Towards Fair, Equitable, and Efficient Peer Review

Ivan Stelmakh
Machine Leaning Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15217
stiv@cs.cmu.edu

Abstract

Peer review is the backbone of academia. The rapid growth of the number of submissions to leading publication venues has identified a need for automation of some parts of the peer-review pipeline and nowadays human referees are required to interact with various interfaces and technologies in this process. However, there exists evidence that if such interactions are not carefully designed, they can exacerbate various problems related to fairness and efficiency of the process. In my research, I aim to design a Human-AI collaboration pipeline in peer review to mitigate these issues and ensure that science progresses in a fair, equitable, and efficient manner.

Despite peer review being the primary mechanism of science dissemination for decades, the rapid growth of the number of submissions to leading AI and ML conferences has challenged its sustainability in two ways:

- It has brought up a call for automated tools to assist human decision-makers.
- It has amplified the shortcomings of the peer-review procedure, making them more visible to the community and stressing the importance of research on peer review.

These issues motivate my thesis research and *I am passionate about working at the intersection of machine learning, operations research, social choice theory, and human-computer interaction, to understand and develop a principled approach towards scientific peer review.* Specifically, I believe that a carefully designed Human-AI collaboration is crucial for sustainability of peer review and in my work I aim at designing tools to support this collaboration.

My research touches both algorithmic and human sides of the Human-AI collaboration and in the sequel I first describe my projects on supporting each of these sides. I then outline a direction for future work on bringing these sides to a closer interaction with a goal of improving the peer-review process.

On a higher level, my work comprises novel theoretical and empirical contributions: I aim to design practical algorithms that are supported by strong theoretical guarantees and are evaluated in a carefully designed real-world experiments. The preliminary results I discuss below have already had a considerable impact in practice with some tools deployed in ICML 2020, and this inspires me to continue my work towards fair, equitable and efficient peer review.

Algorithmic Side. Past research in social science indicates that unfairness of the peer-review process may have far-reaching consequences both on a development of research areas and on career trajectories of individual researchers. Therefore, *my work on the algorithmic side is twofold: first, I aim to ensure that the algorithms used to automate peer review are themselves fair. Second, I aim at designing algorithms that help conference organizers to promote fairness*.

Fairness for Algorithms: The most automated part of the review process is the assignment of submissions to referees and most of the of the top AI and ML conferences rely on a simple and efficient matching algorithm developed by Charlin and Zemel (2013). Simultaneously, assignment is of the utmost importance: one cannot expect good reviews for papers that are assigned to unsuitable reviewers.

In our past work (Stelmakh, Shah, and Singh 2018) we demonstrate that the state-of-the-art algorithm used by NeurIPS and ICML does not necessarily lead to a fair assignment, discriminating against some papers. More importantly, we design a novel assignment algorithm with provable guarantees on the fairness of the assignment that ensures that no paper is discriminated against to improve the assignment of more lucky counterparts.

In addition to strong fairness guarantees, our algorithm is also optimal in terms of the accuracy of final decisions under a popular statistical model, that is, our algorithm theoretically outperforms the state-of-the-art algorithm both in terms of fairness and statistical accuracy. These guarantees are corroborated by an extensive empirical evaluation: in particular, our algorithm was tested and eventually deployed in the assignment of the ICML 2020 conference, improving the fairness by 15-30% while not trading off the conventional measure of the assignment quality.

Algorithms for Fairness: While we can prove that algorithms employed to automate peer review satisfy the requirement of fairness, ensuring fairness of decisions made by humans is a more challenging task. An important direction that I am interested in is a use of algorithms to perform statistical testing for fairness and impartiality of final decisions.

In our work (Stelmakh, Shah, and Singh 2019), we made the progress on this problem by contributing to the long-standing debate on the fairness of the decisions in single-blind peer review. In that, we design a novel semi-randomized experimental procedure that allows to test for

biases in single-blind peer review while requiring only a moderate amount of additional efforts from organizers. We also show that past approaches, including a remarkable and impactful work by Tomkins, Zhang, and Heavlin (2017), rely on the strong and non-testable assumptions and are at risk of being not reliable when using real data.

Continuing the work in this direction, we also designed a test for detection of strategic manipulations of reviewers (Stelmakh, Shah, and Singh 2020). Given that stakes in peer review are high, reviewers (who often have their own works in submission) may be incentivized to rate other works strategically in attempt to improve the final outcome of their own submissions and our test aims at detecting such manipulations. Conceptually, our work complements the line of research on impartial aggregation algorithms as it allows stakeholders to evaluate the presence of strategic behaviour in the system and take an informed decision on whether they should employ an impartial mechanism.

Human Side. Human agents (reviewers) play a key role in peer review. However, a multitude of works on human decision-making and psychology document a susceptibility of human judgements to various cognitive distortions. Therefore, *it is extremely important to ensure that the peer-review process is carefully designed to minimize the impact of the undesirable effects and to foster high-quality reviews*. To this end, in collaboration with the ICML 2020 program committee, I have designed and conducted several human-subject experiments to evaluate novel peer review practices.

Mitigating Cognitive Biases: Human agents are known to suffer from various cognitive biases (e.g., primacy and hindsight biases) and these biases can manifest in various places of the review process; in two recent projects, we have designed and conducted experiments to evaluate the impact of such biases on the final outcome of submissions. First, we target the initial stage of the review process and evaluate the impact of the signal that a paper is a resubmission from some past venue (Stelmakh et al. 2020c). Second, we consider the discussion stage (Stelmakh et al. 2020a) and evaluate the impact of the choice of the discussion initiator. We plan to publish the reports on these trials by early 2021 and the results will help conference organizers to make informed decisions when designing the peer-review process.

Enlarging the Pool of Qualified Reviewers: A surge in the number of submissions received by leading conferences has increased the burden on the pool of qualified reviewers which is growing at a much slower rate. To address the issue, in our recent work (Stelmakh et al. 2020b) we design a procedure for (i) recruiting reviewers from the population not typically covered by major conferences and (ii) guiding them through the reviewing pipeline. In conjunction with the ICML conference, we recruited a small set of reviewers through our procedure and demonstrated that our mechanism allows for a principled enhancement of the reviewer pool and results in reviews of comparable and even superior quality as compared to the conventional pool of reviews.

Interaction Between Humans and Algorithms. Various idiosyncrasies of peer review make its automation very difficult — research papers are very complex and understanding of little nuances is beyond the capabilities of modern NLP algorithms. On the other hand, attempts to build a sustainable system where all the work is performed by humans also did not lead to success due to human subjectivity that results into different submissions being evaluated on different yardsticks, increasing the undesirable randomness of the process.

In my research, *I want to design a novel approach towards peer review that combines the best of both algorithmic and human worlds and promotes the Human-AI collaboration*. The key idea is to use human expertise and reading comprehension for hard parts of peer review which are beyond the strength of machines and then use algorithms to aggregate reviews in an objective manner, thereby decreasing the role of reviewers' subjectivity in final decisions. To achieve this ultimate goal, there are two concrete steps I plan to take:

Augment the role of reviewers: In collaboration with HCI researchers we plan to design a structured interface for reviewers to replace the rudimentary form currently used by most of the conferences. In that, we plan to change the role of a reviewer from a judge who evaluates the submission on several predefined criteria, to the expert who *creates* custom features that characterize strengths and weaknesses of papers and evaluate features proposed by others.

Objectivize subjectivity: The scale of the modern conferences makes it impossible for any single person to get a complete picture of all submissions. As the second step, I plan to design an automated algorithm that aggregates the features created by reviewers using our novel interface across all submissions and assists area and program chairs by providing a bigger picture of the conference. I hope to get preliminary results for this problem by the workshop date in early 2021.

In conclusion, I believe that my work benefits both individual researchers by ensuring equity and fairness, as well as society as a whole by ensuring a judicious progression of scientific research.

References

Charlin, L.; and Zemel, R. S. 2013. TPMS: An Automated Paper-Reviewer Assignment System. In *Proceedings of ICML 2013*.

Stelmakh, I.; Rastogi, C.; Shah, N. B.; Singh, A.; and Daumé III, H. 2020a. A Large Scale Randomized Controlled Trial on Herding in Peer-Review Discussions. *arXiv:2011.15083* .

Stelmakh, I.; Shah, N.; and Singh, A. 2019. On Testing for Biases in Peer Review. In *Proceedings of NeurIPS 32*, 5286–5296.

Stelmakh, I.; Shah, N. B.; and Singh, A. 2018. PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review. *arXiv:1806.06237* Short version in *ALT 2019*.

Stelmakh, I.; Shah, N. B.; and Singh, A. 2020. Catch Me if I Can: Detecting Strategic Behaviour in Peer Assessment. *arXiv:2010.04041* In AAAI 2021.

Stelmakh, I.; Shah, N. B.; Singh, A.; and Daumé III, H. 2020b. A Novice-Reviewer Experiment to Address Scarcity of Qualified Reviewers in Large Conferences. *arXiv:2011.15050* In AAAI 2021.

Stelmakh, I.; Shah, N. B.; Singh, A.; and Daumé III, H. 2020c. Prior and Prejudice: The Novice Reviewers' Bias against Resubmissions in Conference Peer Review. *arXiv:2011.14646* .

Tomkins, A.; Zhang, M.; and Heavlin, W. 2017. Reviewer Bias in Single- Versus Double-Blind Peer Review. *PNAS* 114(48).