of research is gaining increasing popularity [12, 36, 30, 9, 27], and all these considerations will be explored in detail in the forthcoming paper titled “What Can Natural Language Processing Do for Peer Review?”[80] emerging from this Dagstuhl Seminar.

#### 4.5.2 Whitepaper

This working group quickly converged to the understanding that:

- There is a huge opportunity for improving peer review via latest advancements in natural language processing.
- There are also as many challenges as doing it in a safe, fair, and accurate manner, as well as in evaluating the outcomes.
- It is thus important to convey this message to researchers in NLP, ML and related communities, and also make it as easy as possible to step into this research application domain.
- A suitable means of doing so is to write a position paper, accompanied with a repository containing various datasets pertaining to peer review.

With this motivation, the remainder of the sessions focused on planning, organizing, and beginning to write the position paper and compile the datasets. The paper touches upon the following topics:

- Background of the peer-review process.
- Assistance before the review process.
  - Preparing the manuscripts.
    - \* Writing assistance for authors.
    - \* Helping authors form metadata such as keywords and TL;DRs.
    - \* Initial screening of manuscripts for basic checks.
  - Reviewer-paper matching.
    - \* Computing similarities between reviewers and submitted papers.
    - \* Finding conflicts of interest.
    - \* Reducing strategic behavior.
- Assistance during the review process.
  - Evaluating certain aspects of the manuscript.
  - Helping write the review.
  - Discussions with authors and reviewers.
- Assistance after the review process.
  - Helping with the meta review.
  - Final decisions.

- Camera-ready submissions.
- Post-conference analysis.
- Data, privacy, and legal aspects.
- Measurements and experimentations.
- Ethics.

### 4.5.3 Next steps

The whitepaper manuscript is available at [80] and the associated repository is available at <https://github.com/OAfzal/nlp-for-peer-review>.

## 4.6 Working Group on Integrity in Peer Review

*Nihar Shah (Carnegie Mellon University, Pittsburgh, US, nihars@cs.cmu.edu)*

*Iryna Gurevych (TU Darmstadt, DE, iryna.gurevych@tu-darmstadt.de)*

License © Creative Commons BY 4.0 International license  
© Nihar Shah, Iryna Gurevych

This working group focused on issues undermining the integrity of peer review processes. Initially, participants explored a range of challenges affecting peer review integrity (see, e.g., [57, Section 4]). As discussions progressed, the consensus emerged that collusion rings and AI tools to support the integrity of the peer review process represented the most critical issues. Consequently, one subgroup dedicated the rest of their session to this specific problem. Another subgroup worked on developing a roadmap for AI tools and protocols for collecting peer review data as an enabling factor for the envisaged tools.

### 4.6.1 Collusion Rings

The allure of being published in prestigious conferences can sometimes encourage unethical behaviors among participants. One concerning trend that has gained attention is the formation of collusion rings [70, 34]. In these scenarios, groups of researchers manipulate the peer review system to review each other's papers. They then provide favorable evaluations, often disregarding the true merit of the work.

In recent years, program chairs have devoted considerable time and effort to addressing the issue of collusion rings. Tackling this challenge is essential as it directly undermines the integrity and fairness of the peer review process. Alongside this, there is a growing concern over bullying and abuse of power within the academic community. Participants reported instances where senior researchers, including some area chairs at conferences, have pressured junior colleagues to engage in these unethical activities.

One approach to addressing this issue is through technical means: developing algorithmic methods designed to detect or prevent collusion rings. Much research has already been conducted in this area [26, 77, 33, 4, 33, 24, 25, 23]. Further discussions in the working group focused on several potential strategies:

- Introducing additional conditions, such as requiring more reviewer bids. While every intervention has associated costs, it is crucial to consider potential drawbacks. For instance, although the ARR system does not involve bidding, collusion could still occur. Currently, assignments are automated with the option for Area Editors (AEs) to make changes. However, this could lead to issues if an AE is compromised.



- Implementing policies to ensure that the same person does not repeatedly review another individual's papers over multiple years. While such interventions encourages diversity, their disadvantages must also be weighed carefully in terms of reducing expertise of assigned reviewers.
- Approaching the problem as a network issue, where more distant social links might serve as indicators of conflicts of interest. Further analyzing patterns of paper submissions and reviews, identifying discrepancies between poor-quality papers and positive reviews. Although not necessarily indicative of malicious collusion, such patterns could still pose significant problems.
- Conducting automated assessments of papers and, in cases of high disagreement between reviewers, assigning an additional reviewer. This approach must be handled carefully to avoid a high rate of false positives.
- Developing models to quantify the likelihood that certain behaviors are due to chance. This requires careful evaluation to ensure accuracy and effectiveness.

A second, human-centric approach was explored to address this problem, with several strategies proposed:

- *Whistleblower Support*: Participants shared knowledge about instances where collusion rings were exposed through shared communications within chat groups. This highlights the need for robust mechanisms that enable whistleblowers to safely report unethical behaviors.
- *Policy Development Board*: Conferences should establish a board dedicated to policy creation. This board would be responsible for developing protocols and guidelines to manage reports of misconduct effectively, including those of handling whistleblower reports.
- *Guidelines for Program Chairs*: Since program chairs of conferences change every year, it will be useful to develop guidelines for program chairs on what sorts of prevention measures are there for collusion rings.
- *Inter-Conference Data Sharing*: Some preventive measures may require data from multiple conferences to detect recurring patterns of misconduct. However, this poses legal and technical challenges, such as difficulties in transferring data between different conferences or iterations due to current platform limitations. Addressing these issues will be crucial for effective prevention.
- *Education and Training*: Implementing educational programs to inform researchers about unethical practices and their severe consequences can serve as a preventive measure. This approach aims to cultivate a culture of integrity and transparency within the academic community.

**Next steps:** Collusion rings may be addressed either by developing methods to detect them, or by preventing them (during the reviewer assignment phase itself). Although there is ongoing research on these questions [26, 33, 77, 4, 24, 23, 25], finding solutions involves significant trade-offs, and this problem largely remains open [24, 23]. Effective strategies to combat these unethical practices are crucial for maintaining the integrity of the peer-review process. There are a number of subsequent steps that are underway following the Dagstuhl Seminar. The first is to make more researchers aware of this problem and encouraging them to work on technical solutions using the tools at their disposal. For instance, NLP tools have not been used so far to combat against collusion rings, and we argue for doing so in the forthcoming paper on NLP for peer review (discussed in the NLP section of this report). We are also developing guidelines for program chairs and pushing along policies for some sharing

of data across conferences. We envisage further outcomes to emanate in due time based on the aforementioned directions conceived at the Dagstuhl Seminar.

#### 4.6.2 AI Tools

A major bottleneck for developing AI tools to support the integrity of peer review is the lack of data for training and evaluation in this sensitive domain subject to privacy and data protection. Existing large scale data collections suffer from the low-response rate. Therefore, the group participants were interested in helping with the data collection.

This involved tasks such as creating an empirical protocol for data collection, achieving the ethical and legal clearance by writing a proposal for the ACL ethics committee, etc., defining what kind of data to extract, for what purposes and how to implement this technically. Further questions that were covered in the discussion were data ownership, distribution and management, allowing for retraction of consent, backward compatibility with the previous data collection workflow, cross-community data collection, incentivizing data donation and public outreach through blog posts, social media, etc.

At the time of writing, most of the discussed topics have already been addressed by the seminar participants. The result is documented in this website: <https://arr-data.aclweb.org>. The data collection protocol has been successfully implemented and the data collection is now ongoing within the ACL Rolling Review (ARR) platform.

#### References

- 1 David M. Allen and James W. Howell, editors. *Groupthink in Science: Greed, Pathological Altruism, Ideology, Competition, and Culture*. Springer International Publishing, Cham, 2020.
- 2 Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. The NeurIPS 2021 consistency experiment. <https://blog.neurips.cc/2021/12/08/the-neurips-2021-consistency-experiment/>, 2021. Online; accessed 18-April-2024.
- 3 Rebecca M Blank. The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review. *The American Economic Review*, pages 1041–1067, 1991.
- 4 Niclas Boehmer, Robert Bredereck, and André Nichterlein. Combating collusion rings is hard but possible. *arXiv preprint arXiv:2112.08444*, 2021.
- 5 Guillaume Cabanac and Thomas Preuss. Capitalizing on order effects in the bids of peer-reviewed conferences to secure reviews by expert referees. *Journal of the Association for Information Science and Technology*, 64(2):405–415, 2013.
- 6 Laurent Charlin and Richard Zemel. The Toronto Paper Matching System: An automated paper-reviewer assignment system. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, Atlanta, Georgia, USA, May 2013. Journal of Machine Learning Research Workshop and Conference Proceedings.
- 7 Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, 2020.
- 8 Mike D’Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. MARG: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*, 2024.
- 9 Mike D’Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. ARIES: A corpus of scientific paper edits made in response to peer reviews. *arXiv preprint arXiv:2306.12587*, 2023.

- 10 Komal Dhull, Steven Jecmen, Pravesh Kothari, and Nihar B Shah. Strategyproofing peer assessment via partitioning: The price in terms of evaluators’ expertise. In *HCOMP*, 2022.
- 11 Nils Dycke, Iliia Kuznetsov, and Iryna Gurevych. Yes-yes-yes: Proactive data collection for ACL rolling review and beyond. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 300–318, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- 12 Nils Dycke, Iliia Kuznetsov, and Iryna Gurevych. NLPeer: A unified resource for the computational study of peer review. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada, July 2023. Association for Computational Linguistics.
- 13 Tanner Fiez, Nihar Shah, and Lillian Ratliff. A SUPER\* algorithm to optimize paper bidding in peer review. In *Conference on Uncertainty in Artificial Intelligence*, pages 580–589. Proceedings of Machine Learning Research, 2020.
- 14 Chiara Franzoni, Giuseppe Scellato, and Paula Stephan. Changing incentives to publish. *Science*, 333(6043):702–703, 2011.
- 15 Yang Gao, Steffen Eger, Iliia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. Does my rebuttal matter? Insights from a major NLP conference. In Jill Burstein, Christy Doran, and Tamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1274–1290, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- 16 J. A. Garcia, Rosa Rodriguez-Sánchez, and J. Fdez-Valdivia. Confirmatory bias in peer review. *Scientometrics*, 123(1):517–533, April 2020.
- 17 H. Ge, M. Welling, and Z. Ghahramani. A Bayesian model for calibrating conference review scores. Manuscript, 2013. Available online <http://mlg.eng.cam.ac.uk/hong/unpublished/nips-review-model.pdf> Last accessed: April 4, 2021.
- 18 Alexander Goldberg, Giulia Fanti, and Nihar B Shah. Batching of tasks by users of pseudonymous forums: Anonymity compromise and protection. In *ACM SIGMETRICS*, 2023.
- 19 Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B Shah. Peer reviews of peer reviews: A randomized controlled trial and other experiments. *arXiv preprint arXiv:2311.09497*, 2023.
- 20 Judy Goldsmith and Robert H. Sloan. The AI conference paper assignment problem. *WS-07-10:53–57*, 12 2007.
- 21 Jason Harline. The Dissemination Game: Incentives of In-Person vs Virtual Participation – Communications of the ACM, July 2023.
- 22 Jürgen Huber, Sabiou Inoua, Rudolf Kerschbamer, Christian König-Kersting, Stefan Palan, and Vernon L. Smith. Nobel and novice: Author prominence affects peer review. *Proceedings of the National Academy of Sciences*, 119(41):e2205779119, October 2022.
- 23 Steven Jecmen, Nihar B Shah, Fei Fang, and Leman Akoglu. On the detection of reviewer-author collusion rings from paper bidding. *arXiv preprint arXiv:2402.07860*, 2024.
- 24 Steven Jecmen, Nihar B Shah, Fei Fang, and Vincent Conitzer. Tradeoffs in preventing manipulation in paper bidding for reviewer assignment. *arXiv preprint arXiv:2207.11315*, 2022.
- 25 Steven Jecmen, Minji Yoon, Vincent Conitzer, Nihar B Shah, and Fei Fang. A dataset on malicious paper bidding in peer review. In *Proceedings of the ACM Web Conference 2023*, pages 3816–3826, 2023.

- 26 Steven Jecmen, Hanrui Zhang, Ryan Liu, Nihar Shah, Vincent Conitzer, and Fei Fang. Mitigating manipulation in peer review via randomized reviewer assignments. *Advances in Neural Information Processing Systems*, 33:12533–12545, 2020.
- 27 Neha Kennard, Tim O’Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. DISAPERRE: A dataset for discourse structure in peer review discussions. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1234–1249, Seattle, United States, July 2022. Association for Computational Linguistics.
- 28 Ari Kobren, Barna Saha, and Andrew McCallum. Paper matching with local fairness constraints. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1247–1257, 2019.
- 29 Eddie Kohler. HotCRP conference management software, 2013.
- 30 Iliia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. Revise and Resubmit: An intertextual model of text-based collaboration in peer review. *Computational Linguistics*, 48(4):949–986, 12 2022.
- 31 N. Lawrence and C. Cortes. The NIPS Experiment. <http://inverseprobability.com/2014/12/16/the-nips-experiment>, 2014. Online; accessed 17-April-2024.
- 32 Carole J Lee. Commensuration bias in peer review. *Philosophy of Science*, 82(5):1272–1283, 2015.
- 33 Kevin Leyton-Brown, Yatin Nandwani, Hedayat Zarkoob, Chris Cameron, Neil Newman, Dinesh Raghu, et al. Matching papers and reviewers at large conferences. *Artificial Intelligence*, 2024.
- 34 Michael L Littman. Collusion rings threaten the integrity of computer science research. *Communications of the ACM*, 64(6):43–44, 2021.
- 35 Ryan Liu, Steven Jecmen, Vincent Conitzer, Fei Fang, and Nihar B Shah. Testing for reviewer anchoring in peer review: A randomized controlled trial. *arXiv preprint arXiv:2307.05443*, 2023.
- 36 Ryan Liu and Nihar B. Shah. ReviewerGPT? An exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*, 2023.
- 37 Yusha Liu, Yichong Xu, Nihar B Shah, and Aarti Singh. Integrating rankings into quantized scores in peer review. *arXiv preprint arXiv:2204.03505*, 2022.
- 38 Mario Malički and Bahar Mehmani. Structured peer review: Pilot results from 23 Elsevier journals. *bioRxiv preprint bioRxiv:10.1101/2024.02.01.578440*, February 2024.
- 39 Emaad Manzoor and Nihar B Shah. Uncovering latent biases in text: Method and application to peer review. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4767–4775, 2021.
- 40 Reshef Meir, Jérôme Lang, Julien Lesca, Nicholas Mattei, and Natan Kaminsky. A market-inspired bidding scheme for peer review paper assignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4776–4784, 2021.
- 41 Sheshera Mysore, Mahmood Jasim, Andrew McCallum, and Hamed Zamani. Editable user profiles for controllable text recommendation. *arXiv preprint arXiv:2304.04250*, 2023.
- 42 Syavash Nobarany, Kellogg S Booth, and Gary Hsieh. What motivates people to review articles? the case of the human-computer interaction community. *Journal of the Association for Information Science and Technology*, 67(6):1358–1371, 2016.
- 43 Ritesh Noothigattu, Nihar Shah, and Ariel Procaccia. Loss functions, axioms, and peer review. *Journal of Artificial Intelligence Research*, 70:1481–1515, 2021.
- 44 Justin Payan and Yair Zick. I will have order! Optimizing orders for fair reviewer assignment. *arXiv preprint arXiv:2108.02126*, 2021.

- 45 Michael Pearce and Elena A Erosheva. A unified statistical learning model for rankings and scores with application to grant panel review. *arXiv preprint arXiv:2201.02539*, 2022.
- 46 Douglas P Peters and Stephen J Ceci. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, 5(2):187–195, 1982.
- 47 Sukannya Purkayastha, Anne Lauscher, and Iryna Gurevych. Exploring jiu-jitsu argumentation for writing peer review rebuttals. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14479–14495. Association for Computational Linguistics, 2023.
- 48 Charvi Rastogi, Xiangchen Song, Zhijing Jin, Ivan Stelmakh, Hal Daumé III, Kun Zhang, and Nihar B Shah. A randomized controlled trial on anonymizing reviewers to each other in peer review discussions. *arXiv preprint arXiv:2403.01015*, 2024.
- 49 Charvi Rastogi, Ivan Stelmakh, Alina Beygelzimer, Yann N Dauphin, Percy Liang, Jennifer Wortman Vaughan, Zhenyu Xue, Hal Daumé III, Emma Pierson, and Nihar B Shah. How do authors’ perceptions of their papers compare with co-authors’ perceptions and peer-review decisions? *arXiv preprint arXiv:2211.12966*, 2022.
- 50 Charvi Rastogi, Ivan Stelmakh, Xinwei Shen, Marina Meila, Federico Echenique, Shuchi Chawla, and Nihar Shah. To ArXiv or not to ArXiv: A study quantifying pros and cons of posting preprints online. *arXiv preprint arXiv:2203.17259*, 2022.
- 51 Marko A Rodriguez, Johan Bollen, and Herbert Van de Sompel. Mapping the bid behavior of conference referees. *Journal of Informetrics*, 1(1):68–82, 2007.
- 52 Anna Rogers. Field Notes on Hybrid Conferences (EMNLP 2021), November 2021.
- 53 Anna Rogers and Isabelle Augenstein. What can we do to improve peer review in NLP? In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262. Association for Computational Linguistics, November 2020.
- 54 Anna Rogers, Marzena Karpinska, Jordan Boyd-Graber, and Naoaki Okazaki. Program chairs’ report on peer review at ACL 2023. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40–75, Toronto, Canada, July 2023. Association for Computational Linguistics.
- 55 Magnus Roos, Jörg Rothe, Joachim Rudolph, Björn Scheuermann, and Dietrich Stoyan. A statistical approach to calibrating the scores of biased reviewers: The linear vs. the nonlinear model. In *Multidisciplinary Workshop on Advances in Preference Handling*, 2012.
- 56 Martin Saveski, Steven Jecmen, Nihar Shah, and Johan Ugander. Counterfactual evaluation of peer-review assignment policies. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 58765–58786. Curran Associates, Inc., 2023.
- 57 Nihar B Shah. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87, 2022.
- 58 Inna Smirnova, Daniel M. Romero, and Misha Teplitskiy. The bias-reducing effect of voluntary anonymization of authors’ identities: Evidence from peer review, January 2023.
- 59 Inna Smirnova, Daniel M Romero, and Misha Teplitskiy. The bias-reducing effect of voluntary anonymization of authors’ identities: Evidence from peer review. *Available at SSRN 4190623*, 2023.
- 60 Siddarth Srinivasan and Jamie Morgenstern. Auctions and prediction markets for scientific peer review. *arXiv preprint arXiv:2109.00923*, 2021.
- 61 Ivan Stelmakh, Charvi Rastogi, Nihar B Shah, Aarti Singh, and Hal Daumé III. A large scale randomized controlled trial on herding in peer-review discussions. *arXiv preprint arXiv:2011.15083*, 2020.



- 62 Ivan Stelmakh, Nihar Shah, and Aarti Singh. PeerReview4All: Fair and accurate reviewer assignment in peer review. *Journal of Machine Learning Research*, 22(163):1–66, 2021.
- 63 Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4785–4793, 2021.
- 64 Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. Prior and prejudice: The novice reviewers’ bias against resubmissions in conference peer review. volume 5, pages 1–17. ACM New York, NY, USA, 2021.
- 65 Ivan Stelmakh, John Wieting, Graham Neubig, and Nihar B. Shah. A gold standard dataset for the reviewer assignment problem. *arXiv preprint arXiv:2303.16750*, 2023.
- 66 Camillo J Taylor. On the optimal assignment of conference papers to reviewers. 2008.
- 67 Terne Thorn Jakobsen and Anna Rogers. What factors should paper-reviewer assignments rely on? community perspectives on issues and ideals in conference peer-review. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4810–4823, Seattle, United States, July 2022. Association for Computational Linguistics.
- 68 Andrew Tomkins, Min Zhang, and William D. Heavlin. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713, 2017.
- 69 Alexander Ugarov. Peer prediction for peer review: Designing a marketplace for ideas. *arXiv:2303.16855*, 2023.
- 70 T. N. Vijaykumar. Potential organized fraud in ACM/IEEE computer architecture conferences. <https://medium.com/@tnvijayk/potential-organized-fraud-in-acm-ieee-computer-architecture-conferences-ccd61169370d>, 2020. Online; accessed 17-April-2024.
- 71 Jian Wang, Reinhilde Veugelers, and Paula Stephan. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436, October 2017.
- 72 Jingyan Wang and Ashwin Pananjady. Modeling and correcting bias in sequential evaluation. In *Conference on Economics and Computation, (EC)*, 2023.
- 73 Jingyan Wang and Nihar B Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 864–872, 2019.
- 74 Jingyan Wang, Ivan Stelmakh, Yuting Wei, and Nihar Shah. Debiasing evaluations that are biased by evaluations. In *AAAI Conference on Artificial Intelligence*, 2021.
- 75 Iain Xie Weissburg, Mehira Arora, Xinyi Wang, Liangming Pan, and William Yang Wang. Tweets to Citations: Unveiling the Impact of Social Media Influencers on AI Research Visibility, March 2024.
- 76 John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. Simple and effective paraphrastic similarity from parallel translations. In *ACL*, pages 4602–4608, Florence, Italy, July 2019.
- 77 Ruihan Wu, Chuan Guo, Felix Wu, Rahul Kidambi, Laurens van der Maaten, and Kilian Weinberger. Making paper reviewing robust to bid manipulation attacks. In *ICML*, 2021.
- 78 Yuanzhang Xiao, Florian Dörfler, and Mihaela Van Der Schaar. Incentive design in peer review: Rating and repeated endogenous matching. *IEEE Transactions on Network Science and Engineering*, 6(4):898–908, 2018.
- 79 Dennis Zyska, Nils Dycke, Jan Buchmann, Ilia Kuznetsov, and Iryna Gurevych. CARE: Collaborative AI-assisted reading environment. In Danushka Bollegala, Ruihong Huang, and Alan Ritter, editors, *Proceedings of the 61st Annual Meeting of the Association for*

*Computational Linguistics (Volume 3: System Demonstrations)*, pages 291–303, Toronto, Canada, July 2023. Association for Computational Linguistics.

- 80 Iliia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, Jingyan Wang, Xiaodan Zhu, Anna Rogers, Nihar B. Shah, Iryna Gurevych. What Can Natural Language Processing Do for Peer Review?”, CoRR, Vol. abs/2405.06563, 2024.

## Participants

- Osama Mohammed Afzal  
MBZUAI – Abu Dhabi, AE
- Koen Dercksen  
Radboud University  
Nijmegen, NL
- Nils Dycke  
TU Darmstadt, DE
- Alexander Goldberg  
Carnegie Mellon University –  
Pittsburgh, US
- Iryna Gurevych  
TU Darmstadt, DE
- Jason Hartline  
Northwestern University –  
Evanston, US
- Tom Hope  
The Hebrew University of  
Jerusalem, IL
- Dirk Hovy  
Bocconi University – Milan, IT
- Eddie Kohler  
Harvard University – Allston, US
- Jonathan Kummerfeld  
The University of Sydney, AU
- Ilia Kuznetsov  
TU Darmstadt, DE
- Anne Lauscher  
Universität Hamburg, DE
- Kevin Leyton-Brown  
University of British Columbia –  
Vancouver, CA
- Sheng Lu  
TU Darmstadt, DE
- Dorsa Majdi  
Sharif University of Technology –  
Tehran, IR
- Mausam  
Indian Institute of Technology –  
New Delhi, IN
- Bahar Mehmani  
Elsevier BV – Amsterdam, NL
- Margot Mieskes  
Hochschule Darmstadt, DE
- Aurélie Névéol  
CNRS – Orsay, FR
- Danish Pruthi  
Indian Institute of Science –  
Bangalore, IN
- Lizhen Qu  
Monash University –  
Clayton, AU
- Anna Rogers  
IT University of  
Copenhagen, DK
- Roy Schwartz  
The Hebrew University of  
Jerusalem, IL
- Nihar Shah  
Carnegie Mellon University –  
Pittsburgh, US
- Noah A. Smith  
University of Washington –  
Seattle, US
- Tamar Solorio  
MBZUAI – Abu Dhabi, AE
- Jingyan Wang  
Georgia Institute of Technology –  
Atlanta, US
- Xiaodan Zhu  
Queen's University –  
Kingston, CA

