Simple Measures to Improve Accountability in Peer Review for Large AI Conferences

Position Paper

Ulle Endriss
ILLC, University of Amsterdam
The Netherlands

ABSTRACT

I discuss a number of simple measures designed to improve the quality of peer review for large conferences, specifically regarding the accountability of reviewers for their crucial contributions. These measures seem to have had a positive impact for the two large AI conferences I chaired in the recent past, AAMAS-2021 (with 600+ submissions) and ECAI-2024 (with 2,300+ submissions).

INTRODUCTION

Complaints about the quality of reviews must be about as old as the practice of peer review itself, and the subjective impression that things are going downhill—especially during one's own lifetime—is part of human nature. But due to the explosive growth of conferences in AI in recent years, chances are that in the specific context of reviewing for such conferences at least some of these concerns are grounded in objective truths. Indeed, running the reviewing process for a large conference is significantly harder than doing so for a a more human-sized event. Even under the—in my opinion perfectly realistic—assumption that most reviewers will do a good job most of the time, dealing with the remaining cases—and even just *identifying* those cases—can become an almost insurmountable task for a PC chair once a conference exceeds a certain size.

In this position paper, I want to describe four very simple improvements to the reviewing process we used for AAMAS-2021 and ECAI-2024, two conferences for which I served as one of two PC chairs. The measures described, in one way or another, all make PC members more immediately *accountable* for the work they do—and committed to do when they joined the PC—by turning certain things we implicitly tend to assume good people will do anyway into explicit steps in the overall reviewing process. An important side effect (of three of the four measures) is that it becomes easier for PC chairs to identify those cases where a reviewer underperformed (or must be expected to underperform in the future), so they can focus their own energies on these most critical cases.

My proposals are to (*i*) ask reviewers to explicitly acknowledge their review assignment upon receipt, to (*ii*) work with two rather than a single reviewing deadline, to (*iii*) require SPC members to take explicit responsibility for the reviews written (by others) for the papers they handle, and to (*iv*) ditch triple-blind reviewing in favour of double-blind reviewing. Importantly, each of these improvements can be implemented independently from the others and can be expected to have a positive impact also in isolation.

FOUR PROPOSALS

For each of the four proposals for improving accountability in peer review I put forward here, I briefly describe the *problem* it is meant to address, I then sketch the *solution* I propose and discuss some of the *practicalities* involved with implementing that solution, and I finally report on the *impact* it had at AAMAS-2021 and ECAI-2024.

Explicit Confirmation of Review Assignment

Problem. When PC chairs send out the review assignment, they cannot be certain that every reviewer takes notice. Indeed, a significant number of such emails get caught in spam filters, reviewers might overlook an important email, some might experience health issues, and others might have forgotten what they committed to several months earlier. The PC chairs will typically realise that there is a problem only when the reviewing deadline has passed. It also is not uncommon that a reviewer realises they have a conflict of interest for a specific paper only once they start reading it, and they might do so only a couple of days before the deadline.

Solution. Ask every PC member to explicitly confirm—within 72 hours of the review assignment having been sent out—that they will be able to review each one of the papers assigned to them. Ask them to download the papers before responding and include an explicit question about possible conflicts of interest. Follow up with individuals who are not responsive. Assign alternative reviewers in case people cannot be reached within a week or so. Consider doing something similar for SPC members and area chairs.

Practicalities. As far as I know, reviewing platforms currently do not fully support this measure. For AAMAS-2021, we used a simple Google form, which worked but is cumbersome and might not scale to larger conferences. For ECAI-2024, we used a new feature of EasyChair, called 'reviewer expertise', where PC members can rate their own expertise for reviewing a given paper assigned to them. This is not perfect (asking for too much information and not being explicit about PC members *committing* to reviewing papers they do not ask to be reassigned), but it still worked reasonably well.

Impact. For AAMAS-2021, this allowed us to catch several dozens of conflicts of interest within a week of the review assignment, meaning we were able to assign new reviewers at that time rather than during the hectic days around the final reviewing deadline. For this conference the number of reviewers found to be completely inactive at that time was small, maybe because AAMAS is still a relatively closely-knit community. For ECAI-2024, the situation was inverted. We only discovered a small number of previously undetected conflicts of interest but we decided to expel a significant number of individuals from the PC due to being unresponsive. Both of these differences with respect to AAMAS might be explained by ECAI being a much larger and broader conference.

A made-for-purpose confirmation feature that is directly integrated into the reviewing platform would further improve the usefulness of this measure, and make it less onerous for reviewers and PC chairs alike. I note that a tiny minority of reviewers might get irritated about being asked to do something they are not used to.

Staggered Reviewing Deadlines

Problem. Some reviews will be of poor quality and some will not arrive on time or at all. In the first instance, it is the task of the SPC member assigned to a given paper to ensure such problems are resolved before the start of the rebuttal phase. But typically around 50% of all reviews will arrive on the day of the reviewing deadline, making it impossible for the SPC member to predict earlier where an intervention might be required and then giving them very little time to deal with those cases that actually require an intervention.

Solution. Introduce two separate reviewing deadlines, one week apart, and ask each individual reviewer to submit at least 50% of their reviews by the first deadline. Reviewers who deliver as requested by the first deadline should receive a thank-you note acknowledging this fact. Reviewers who submit at least one review but fall short of the 50% target should get a friendly message noting this fact and expressing confidence that all remaining reviews will arrive by the second deadline. Reviewers who do not submit any reviews by the first deadline should receive an urgent request to contact the PC chairs (possibly through a form) to clarify their intentions for completing their review assignment.

Practicalities. While your reviewing platform of choice might not directly support sending targeted messages to reviewers based on the proportion of assigned reviews they have submitted, it is fairly easy to download the reviewing data and run a script on it to construct lists of addresses to send each of the three messages to.

Impact. Of course, not all reviewers will comply and deliver 50% by the first deadline, but for both for AAMAS-2021 and ECAI-2024 we still had in the order of 50% of all reviews in the system by that time (as some reviewers over-performed). So this important milestone was reached one full week early, giving SPC members and PC chairs one extra week to deal with difficult cases. This makes a huge difference. You have extra time to get individual reviews improved. You can alert reviewers about issues in their first batch of reviews before they submit the second batch. You can assign emergency reviewers to papers assigned to reviewers who turn out to be unresponsive one week earlier than you would otherwise.

For both conferences, by the start of the rebuttal phase every single paper had the required number of 3 reviews in the system (while the average number of reviews written was below 3.1 per paper, meaning the target of 3 reviews per paper was reached without simply commissioning way too many reviews and thereby making the community work overtime). I'm not aware of any other large AI conference in recent years that achieved the same, and I believe the main reason why it worked for these two conferences was the use of staggered deadlines (with the first measure discussed above also playing a significant role).

But how does this measure affect the reviewers themselves? Also here, a tiny minority of them, both good and bad ones, might get irritated by this attempt to take over their personal time management. Hopefully, this is just a matter of getting used to staggered deadlines. In any case, a conscientious reviewer will start their work some time in advance of the final deadline anyway, so this measure should have but a very minor impact on their way of working.

Explicit Approval of Reviews

Problem. A good SPC member will try to follow up with reviewers who deliver substandard reviews on their own accord, but it can be difficult to judge when to do so. And for PC chairs it is almost impossible to get an overview of which reviews are fine, for which ones an SPC member will ensure whatever issues are there will get resolved, and for which ones they need to intervene themselves.

Solution. Right after each reviewing deadline, for each review submitted, ask the SPC member handling the paper whether they *approve* the review, i.e., whether they consider it to be of sufficient quality to be released to the authors. This makes them accountable. If they are not prepared to approve a given review, the SPC member should describe the problem in a sentence or two and indicate whether they expect to be able to get the problem resolved themselves or whether they require help from the PC chairs. For very large conferences, it makes sense to also ask area chairs for approval of the metareviews written by SPC members.

Practicalities. I'm not aware of any reviewing platform that fully supports this measure. Some, such as CMT, allow SPC members to *rate* reviews. But this is not the same: you need an explicit statement from the SPC member that they approve a review, you need a short description of the problem if they do not, and you need the option for them to ask for help or to say explicitly that they do not need help. For AAMAS-2021 we implemented review approval using Google forms and for ECAI-2024 using a custom-made website. The latter approach is clearly superior but requires nontrivial technical support to be set up. Of course, integrating this step into the reviewing platform would be much better.

As an aside, let me note that reviewing platforms typically do not permit anyone but PC chairs to edit reviews, and that this is an important feature—helping safeguard accountability and transparency in its own way—that should not be over-ridden just to simplify the task of SPC members to get problematic reviews fixed.

Impact. This measure clearly improved the quality of reviews for both AAMAS-2021 and ECAI-2024. Still, it is not uncommon that reviewers do not react, or do not react adequately, to prompts received from SPC members and this can be a frustrating experience for the latter. Also, not all SPC members will perform this task (in a timely fashion or at all). But even if just 80–90% of them do, as a PC chair you can turn the task of monitoring review quality from something that is absolutely overwhelming into something almost feasible, by trusting the judgments of SPC members who participate and focusing your own attention on the remaining reviews.

Reviewer-Reviewer Visibility

Problem. In an attempt to empower junior reviewers to speak their mind (and also to combat fraudulent behaviour such as the coercion of fellow reviewers), in recent years there has been a trend at large AI conferences to replace *double-blind reviewing* (reviewers do not see authors, authors do not see reviewers) with *triple-blind reviewing* (reviewers also do not see each other). The downside of this is that it makes reviewers less accountable for the work they deliver. And reviewers now miss out on the opportunity to earn a reputation of being a good reviewer. PC discussions have become unengaging, unrewarding, and often plain confusing.

Solution. Revert back to double-blind reviewing, which used to be the standard model for AI conference reviewing until recently.[1] To address the concern of junior reviewers possibly not being able to freely express their views when not protected by anonymity, issue adequate guidelines for proper conduct (especially by senior members of the community) during PC discussions.

Practicalities. This would be straightforward to implement, as presumably all reviewing platforms give PC chairs the option to customise who can see who within the PC. Note that it is important that (regular) PC member A gets to see B (and B's review) only *after* A has submitted their own review (but I believe that essentially all reviewing platforms implement this rule). A possible refinement worth considering (and as far as I know not currently implemented in common reviewing platforms) would be to only show A and B to each other once *both* of them have submitted their reviews.[2] This would make coercion much harder, as it would become only possible after the target has submitted their own review and any later updates to that review would have to happen in plain sight of all PC members assigned to the paper (assuming of course that the reviewing platform keeps previous versions of a review visible to the PC—as it anyway should and as, e.g., EasyChair does).

Impact. In my personal experience of serving in a variety of roles on the committees of many conferences that have used either double-blind reviewing or triple-blind reviewing, the former with its possibility to address each other by name leads to more engaging and productive PC discussions. This is not surprising. Being able to see who you are talking to makes participating in a discussion much more interesting. Being able to put names to arguments also makes it easier to remember what the current state of a discussion is (these discussions typically extend over a week or ten days, with long breaks between turns, and most interlocutors are part of several such discussions running in parallel). Besides improving the quality of these discussions and thereby the quality of the decisions taken on their basis, the double-blind model also provides each participant with the opportunity to distinguish themselves through exemplary reviews and exemplary engagement during PC discussions. And it makes it easier to learn from others (as this works much better when one can put a name to a review or a comment).

This is not to say that the aforementioned arguments in favour of triple-blind reviewing do not have merit. But for any one conference they concern only a small minority of papers and PC members. On the other hand, the quality of PC discussions affects essentially all papers and—let's not forget—all PC members. I believe that the concern of senior reviewers misbehaving and junior reviewers not feeling at ease can be addressed, at least to a good extent, by issuing appropriate guidelines and by having 'people skills' in mind as a relevant criterion when appointing SPC members and area chairs. The concern regarding coercion cannot be fully eliminated, but the above-mentioned idea of hiding reviewers' identities from each other until both of them have uploaded a review should significantly reduce it and it presumably—and hopefully—only affects a tiny minority of papers and PC members anyway.

CONCLUDING REMARKS

While the fourth proposal, to ditch triple-blind reviewing in favour of double-blind reviewing, might be controversial, the other three proposals surely are not. They are straightforward practical improvements to the reviewing process and I urge every PC chair to implement them. As reported earlier, all four measures had noticeably positive effects for both AAMAS-2021 and ECAI-2024.

Let me conclude with three remarks on good practice that do not directly fall under the header of 'accountability' but that nonetheless relate to the concrete measures discussed here.

First, reviewers who do everything right should not get spammed with endless reminders. What usually happens is that PC chairs start to panic when they realise just how long it takes for reviews to come in. They then send out ever-more desperate blanket reminders to *all* reviewers. This is unpleasant for recipients, can have a negative effect on their motivation to do a good job, and makes it more likely they overlook a future reminder that actually is intended for them. These reminders also contribute to normalising the notion that being late on a review is somehow socially acceptable (which it is not). Instead, reminders should be targeted. Anyone who submitted at least one review on time can be assumed to be aware of the deadline and thus does not need to be reminded. Anyone who made the deadline should receive a thank-you note rather than one of those conditional reminders ("in case you did not yet . . . ").

Second, we should do away with the false narrative that reviewing somehow is 'voluntary' work. It is not. Everyone who submits papers that get reviewed by others also as an obligation to contribute to reviewing themselves. There can be exceptions (still being in training, contributing to service in other ways, being on parental leave, etc.), but these are only ever temporary in nature. Every employer, also outside of academia, who pays their staff to do research and expects them to submit papers also has an obligation to pay for their time when they do reviewing work.

Third, authors should get clear instructions about when it is acceptable to formally complain about review quality (and when it is not) and they should receive an explicit invitation to consider doing so. We did this for AAMAS-2021 and ECAI-2024. While review quality was far from perfect for these conferences, we received official complaints for only about 1% of papers, of which roughly half were deemed to warrant some kind of intervention from us. This approach not only allows you to correct some major mistakes but it helps authors feel heard and, hopefully, improves people's attitude towards peer review—despite its many imperfections.

[1] The question of whether reviewers should be made aware of the identity of the authors of they papers they review (as in *single-blind reviewing*) is orthogonal to the question of whether reviewers should learn about each others' identities. I shall not discuss this point here and only note that, in the AI community, the arguments against single-blind reviewing are well known and most people agree with them, also in the face of the obvious realisation that perfect anonymisation is hardly possible.

[2] What I do see being in done at some conferences is that no reviewer can see *any* reviews (or fellow reviewers) until the end of the rebuttal phase. While this addresses the coercion concern, it has other disadvantages: fewer opportunities to correct wrong or unprofessional reviews, and more waiting time for good reviewers who delivered on time and are simply curious what their colleagues had to say.