

Hands-Off or Hands-On? Variation in Area Chair Practices and Implications for AI Support

Ines Arous¹, Neha Nayak Kennard², Andrei Mircea³, Emily Kuang¹, Jackie Chi Kit Cheung⁴,
Andrew McCallum²

¹York University

²University of Massachusetts Amherst

³University of Montreal, Mila

⁴McGill University, Canada CIFAR AI Chair, Mila

inesar@yorku.ca, kennard@cs.umass.edu, mirceara@mila.quebec, ekuang@yorku.ca, cheungja@mila.quebec,
mccallum@cs.umass.edu

Abstract

Area chairs (ACs) play a critical role in the peer-review process, managing conflicts and ensuring fair outcomes. Although AI tools have been proposed to support ACs, little is known about the challenges they face and their perceptions of these technologies. In this paper, we conduct interviews including a design probe with 27 ACs in AI to explore their challenges, strategies, and perspectives on potential AI tools. Through thematic analysis, we identify key tensions arising from the growing volume of submissions, uneven reviewer expertise, and the complex task of managing the relationship between reviewers and authors. Most importantly, we find substantial variation in how ACs engage with submissions and influence outcomes: some adopt a largely hands-off approach, while others take a more hands-on role in guiding discussions and decisions. This variation challenges the notion of a single, universal AC practice and highlights the need to account for diverse approaches. When reflecting on the potential use of AI tools, ACs expressed a cautious stance, drawing on their domain knowledge and heightened awareness of AI limitations. From these findings, we derive three design implications: tailoring AI assistance to diverse AC practices, design assistance for discussion moderation, and embedding human-centered AI principles that preserve human agency in decision-making.

1 Introduction

Scientific peer review is pivotal to maintaining quality standards for academic publication. Within this process, area Chairs (ACs) play a critical role: they mediate conflicts, synthesize divergent reviews, and ensure a fair decision for each submission (Ailamaki et al. 2019; Resnik and Elmore 2016). All of these responsibilities are carried out under tight deadlines and an ever-growing number of submissions, making their role both demanding and time-sensitive (Galipeau et al. 2016; Tran et al. 2020). For instance, the International Conference on Learning Representations (ICLR) 2025 received 11,603 submissions (Miley, Jill 2025), which is an exponential increase from just 490 submissions in 2017. At AAAI, submissions reached 29,000 in 2025 compared with just 1,991 in 2015 (AAAI 2025). This unprecedented scale

places compounding pressures on the review system, from limited workflow support and reviewer scarcity to the difficulty ACs face in maintaining fairness under time pressure.

Various tools have been designed to support the peer review system by providing scaffolding for *reviewers* to enhance the quality of their feedback (Hicks et al. 2016; Ngoon et al. 2018; Sun et al. 2024a) and to support *authors* in interpreting peer reviews (Yang et al. 2025; Yen, Kim, and Bailey 2020; Yen et al. 2017). Comparatively little attention has been paid to designing tools for the *distinctive needs of ACs*, whose role requires making high-stakes decisions under tight time constraints. At the same time, substantial work has explored the use of AI, and more recently, large language models (LLMs), to support or mimic components of the AC decision-making process, such as predicting acceptance decisions (Checco et al. 2021) and summarizing reviews (Li, Hovy, and Lau 2023; Li, Lau, and Hovy 2024; Zeng et al. 2025). Some of these techniques have been deployed in a recent pilot program in AAAI-26 (Ellison 2025). While these approaches promise efficiency gains, they often overlook existing AC practices and willingness to adopt such tools, as well as the risks posed by poor integration, including the reinforcement of existing biases (Fecher et al. 2025).

Our work seeks to understand ACs’ challenges, strategies, and expectations, and explore how potential designs for AI assistance might align or conflict with their real-world practices. Unlike prior work that proposed automated methods without engaging ACs, this study centers their perspectives to inform the design of future AI assistants. Specifically, our study was guided by the following research questions (RQs):

- RQ1: How do ACs experience and navigate their role?
 - RQ1a: What challenges do ACs currently face?
 - RQ1b: What are their current strategies for addressing these challenges?
- RQ2: What are ACs’ perceptions of AI tools, and what ethical concerns do they associate with their use?

We conducted 27 semi-structured interviews with a design probe with ACs from ICLR, a uniquely suitable context for this study. ICLR is the only major AI venue that publicly releases reviews and AC feedback for both accepted and rejected papers, making it a frequent testbed for

AI-based peer-review tools (Arous et al. 2021; Zeng et al. 2025). Through these interviews, we observed substantial variation in the extent of AC engagement with submissions: some ACs took a hands-on approach and actively advocated for promising work, whereas others were hands-off, adopting a judge-like stance and deferring to reviewer consensus. We also found that nudging reviewers was widely seen as the most challenging task, described as frustrating and misaligned with ACs' seniority. We further employed a design probe using wireframes of four potential AI-assistive tools to elicit ACs' design considerations. ACs expressed cautious optimism about AI assistance, recognizing its potential to alleviate cognitive burden while voicing concerns about bias amplification, misinterpretation, and over-reliance without appropriate safeguards. They further emphasized the importance of personalizing AI support, as there is no one-size-fits-all solution for assisting ACs. Drawing on these perspectives, we distill key design implications for future AI assistants that support, rather than undermine, AC judgment.

2 Related Work

2.1 AI methods for Scientific Peer Review

Various tools have been designed to support the peer review process by assisting reviewers and authors. For instance, some methods have explored introducing interactive techniques such as reusable suggestions and adaptive guidance (Ngoon et al. 2018), or scaffolding novice reviewers through contextual reflections (Sun et al. 2024a). Other efforts have focused on supporting authors in interpreting and responding to peer reviews, by generating a positive summary of the reviews (Yang et al. 2025) or by assisting authors in identifying priorities and reflecting on their work (Yen, Kim, and Bailey 2020; Yen et al. 2017).

While various tools have been developed to assist both reviewers and authors, few design studies have addressed the *unique and often conflicting priorities that ACs must navigate when balancing the needs of both groups*. Recent work has begun to explore AI-driven approaches to assist ACs. One such approach involves calibrating reviewers' scores (Arous et al. 2021; Flach et al. 2010; Tan et al. 2021), which helps ACs assess the reliability of each review. Other approaches analyze review and rebuttal discourse structure to provide ACs with an overview of the main arguments (Wang and Wan 2018; Ghosal et al. 2019; Darrin et al. 2024; Kennard et al. 2022). More recently, various efforts have been dedicated to mimicking the ACs' role by predicting the acceptance decisions of submissions (Checco et al. 2021) and summarizing the reviews with LLMs (Li, Lau, and Hovy 2024; Zeng et al. 2025; Sun et al. 2024b). Some of these methods have been integrated into the pilot AI Review program at AAAI-26, demonstrating AI's potential to make meaningful contributions to peer review (Biswas et al. 2026).

However, prior efforts have largely relied on speculative assumptions about ACs' challenges rather than insights from design research or direct engagement with ACs. As a result, these approaches often fail to reflect the nuanced realities of AC work. To address this gap, our study engages directly

with ACs through interviews and a design probe, uncovering their challenges and collaboratively identifying opportunities for AI-driven design to support their roles.

2.2 Human-AI Collaboration for Decision Making in High-Stakes Domains

Many human-AI collaboration tools have been designed to assist decision-makers in high-stakes domains, including education, medicine, and law (Fok et al. 2023; Norkute et al. 2021; Fogliato et al. 2022; Grgić-Hlača, Engel, and Gummadi 2019). Some studies have demonstrated that effective collaboration between decision-makers and AI can improve decision-making processes under time constraints (Jacobs et al. 2021a; Bansal et al. 2021; Kawakami et al. 2022b). These tools can provide invaluable support by highlighting relevant information (Yang et al. 2023; Burgess et al. 2023; Han et al. 2024) and effectively mitigating potential biases (Zhang, Bellamy, and Kellogg 2015; Wang, Harper, and Zhu 2020). However, other studies show that human-AI decision-making may fail to improve performance or even lead to incorrect decisions in high-stakes domains compared with either human or AI decisions alone (Cheng et al. 2022; Green and Chen 2019a; Kawakami et al. 2022a; Green and Chen 2019b; Grgić-Hlača, Engel, and Gummadi 2019). In several studies, decision-makers' attitudes toward AI contribute to shaping the outcome: excessive skepticism can lead to under-reliance on useful predictions, while over-confidence can result in over-reliance on erroneous recommendations (Cabitza et al. 2023; Dietvorst, Simmons, and Massey 2015; Kawakami et al. 2022b).

To mitigate these risks, prior research has explored design strategies that empower decision-makers to critically evaluate and, when necessary, override AI predictions. These designs aim to preserve human agency by presenting AI-generated recommendations while leaving the final judgment to humans (Chandrasekharan et al. 2019; Grgić-Hlača, Engel, and Gummadi 2019). Other approaches seek to mitigate the risk of decision-makers' over-reliance on these systems through a range of design strategies. These include timing AI assistance (Fogliato et al. 2022), encouraging active participation of decision-makers (Solomon 2014), and seeking support from external sources, such as color-coded tags (Bansal et al. 2021; Fok et al. 2023; Fogliato et al. 2022), or presenting additional context relevant to the decision (Head et al. 2021; Cai et al. 2019; Yang et al. 2023). These strategies help users quickly identify key information without searching or reading entire documents. An alternative approach is to provide concise summaries of relevant information (Norkute et al. 2021; Zhang et al. 2023), which help reduce cognitive load and improve decision-making efficiency (Han et al. 2024). Other studies highlight the importance of personalizing AI assistance to decision-makers' workflows and needs (Corti et al. 2024; Bertrand et al. 2023; Tankelevitch et al. 2024). For instance, Tankelevitch et al. (2024) argues that personalization in generative AI systems supports users' metacognitive processes by enabling them to adapt outputs to their workflows while calibrating their confidence in the system. Similar findings have been reported in clinical settings (Corti et al. 2024; Choi and Lee 2026),

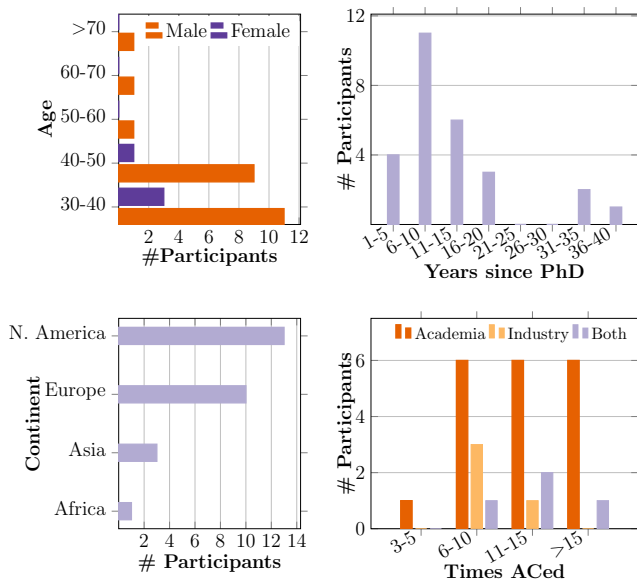


Figure 1: Demographics of the ACs we interviewed. Our participants represented diverse ages, genders, affiliations, and seniority levels.

where personalization mechanisms help preserve human agency by aligning AI systems with local practices and different stages of decision-making workflows. Despite these efforts, it remains an open question how to effectively foster human-AI decision-making, particularly in high-stakes, multi-party contexts like peer review. AC decision-making is distinctive in that it requires balancing domain expertise with deference to reviewer consensus, all while preserving reviewer anonymity and operating under tight deadlines.

3 Method

We conducted semi-structured interviews with 27 ACs from machine learning conferences. To elicit ACs’ intuitions and perceptions about AI support, the interviews incorporated a design probe using wireframes of four potential AI-assistive tools. This approach allowed participants to envision how AI could fit into their workflows without the constraints of fully implemented prototypes. The speculative nature of the wireframes enabled exploration of broader challenges and perspectives beyond any single design. Each wireframe depicted different modes of AI use, based on existing methods across various reviewing sub-tasks. This research project was reviewed and approved by our institution’s IRB.

3.1 Participants

Our intended participant pool was researchers who had previously served as ACs for machine learning conferences. To reach this specific group, we contacted ACs for ICLR, a top-tier machine learning conference (see the appendices for details on the ICLR reviewing process¹). Although the

inclusion criterion was having previously served as an AC for ICLR, it is worth noting that all participants had served as ACs for multiple machine learning conferences. During the interviews, we encouraged them to draw on their experiences across these various conferences. Using ICLR as an inclusion criterion enabled us to leverage the publicly available lists of ACs on ICLR websites² to retrieve a list of 561 former ACs. Since ICLR is hosted on OpenReview, we also used the OpenReview API³ to retrieve the homepages of these researchers. We manually collected publicly available email addresses of as many researchers on this list as possible, found either on their homepages or recent publications.

This process resulted in a collection of 540 email addresses. Expecting a very low response rate, we contacted all 540 candidates by email, explaining the goal of our study and requesting their participation in a semi-structured interview. In total, 58 potential participants responded to our email, of whom 28 consented to participate in the study. The demographics of all participants are depicted in Figure 1. We list aggregated characteristics of study participants to prevent deanonymization. All participants reported their gender identity as either *man* or *woman*, and their ages ranged from 30 to 70 years or older. They were based across four continents: Africa, Asia, Europe, and North America. All had served as ACs at least three times. Nineteen participants were affiliated with academia, four with industry, and four held joint appointments in both sectors. The interviews were conducted in August and September 2023.

3.2 Interview Methodology

We conducted 30 to 45-minute remote semi-structured interviews via Zoom, which were recorded with participants’ permission and transcribed. The interviews were structured into two main parts. The first half focused on understanding each participant’s experience as an AC. The second half discussed how ACs envisioned using AI-assistive tools. The interview transcript is provided in an appendix.

AC Experience and Workflow: Interviews began by asking each participant about their experience as an AC for ICLR. Next, we asked ACs how they viewed their role and the specific tasks in their workflow. These open-ended questions allowed participants to define their responsibilities on their own terms and to surface the challenges they encountered at each stage of the process. They also discussed their strategies and shared their perceptions of what is effective and less effective when managing different reviewing cases.

Design Probe: In the second part of the interview, participants were asked to describe their view of a potential AI-assistive tool and the capabilities they envisioned for it. To elicit design considerations, we employed a design probe using four wireframes (see Figure 2) and presented them to participants as static slides in an electronic slideshow. These wireframes were not intended as final system designs but rather to spark discussion with participants (Yudhanto, Pryhatyanto, and Sulandari 2022; Zaidi and Karahalios 2025;

¹ Appendices can be found in the extended manuscript, which is available on arXiv.

² e.g., <https://iclr.cc/Conferences/2022/ProgramCommittee>

³ <https://docs.openreview.net/>

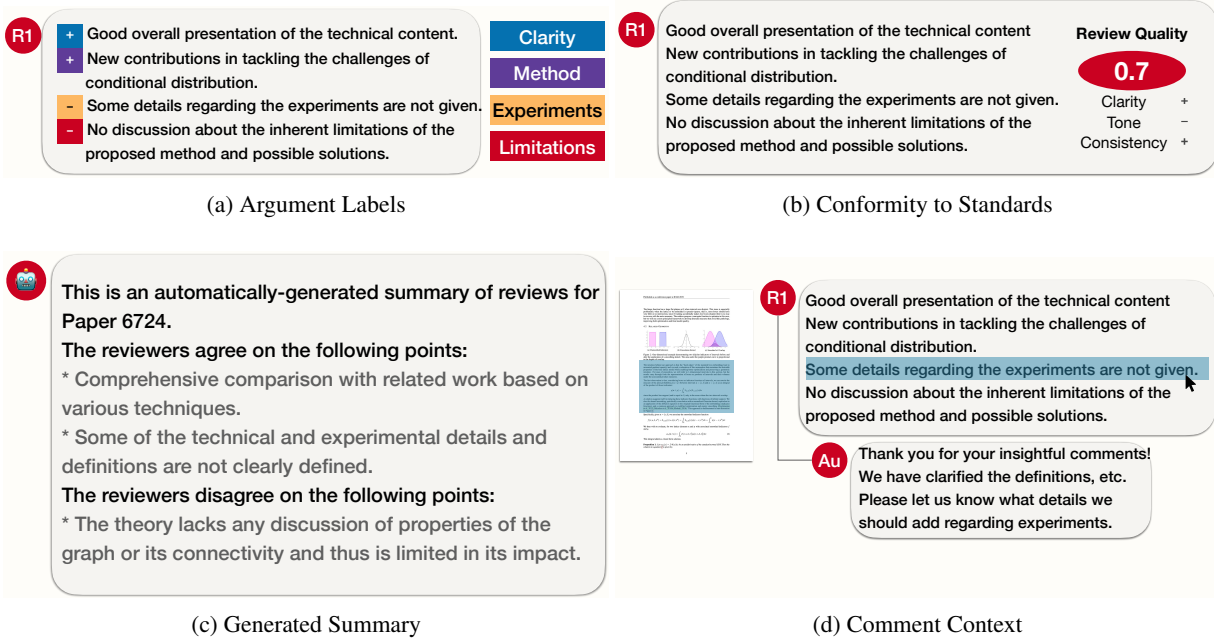


Figure 2: The wireframes shown to participants: (a) Argument labels: each argument within a review is associated with a color indicating the discussed aspect (e.g., clarity, method), and is accompanied by a (+) or (-) sign to denote its sentiment, (b) Conformity to standards: each review is associated with a number reflecting its conformity to a conference standards, (c) Generated summary: an automatically generated summary from all reviews, (d) Comment context: highlights the context of each argument within a review from the related paper.

Biermann, Ma, and Yoon 2022). We adopted this approach over interactive evaluation to focus on conceptual possibilities and prevent participants from being distracted by technical issues such as latency or interface details.

The order in which each wireframe was shown to participants was randomized to maximize the breadth of design discussions and mitigate order effects. For each wireframe, the interviewer explained the design concept and presented corresponding slides illustrating the features of an AI-assistive tool. After each design concept, the interviewer used probing questions to deepen understanding of how ACs perceived the tool’s advantages and disadvantages and how it fits within their workflows. Participants then ranked the designs according to perceived usefulness. The interviewer then asked follow-up questions about design considerations for assistive tools for ACs.

3.3 Creating the Design Probe

We constructed our design probe to present four major conceptual AI-assistive tools based on common challenges faced by ACs in prior work and grounded in AI and HCI literature for peer review. We then conducted brainstorming sessions (approx. 4 hours in total) to sketch the initial wireframes. During brainstorming, our research questions served as a starting point and an anchor. While we incorporated several assistance methods, we did not include those that fully automate AC decision-making, as we target tools that augment rather than replace human judgment. The initial sketches were iteratively refined into the final versions

presented to participants (see the appendices for a summary of key changes and an illustration of the sketches before and after our iterations). The wireframes are discussed below:

Argument labels (Fig. 2a) Several systems have supported decision-making by visually highlighting task-relevant features and annotating them with descriptive labels via color-coded tags (Bansal et al. 2021; Fok et al. 2023; Fogliato et al. 2022). This design feature is particularly valuable when users must process long documents, as it enables quick identification of relevant sections without exhaustive reading. Empirical evidence supports these benefits: in one study, it was found that color-coded tags helped participants locate relevant information faster and maintain focus (Fok et al. 2023). In our context, ACs navigate long reviews at different stages, first to identify topics discussed during rebuttal, and later to synthesize reviewers’ arguments when drafting their meta-review (Wang and Wan 2018; Ghosal et al. 2019; Hua et al. 2019). Thus, we developed a wireframe that color-codes each review argument by the aspect discussed (e.g., method) and marks its associated sentiment with a positive (+) or negative (−) sign. The wireframe offers an alternative view (shown in an appendix, Fig. 5-C) that clusters all arguments related to the same aspect, with each argument accompanied by a reviewer’s identification.

Conformity to standards (Fig. 2b) A substantial part of the AC’s decision-making process involves weighing reviewers’ arguments and flagging potential issues. To address these challenges, AI methods have been developed

to evaluate review quality (Arous et al. 2021; Flach et al. 2010; Tan et al. 2021). Similar AI-generated predictions have been used to assist decision-makers in diverse domains: for instance, systems have shown predicted biometrics to physicians when prescribing medications (Burgess et al. 2023), toxicity scores to content moderators (Chandrasekharan et al. 2019), risk assessments to judges making bail decisions (Grgić-Hlača, Engel, and Gummadi 2019). Therefore, we design a wireframe that shows a score reflecting a review’s conformity to conference standards that covers three dimensions: 1. **Clarity** evaluates if a review is well-structured and includes supporting arguments for the recommended score; 2. **Consistency** checks alignment between the recommended score and the review content; and 3. **Tone** evaluates the politeness of a review.

Generated summary (Fig. 2c) Summaries have been shown to reduce cognitive load and allow decision-makers to complete tasks more efficiently (Han et al. 2024). They have been applied in domains such as legal case analysis (Norkute et al. 2021), research ideation through literature-based summaries (Pu et al. 2025), and clinical review by condensing patient–doctor dialogues (Han et al. 2024). Building on this principle, various LLM-based methods have been developed for meta-review generation (Li, Hovy, and Lau 2023; Li, Lau, and Hovy 2024; Zeng et al. 2025). In our design, we opted for extractive summarization based on early brainstorming discussions with team members who had AC experience. They emphasized the risks of hallucinations associated with abstractive summarization and preferred to retain the original wording of reviewers’ arguments. Therefore, our wireframe proposes a method that clusters reviewers’ arguments to highlight areas of agreement and disagreement, while separately compiling additional comments that appear when a reviewer raises a unique point.

Comment context (Fig. 2d) Various reading interfaces aim to enhance decision-making by presenting contextual evidence that users would otherwise need to search for or extract through extensive reading. For example, some tools highlight relevant patient histories in clinical settings (Cai et al. 2019), while others provide in-context definitions of technical terms in academic papers (Head et al. 2021). Inspired by these designs, we illustrated a wireframe that enables navigation of reviewer–author discussions by highlighting the paper’s context for each argument and displaying the corresponding author response.

3.4 Analysis

We employed reflexive thematic analysis to analyze the interview data (Braun and Clarke 2021). Our process involved several iterative stages. First, three of the authors read the transcripts to familiarize themselves with the data. They then inductively developed codes through open coding, identifying segments of data potentially relevant to the research questions and assigning descriptive labels. Second, they iteratively refined existing codes by comparing them to the interview transcripts. Next, the coders clustered related codes into preliminary themes that captured recurring patterns or concepts. To strengthen the analysis, all authors engaged in

multiple synchronous meetings to discuss the themes that were actively created and iteratively refined them, ensuring that each one was “clearly demarcated and built around a strong core concept” (Braun and Clarke 2021). Ultimately, we created 162 codes and 6 main themes, such as: ‘a challenging dynamic between passive reviewers and persistent authors, and ACs’ frustration with chasing responses’; and ‘the scalability challenges in peer review with ACs’ cognitive overload’. We continually assessed the alignment of the developed themes with our research questions. We organized the findings according to our research questions. For RQ1 (challenges and strategies), we iterated until we had a set of distinct themes that captured participants’ experiences. For RQ2 (feedback on the design probe), we grouped responses into advantages, concerns, and desired design features, refining these categories until we obtained a representative set.

4 Findings

4.1 Existing Challenges (RQ1a)

Navigating Tensions in Managing Reviewers and Authors ACs are tasked with managing a challenging dynamic between reviewers and authors: reviewers contribute their time and expertise without direct incentives, while authors are eager for their papers to be accepted.

Reviewers’ lack of commitment. Participants admitted that reviewers often lacked commitment during the review process. They frequently submitted their reviews late, or not at all, often without notifying ACs. This lack of commitment was also reflected in the quality of reviews: reviewers sometimes overlooked key aspects of a paper’s contribution or confidently judged it without clear justification. Such shortcomings were evident in short, generic, and poorly justified reviews. Among the 27 interviewed ACs, 22 reported that reviewers often disengaged from the process after submitting their reviews, despite AC reminders. Their opinions remained unchanged after the rebuttal or even “ghosted” the authors’ responses. One AC admitted:

P1: It’s sort of running back and forth digitally between folks, and I think it’d be much nicer if reviewers engaged more without the AC having to mediate.

Some ACs attributed reviewers’ lack of engagement to the limited time reviewers have, as “everything is as rushed as possible.” They also hypothesize that these issues stem from a lack of incentives, as reviewers felt obligated to give back to the community but did not receive tangible benefits.

Authors’ persistence. ACs noted that authors needed to push their papers for publication, as it was required to advance their careers. This created an asymmetric situation in which authors have strong incentives to engage in extensive discussion, whereas reviewers do not. Additionally, ACs observed a range of strategies employed by authors to persuade reviewers to reconsider their scores or to convince ACs that reviewers had misunderstood their work.

P19: You [as an author] are very polite when writing to the reviewers. But then you’re mean when writing to the AC, because your goal is either to convince the reviewer or to discredit the reviewer with the AC.

Herding cats. Managing this challenging dynamic between reviewers and authors came down to the ACs. They reported various strategies to “chase” reviewers, including sending reminders (*‘it’s due in two days!’*, *‘it’s due tomorrow!’*) and contacting reviewers through alternate channels such as Slack or their personal email. Participants found this task frustrating, noting that it did not require their level of seniority and left them feeling undervalued.

P13: You don’t need actually a lot of subject-specific knowledge, and it is a bit of a waste of time.

Scalability Challenges in Peer Review The rapid increase in submissions poses challenges for peer review.

Scarcity of expert reviewers. The disparity between the number of expert reviewers and the number of submissions frustrated ACs, as they received fewer qualified reviews than needed for a fair evaluation. To address this issue, various conferences recruited junior researchers as reviewers, which ACs viewed as “suboptimal”. For instance, 5 ACs found that junior reviewers tend to focus excessively on novelty and often hold extreme opinions: they either judged harshly on minor details, expecting groundbreaking contributions, or were overly positive, overlooking fatal flaws or existing work.

Cognitive overload. Participants described the difficulty of managing a large volume of submissions, sometimes being assigned over 20 papers for a conference. This heavy workload, combined with their busy schedules and the varying quality of reviews, made it difficult for them to track each paper. They highlighted the challenges of navigating the extensive information required for a fair and informed decision. Among them, 10 ACs aimed to collect as much information as possible and ensured that everyone involved had a fair opportunity to articulate their views. Gathering all this information involved facilitating productive discussions and navigating numerous detailed technical responses across different interfaces. Handling this information overload, particularly in conferences with lengthy discussions, was overwhelming.

P7: It’s sometimes kind of a cognitive overload to see what’s going on in the discussion.

4.2 ACs’ Strategies without Tool Support (RQ1b)

Nudging Reviewers ACs admitted that their most frustrating task was coordinating between reviewers and authors. They primarily sent reminders and prompted reviewers to update their reviews and participate in discussions.

P6: On the one hand, kind of bureaucratically, there’s like managing this, and part of it is like all the herding cats and making sure that the reviewers do their job.

In these communications, many ACs used template messages, either provided by the conference or developed from their prior experience, to save time. Among our participants, 16 ACs aimed to resolve disagreements and reach a consensus by encouraging reviewers with strongly divergent views or weaker arguments to read the other reviews, sometimes by explicitly prompting them. They also noted that the discussion would sometimes veer off course, necessitating their intervention to maintain a professional tone.

Varying Strategies in Interpreting Reviews In cases of disagreement among reviewers, we found that ACs used different strategies to weigh reviewers’ arguments. For instance, 5 ACs relied on the opinions of reviewers whose expertise closely matched the paper’s topic. Another 18 ACs relied on long and detailed reviews that provide concrete evidence for their claims and actively participate in the discussion. 3 ACs used authors’ responses to assess the quality of reviews. Some ACs noted that some reviews reflected personal opinions or biases toward specific research areas rather than objective assessments. To address this, ACs described actively filtering facts from opinions and copying comments from reviews into a document or spreadsheet, organizing them by topic or severity. They also evaluated the extent to which the rebuttal addressed reviewers’ concerns.

Selective Engagement with Submissions and Decisions

Across the interviews, we observed that ACs vary widely in how much they engage with submissions and in their level of involvement in the outcomes. For instance, 14 ACs leveraged their seniority and network in the decision-making process. They tended to review papers, advocating for promising ideas that can stimulate engaging discussions at conferences. They also sought to educate reviewers and help them advance in their roles. Other ACs distinguished their role from that of reviewers and were reluctant to read papers or provide their own opinions on submissions. Instead, they attempted to be impartial, listening to all sides and making a decision based on consensus and what they found most convincing. Several ACs admitted struggling with the extent to which they need to remain impartial rather than injecting their own opinions, where 5 ACs explicitly expressed concern about whether it was appropriate to overrule reviewers, especially when they disagreed with reviewers’ consensus.

P17: There are times when I think all the reviewers are wrong, and I disagree with them. Is it still my job to, like, maybe try to overturn the reviewers, or just ignore them and do what I want?

These variations were also reflected in the reported time spent on AC duties: some reported spending about 10 minutes per paper, while others estimated 45 minutes to several hours, with one AC dedicating a full week of work to these duties. Some ACs were aware of these variations and saw potential in using tools to make the process more effective.

P13: Maybe someone is hands-off, and they just don’t really do anything until the very end, or someone is hands-on, and maybe they’re spending a disproportionately large amount of time on the task [...] tools, slash training, and all that, presumably could play some role in making the process a bit more effective.

4.3 Perceptions of AI Tools and Ethical Concerns Among ACs (RQ2)

While reviewing the wireframe prototypes (Fig. 2), ACs discussed potential ways in which AI-assistive tools could support them. During these discussions, many ACs envisioned tools and workflows that extended beyond the presented designs. Alongside these opportunities, ACs expressed var-

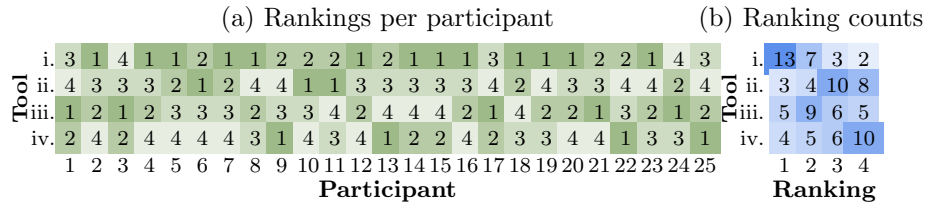


Figure 3: Stated preference rankings for the wireframes presented (i. argument labels; ii. conformity to standards; iii. generated summary; iv. comment context) for individual participants (a) and aggregated across participants (b).

ied preferences and raised ethical concerns about deploying such tools. We therefore report both wireframe-elicited responses and emergent design ideas, as well as participants’ concerns. We report the ACs’ ranking of the wireframes by usefulness in Fig. 3.

Wireframe-Elicited Visions We first present design concepts elicited through the four wireframe prototypes:

Argument labels. When shown this wireframe (Fig. 2a), the majority of ACs responded positively to the visual representation of reviewers’ arguments. Several noted that the interface appeared to be “replacing what an AC would do [...] barring the actual decision”. ACs appreciated that the interface streamlined the review process by enabling quick skimming and providing a structured view of reviewers’ concerns. They found that such an interface could assist in the decision-making phase by identifying common weaknesses in a submission, thereby making it more efficient to form their opinion. Additionally, they recognized its potential to simplify writing meta-reviews, as they would use the highlighted arguments from the reviews as evidence for their decision. At the same time, ACs expressed concerns about *over-reliance* on the tool. As experts in AI, they cautioned that errors in argument classification could lead ACs to misinterpret reviewers’ intent, potentially resulting in flawed judgments. Moreover, several ACs worried that presenting arguments in a structured, summarized form might reduce the decision-making process to counting pros and cons rather than thoroughly assessing the reviews.

P18: I’m not sure if the algorithm has grouped one sentence to a wrong class which leads, to a wrong direction, then lead me to a wrong decision. Who will be responsible for such a mistake?

ACs’ design feedback emphasized the importance of visualizing reviews “as-is,” rather than interrupting or restructuring their narrative flow. Several participants expressed concern that the alternative interface shown Fig.5-C in the appendices, which segmented and reordered review content, could increase the risk of misinterpreting reviewers’ arguments. Participants also suggested enabling both ACs and reviewers to modify the categorizations of reviewers’ arguments. They emphasized that such edits were necessary to ensure that the overall process remained trustworthy:

P09: I would prefer that any automated judgment come as early as possible in the process, so that there’s as many eyeballs as possible—human eyeballs.

Conformity to standards. 8 ACs expressed enthusiasm for the conformity to standards interface (Fig. 2b), which assigns a score to review quality based on multiple criteria. In terms of workflow integration, ACs noted that the interface could help them proactively identify reviews with toxic language or low-quality content. Others described using the scores to identify submissions that may require additional review, particularly when existing reviews scored poorly. By surfacing such issues early, the interface was seen as a way to intervene before reviews progressed to later stages of the process. Beyond the writing phase, ACs described using the aggregated conformity scores to “upweight or downweight” individual reviews, treating the signal as contextual information when synthesizing reviews. However, ACs were concerned that scoring inaccuracies could lead to unfair outcomes by downweighting certain reviews. Others raised concerns about the potential for *bias propagation*, particularly if existing reviewer biases were amplified through automated scoring mechanisms.

P1: Especially as somebody who does machine learning, I wanna know where these numbers came from.

Drawing on their expertise in machine learning, participants described how such a tool could benefit from *interpretable* models that allow ACs to understand the generated evaluations, rather than treating scores as opaque judgments. In addition, they suggested adding a *personalization* feature that enables the tool to adapt to individual ACs’ preferences. For instance, several ACs suggested incorporating additional signals, such as review length, technical soundness, or reviewers’ expertise. ACs also suggested adding a feature to enable them to calibrate the system based on their own assessments of review quality. For example, some envisioned adjusting weights in response to how closely the system’s evaluations aligned with their personal judgments over time.

Generated Summary. 12 ACs considered the generated summary a useful tool that could substantially save them time if it faithfully conveys reviewers’ concerns. They found that they could use it to have a first glance at all reviews, as a starting point to writing the meta-review, or to verify that they did not miss any comments. Despite its potential to save time, ACs were concerned about the possibility of it becoming an “automatic AC.” They believed that the meta-review should go beyond summarizing reviews and instead provide feedback to authors. 6 ACs were concerned with the risk of hallucination, where the generated content might deviate from the original reviews and fail to convey the signif-

icance of each argument. Moreover, many expressed apprehension that relying too heavily on such a tool could foster “laziness,” with ACs relying on it exclusively for their meta-reviews, potentially resulting in unfair decisions.

To mitigate these risks, participants emphasized the importance of ensuring that ACs remain the central decision-makers, with a tool that provides assistance, akin to *human-in-the-loop* systems. They stressed that automation should assist sense-making without replacing expert judgment. Participants recommended interface designs that require ACs to actively edit or revise the generated summary rather than allowing direct copy-and-paste, to encourage deliberation and accountability. Additionally, they suggested enabling reviewers to validate the generated summary to ensure it accurately reflects their arguments.

P14: Making sure you force them [ACs] to be an active component of the loop [human in the loop].

Comment context. 12 ACs identified the *comment context* interface as a valuable time-saving tool, particularly for navigating lengthy review discussions. They appreciated that it reduced the need to manually search through extensive exchanges to locate the context surrounding reviewers’ concerns. Additionally, they noted its usefulness in monitoring the discussion between authors and reviewers. At the same time, ACs cautioned that this solution could not be “one-size-fits-all.” Several emphasized that the interface would need to integrate with their existing reading practices, such as PDF viewers or printed submissions, rather than imposing a separate workflow. Others raised concerns about the reliability of AI-based contextualization, noting that inaccuracies could lead the system to omit relevant text or link comments to incorrect contexts.

In terms of design features, participants described how such a tool could help them keep up with ongoing discussions by explicitly highlighting arguments that require AC attention. Rather than scanning entire threads, ACs envisioned being proactively directed to comments or exchanges that raised unresolved concerns, requested clarification, or introduced new substantive points through the interface. Others suggested augmenting this functionality by visually marking “hot spots” within a submission, such as areas where multiple reviewers raised concerns. By surfacing these regions, the tool could help ACs focus their reading on contested areas, supporting more efficient sense-making.

Emergent Design Ideas Beyond the Wireframes In addition to responding to the wireframe prototypes, ACs articulated new design ideas. These emergent concepts arose organically through reflection and discussion.

Assisting in reviewers’ communication. When discussing the potential of AI-assistive tools, seven ACs identified reviewer communication as one of the most frustrating and time-consuming tasks that can be automated. In particular, they described the need to repeatedly follow up with reviewers to confirm responsiveness as largely “procedural” work that does not require ACs’ expertise and could be “easily automated.” ACs envisioned tools that support reviewer communication throughout multiple phases of the review process. During the initial review period, they suggested

lightweight responsiveness signals that allow reviewers to indicate continued commitment, such as a simple acknowledgment: “I’m going to do this, but I just haven’t done it yet”, which one AC noted could substantially reduce anxiety around missing reviews. During the rebuttal phase, participants proposed automated reminders prompting reviewers to read and respond to authors’ rebuttals. For example, one AC described a web-based form with a checkbox that would allow reviewers to indicate whether the rebuttal changed their position. One AC framed such a system humorously, revealing underlying frustration with the process:

P06: A robot that could hunt down my reviewers and beat them or a way more annoying alert system to get people to do the work they’re supposed to do.

Assisting reviewers and fostering accountability. Given the need to recruit junior reviewers, 10 ACs suggested training programs or tooling to scaffold review writing, arguing that supporting reviewers earlier in the process could reduce the burden on ACs later. At the same time, participants highlighted challenges stemming from the voluntary nature of peer review, which led to inconsistent commitment among reviewers. To address these challenges, 6 ACs suggested leveraging conference records or a tool such as the conformity to standards (Fig. 2b) to track reviewers’ reliability and inform future reviewer selection.

5 Discussion

5.1 Beyond One-Size-Fits-All (RQ1)

Our findings reveal that ACs vary widely in how much they engage with submissions and in their level of involvement in the outcomes. This variability resonates with prior work on peer review, which frames it as a constellation of practices encompassing consultative, decision-making, administrative, and moderating activities (Reinhart and Schendzielorz 2024). These activities are often led or supervised by ACs (Roth 2006; Hall, Hundley, and van Teijlingen 2015; Galipeau et al. 2016; D’Andrea and O’Dwyer 2017). However, our study extends this literature by demonstrating that AC engagement with these practices is not uniform but somewhat situational, with some adopting a hands-on approach while others are more hands-off. While previous research has examined reviewer engagement and the challenges they face (Lee et al. 2020b; Wilson and Nacke 2022), our work is the first to foreground the nuanced challenges ACs encounter in navigating these activities.

Design Implications: Develop Support Tools that Offer Personalized Assistance Rather than prescribing a single “ideal” practice, future systems should acknowledge and legitimize the variability in ACs’ practices while offering strategies to manage tensions, such as balancing advocacy with impartiality. This perspective also helps explain why many ACs emphasized personalization features and why preferences varied across wireframes. Consequently, we argue that *there is no one-size-fits-all solution for assisting ACs*. Individual differences, such as prior experience, familiarity with the submission, and the level of reviewer engagement, likely shape ACs’ needs. While previous AI tools

for ACs have primarily focused on scaffolding novices (Sun et al. 2024b), our findings point to an opportunity for personalized, context-aware assistance. Findings from Sect. 4.3 show that ACs' preferences for particular wireframes were often grounded in how well these designs aligned with their existing workflows. For instance, the argument-labeling interface was positively received because it supports the structuring work ACs already do, without replacing their final judgment. Accordingly, future designs should account for ACs' established work practices and seek to integrate AI support in ways that complement, rather than disrupt, these workflows. This need for alignment was also evident in discussions of the comment-context interface, where ACs emphasized the importance of integrating such features into their existing reading practices. This observation is consistent with other studies on AI assistance (Tankelevitch et al. 2024; Zhang et al. 2024; Jacobs et al. 2021a), which show that misalignment between AI assistance and users' decision-making processes can lead to ineffective use or even abandonment of the system.

Drawing on findings from the different wireframes, especially 'conformity to standard', we argue that AI assistance for ACs should strike a balance between being simple yet informative, allowing ACs to get an overview at a glance while also providing explanations of how the results were generated on demand. This finding resonates with explainable AI assistance systems used in clinical settings, where explanations should be selective and mutable (Corti et al. 2024; Bertrand et al. 2023), allowing decision-makers to tweak the granularity of the information.

The discussion on the summarization tool revealed a divide between participants who valued its efficiency gains and those who expressed concern about over-reliance. To balance these perspectives, it is critical to ensure ACs engage critically with AI-generated content. This could be by restricting access to AI-generated content until users have formed their initial judgments, as it was shown to promote active engagement (Fogliato et al. 2022) or through personalization mechanisms by enabling users to toggle systems' features and adjust their output to their preferences (Zhang et al. 2023). All these potential personalization features should transparently communicate what historical data is used and offer flexible controls that allow ACs to adjust or limit data usage according to their privacy preferences (Meng et al. 2018; Rezk 2023; Choi and Lee 2026).

5.2 Facilitating Reviewer Coordination (RQ1)

Our findings reveal that ACs face persistent coordination challenges, including chasing reviewers for timely responses, managing uneven participation, and navigating disagreements to reach consensus. These dynamics suggest that current communication channels are insufficient for fostering productive discussions and timely decision-making. As discussed in Sect. 4.3, ACs found this coordination task the most frustrating, noting that it did not require their seniority. Prior research on group decision-making highlights the importance of structured facilitation, such as the "diamond of participation" framework (Kaner 2014), which encourages divergence of ideas before converging toward consen-

sus. Similar challenges have been addressed in online deliberation settings, where conversational agents have successfully moderated discussions by balancing contributions across participants, promoting topic coverage, and guiding groups toward consensus (Bagmar et al. 2022; Kim et al. 2020, 2021). These strategies have been shown to enhance deliberative quality and align group consensus with individual opinions, leading to more authentic outcomes.

Design Implications: Develop Conversational Agents to Act as Moderators

Building on these insights, we see an opportunity to explore conversational assistant moderators for peer review. Such an assistant could help ACs by structuring reviewer discussions, encouraging participation from less active reviewers, and preventing dominance by a few voices (Bagmar et al. 2022; Kim et al. 2020). By offloading coordination tasks to the assistant, ACs could focus on higher-level judgment and decision-making, aligning their efforts with their expertise and reducing feelings of frustration and undervaluation (expressed in Sect. 4.1). Beyond facilitation, the assistant could visualize discussion stages, highlight key points of agreement and disagreement, and recommend contextually appropriate moderator messages (Lee et al. 2020a), as suggested by ACs in Sect. 4.3. However, deploying such an agent requires caution. Prior work shows that errors or poorly timed interventions by conversational agents can significantly undermine participants' willingness to engage (Do et al. 2023; Kuang et al. 2024). For example, if a conversational agent incorrectly flags a reviewer as unresponsive, reviewers may feel undervalued, leading to disengagement. Conversely, failing to identify unresponsive reviewers may incentivize them to game the system and avoid detection. Other repair strategies could provide explanations, for example, by notifying reviewers when an author has responded and a specified amount of time has passed without a reviewer reply (Ashktorab et al. 2019).

5.3 Tensions Underlying Informed Caution Toward AI Support (RQ2)

Our findings reveal a persistent cautiousness among ACs toward AI-based support tools. Participants recognized the potential of AI tools to streamline decision-making, while cautioning that model inaccuracies could undermine the fairness of final decisions. There was a strong consensus that human oversight must remain central to peer review. This cautious behavior may be explained by two phenomena documented in prior research. First, decision-makers with higher AI literacy tend to exhibit greater skepticism toward AI recommendations compared to those with lower familiarity (Jacobs et al. 2021b; Burgess et al. 2023). Similarly, our participants are AI experts and are acutely aware of risks and unintended consequences, which likely shaped their cautious stance (as stated by P1 in Sect. 4.3). Second, prior work consistently emphasizes the importance of safeguarding human agency: decision-makers are more likely to adopt AI tools when these tools assist rather than replace their judgment (Corti et al. 2024; Fogliato et al. 2022; Kawakami et al. 2022a; Ma et al. 2023; Salimzadeh, He, and Gadiraju 2023). Our findings reinforce this principle in the context of peer

review, highlighting the need for AI systems that augment rather than automate AC decision-making.

Design Implications: Implement Human-Centered AI Principles that Safeguard Human Agency Participants’ caution toward AI and their emphasis on human agency have important implications for the design of support tools for ACs. Rather than automating or mimicking the AC role, future systems should focus on augmenting human judgment while maintaining transparency and control. This means ensuring that AI outputs are interpretable and traceable to original sources. This aligns with prior research showing that providing explanations fosters trust and accountability in AI-assisted decision-making (Fogliato et al. 2022; Bućinca, Malaya, and Gajos 2021; Veale, Van Kleek, and Binns 2018; Cobbe, Lee, and Singh 2021; Holstein et al. 2023). For example, extractive summaries that preserve reviewers’ arguments should be favored over abstractive summaries that risk introducing distortions. Unlike authors, who appreciated positively framed summaries to help them accept criticism (Yang et al. 2025), ACs prefer avoiding distortions introduced by abstractive summaries. Support tools should also help ACs identify overlooked aspects of reviews and discussions. For example, comparing meta-reviews with individual reviews could highlight inconsistencies or gaps that may otherwise go unnoticed. This resonates with recent work on human–AI collaboration in peer review, where AI was seen as useful for broadening perspectives and reducing reviewers’ blind spots (Chen, Brumby, and Cox 2025). Most importantly, future designs must maintain ACs’ active engagement and ensure they remain the final decision-makers to prevent over-reliance, for example, by carefully timing when assistance is introduced (Fogliato et al. 2022), by highlighting key arguments rather than abstracting content in summaries (Solomon 2014), and surfacing relevant passages from reviewer–author discussions to facilitate conflict mediation (De-Arteaga, Fogliato, and Chouldechova 2020).

5.4 Venue-Specific Dynamics and Broader Implications

While participants offered valuable insights, they do not represent the full diversity of peer-review practices across venues. Accordingly, we differentiate generalizable patterns from those specific to ICLR.

What likely generalizes. Many challenges highlighted by ACs, most notably the increasing number of submissions, the scarcity of expert reviewers, and the resulting cognitive load on ACs, are not unique to ICLR. Similar pressures have been documented across conference and journal ecosystems, including AAAI, AIES, and adjacent research communities (Ellison 2025; AIES 2025; Stefanidi et al. 2023). In response, ACs described developing personal heuristics and lightweight methods for interpreting and reconciling reviewers’ and authors’ arguments, a strategy that plausibly translates to other venues. Likewise, the wireframe concepts we explored are designed to streamline sense-making and may help reduce cognitive burden independently of venue-specific norms. As such, these designs are promising options for broader adaptation.

What appears venue-specific. Dynamics observed in our study reflect ICLR’s particular review context, include lengthy reviewer–author discussions during rebuttals, the public nature of reviews, and the use of the OpenReview platform. Together, these characteristics can amplify tensions between reviewers and authors and required increased coordination and moderation work. These dynamics may be less pronounced in venues where reviewer–author interaction is more limited. Additionally, because our participants are AI experts, their skepticism regarding model reliability, bias, and inaccuracies may be more pronounced than that of ACs in other domains. This expertise likely sharpened critiques of prediction-driven features and raised expectations around personalization, transparency, and control.

Implications for transfer to other domains. Taken together, these observations suggest that AI support for AC work should be generalizable but adaptable: core assistance for coordination and sense-making (e.g., argument structuring, summarization) can generalize, while levels of automation and interpretability should be adapted to venue norms, and personalization features adjusted to individual AC preferences. The cautious stance voiced by ICLR ACs can be read as a safety-first baseline. Even if other venues exhibit more moderate skepticism, the concerns and desired features identified here (e.g., preserving human agency, enabling personalization, and offering explanations) represent conservative defaults likely to be appropriate across domains. Finally, we observed substantial variability in how ACs approach their work. While we expect this variability to extend beyond ICLR, future research is needed to empirically examine how it manifests across different venues.

6 Conclusion and Future Work

In this paper, we investigate ACs’ challenges, strategies, and perceptions of AI assistance. Our analysis reveals that ACs vary widely in how much they engage with submissions and in their level of involvement in the outcomes. While some ACs adopt a hands-off approach and prefer to remain impartial, others were hands-on, reading submissions and actively advocating for promising ideas. Through a design probe with wireframes, we found that they approached AI support cautiously: while they recognized its potential to reduce cognitive overload and make decisions faster, they were concerned about the risks of over-reliance, misinterpretation, and bias amplification. From these insights, we distilled three design considerations: personalized assistance aligned with ACs’ diverse workflows; conversational agents to act as moderators between reviewers and authors; and human-centered AI principles that safeguard human agency. By centering ACs’ lived practices, this study provides an empirical foundation for assistive, accountable AI in peer review. Future research could investigate interactive prototypes that accommodate diverse needs and review stages. Further, ACs’ experience using these tools over time should be understood using a longitudinal study.

Acknowledgments

We are grateful to the area chairs who participated in this study for their time and insights. The research was undertaken thanks in part to funding from the Connected Minds Program, supported by the Canada First Research Excellence Fund, Grant #CFREF-2022-00010, and the Canada Research Chairs Program. We also acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant #RGPIN-2026-06542 and #RGPIN-2026-06657. This work was also supported in part by the National Science Foundation (NSF) under award number 2244805, as well as a FRQNT Master's (B1) research scholarship, a FRQNT Doctoral (B2X) research scholarship, and a NSERC Post-graduate doctoral (PGS D) scholarship. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.

References

- AAAI. 2025. AAAI-26 Review Process Update: Scale, Integrity Measures, and Pathways to Sustainability. <https://aaai.org/conference/aaai/aaai-26/review-process-update/>. Accessed: 2026-05-12.
- AIES. 2025. The Program Guide For The Eighth AAAI/ACM Conference on Artificial Intelligence Ethics & Society. <https://www.aies-conference.com/2025/wp-content/doc/AIES-2025-Program-10.20.pdf>. Accessed: 2026-05-12.
- Ailamaki, A.; Chrysogelos, P.; Deshpande, A.; and Kraska, T. 2019. The SIGMOD 2019 Research Track Reviewing System. *ACM SIGMOD Record*, 48(2): 47–54.
- Arous, I.; Yang, J.; Khayati, M.; and Cudre-Mauroux, P. 2021. Peer Grading the Peer Reviews: A Dual-Role Approach for Lightening the Scholarly Paper Review Process. In *Proceedings of the Web Conference 2021*, WWW '21, 1916–1927. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383127.
- Ashktorab, Z.; Jain, M.; Liao, Q. V.; and Weisz, J. D. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–12. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-5970-2.
- Bagmar, A.; Hogan, K.; Shalaby, D.; and Purtilo, J. 2022. Analyzing the Effectiveness of an Extensible Virtual Moderator. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP): 1–16.
- Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, 1–16. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8096-6.
- Bertrand, A.; Viard, T.; Belloum, R.; Eagan, J. R.; and Maxwell, W. 2023. On Selective, Mutable and Dialogic XAI: a Review of What Users Say about Different Types of Interactive Explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–21. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9421-5.
- Biermann, O. C.; Ma, N. F.; and Yoon, D. 2022. From Tool to Companion: Storywriters Want AI Writers to Respect Their Personal Values and Writing Strategies. In *Designing Interactive Systems Conference*, 1209–1227. Virtual Event Australia: ACM. ISBN 978-1-4503-9358-4.
- Biswas, J.; Schoepp, S.; Vasan, G.; Pipari, A.; Zhang, A.; Hu, Z.; Joseph, S.; Lease, M.; Li, J. J.; Stone, P.; Wagstaff, K. L.; Taylor, M. E.; and Jenkins, O. C. 2026. AI-Assisted Peer Review at Scale: The AAAI-26 AI Review Pilot. ArXiv:2604.13940 [cs].
- Braun, V.; and Clarke, V. 2021. *Thematic Analysis: A Practical Guide*. SAGE. ISBN 978-1-5264-1729-9. Google-Books-ID: eMArEAAAQBAJ.
- Burgess, E. R.; Jankovic, I.; Austin, M.; Cai, N.; Kapuścińska, A.; Currie, S.; Overhage, J. M.; Poole, E. S.; and Kaye, J. 2023. Healthcare AI Treatment Decision Support: Design Principles to Enhance Clinician Adoption and Trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–19. Hamburg Germany: ACM. ISBN 978-1-4503-9421-5.
- Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 188:1–188:21.
- Cabitz, F.; Campagner, A.; Angius, R.; Natali, C.; and Reverberi, C. 2023. AI Shall Have No Dominion: on How to Measure Technology Dominance in AI-supported Human decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20. Hamburg Germany: ACM. ISBN 978-1-4503-9421-5.
- Cai, C. J.; Winter, S.; Steiner, D.; Wilcox, L.; and Terry, M. 2019. “Hello AI”: Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 104:1–104:24.
- Chandrasekharan, E.; Gandhi, C.; Mustelier, M. W.; and Gilbert, E. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 174:1–174:30.
- Checco, A.; Bracciale, L.; Loreti, P.; Pinfield, S.; and Bianchi, G. 2021. AI-assisted peer review. *Humanities and Social Sciences Communications*, 8(1): 1–11.
- Chen, S.; Brumby, D.; and Cox, A. 2025. Envisioning the Future of Peer Review: Investigating LLM-Assisted Reviewing Using ChatGPT as a Case Study. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, CHIWORK '25, 1–18. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1384-2.
- Cheng, H.-F.; Stapleton, L.; Kawakami, A.; Sivaraman, V.; Cheng, Y.; Qing, D.; Perer, A.; Holstein, K.; Wu, Z. S.; and

- Zhu, H. 2022. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *CHI Conference on Human Factors in Computing Systems*, 1–22. New Orleans LA USA: ACM. ISBN 978-1-4503-9157-3.
- Choi, S.-Y.; and Lee, K. 2026. Human-Centered Personalization in Radiology AI: Evaluating Trust, Usability, and Cross-Hospital Robustness. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, 1–23. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-2278-3.
- Cobbe, J.; Lee, M. S. A.; and Singh, J. 2021. Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 598–609. Virtual Event Canada: ACM. ISBN 9781450383097.
- Corti, L.; Olthmans, R.; Jung, J.; Balayn, A.; Wijsenbeek, M.; and Yang, J. 2024. “It Is a Moving Process”: Understanding the Evolution of Explainability Needs of Clinicians in Pulmonary Medicine. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–21. Honolulu HI USA: ACM. ISBN 979-8-4007-0330-0.
- Darrin, M.; Arous, I.; Piantanida, P.; and Cheung, J. 2024. GLIMPSE: Pragmatically Informative Multi-Document Summarization for Scholarly Reviews. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12737–12752. Bangkok, Thailand: Association for Computational Linguistics.
- De-Arteaga, M.; Fogliato, R.; and Chouldechova, A. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12. Honolulu HI USA: ACM. ISBN 9781450367080.
- Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology. General*, 144(1): 114–126.
- Do, H. J.; Kong, H.-K.; Tetali, P.; Lee, J.; and Bailey, B. P. 2023. To Err is AI: Imperfect Interventions and Repair in a Conversational Agent Facilitating Group Chat Discussions. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1): 99:1–99:23.
- D’Andrea, R.; and O’Dwyer, J. P. 2017. Can editors save peer review from peer reviewers? *PLOS ONE*, 12(10): e0186111.
- Ellison, M. 2025. AAAI Launches AI-Powered Peer Review Assessment System. <https://aaai.org/aaai-launches-ai-powered-peer-review-assessment-system/>. Accessed: 2026-08-07.
- Fecher, B.; Hebing, M.; Laufer, M.; Pohle, J.; and Sofsky, F. 2025. Friend or foe? Exploring the implications of large language models on the science system. *AI & SOCIETY*, 40(2): 447–459.
- Flach, P. A.; Spiegler, S.; Golénia, B.; Price, S.; Guiver, J.; Herbrich, R.; Graepel, T.; and Zaki, M. J. 2010. Novel tools to streamline the conference review process: experiences from SIGKDD’09. *ACM SIGKDD Explorations Newsletter*, 11(2): 63–67.
- Fogliato, R.; Chappidi, S.; Lungren, M.; Fisher, P.; Wilson, D.; Fitzke, M.; Parkinson, M.; Horvitz, E.; Inkpen, K.; and Nushi, B. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, 1362–1374. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9352-2.
- Fok, R.; Kambhamettu, H.; Soldaini, L.; Bragg, J.; Lo, K.; Hearst, M.; Head, A.; and Weld, D. S. 2023. Scim: Intelligent Skimming Support for Scientific Papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 476–490. Sydney NSW Australia: ACM. ISBN 979-8-4007-0106-1.
- Galipeau, J.; Barbour, V.; Baskin, P.; Bell-Syer, S.; Cobey, K.; Cumpston, M.; Deeks, J.; Garner, P.; MacLehose, H.; Shamseer, L.; Straus, S.; Tugwell, P.; Wager, E.; Winker, M.; and Moher, D. 2016. A scoping review of competencies for scientific editors of biomedical journals. *BMC Medicine*, 14(1): 16.
- Ghosal, T.; Verma, R.; Ekbal, A.; and Bhattacharyya, P. 2019. DeepSentiPeer: Harnessing Sentiment in Review Texts to Recommend Peer Review Decisions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1120–1130. Florence, Italy: Association for Computational Linguistics.
- Green, B.; and Chen, Y. 2019a. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* ’19, 90–99. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6125-5.
- Green, B.; and Chen, Y. 2019b. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–24.
- Grgić-Hlača, N.; Engel, C.; and Gummadi, K. P. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 178:1–178:25.
- Hall, J.; Hundley, V.; and van Teijlingen, E. 2015. The journal editor: Friend or foe? *Women and Birth*, 28(2): e26–e29.
- Han, J.; Park, J.; Huh, J.; Oh, U.; Do, J.; and Kim, D. 2024. AscleAI: A LLM-based Clinical Note Management System for Enhancing Clinician Productivity. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7. Honolulu HI USA: ACM. ISBN 979-8-4007-0331-7.
- Head, A.; Lo, K.; Kang, D.; Fok, R.; Skjonsberg, S.; Weld, D. S.; and Hearst, M. A. 2021. Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. In *Proceedings of the 2021 CHI Conference*

- on *Human Factors in Computing Systems*, 1–18. Yokohama Japan: ACM. ISBN 978-1-4503-8096-6.
- Hicks, C. M.; Pandey, V.; Fraser, C. A.; and Klemmer, S. 2016. Framing Feedback: Choosing Review Environment Features that Support High Quality Peer Assessment. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 458–469. San Jose California USA: ACM. ISBN 978-1-4503-3362-7.
- Holstein, K.; De-Arteaga, M.; Tumati, L.; and Cheng, Y. 2023. Toward Supporting Perceptual Complementarity in Human-AI Collaboration via Reflection on Unobservables. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 152:1–152:20.
- Hua, X.; Nikolov, M.; Badugu, N.; and Wang, L. 2019. Argument Mining for Understanding Peer Reviews. In *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)*, 2131–2137. Minneapolis, Minnesota: Association for Computational Linguistics.
- Jacobs, M.; He, J.; F. Pradier, M.; Lam, B.; Ahn, A. C.; McCoy, T. H.; Perlis, R. H.; Doshi-Velez, F.; and Gajos, K. Z. 2021a. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. Yokohama Japan: ACM. ISBN 9781450380966.
- Jacobs, M.; Pradier, M. F.; McCoy, T. H.; Perlis, R. H.; Doshi-Velez, F.; and Gajos, K. Z. 2021b. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry*, 11(1): 108.
- Kaner, S. 2014. *Facilitator's Guide to Participatory Decision-Making*. John Wiley & Sons. ISBN 978-1-118-40495-9.
- Kawakami, A.; Sivaraman, V.; Cheng, H.-F.; Stapleton, L.; Cheng, Y.; Qing, D.; Perer, A.; Wu, Z. S.; Zhu, H.; and Holstein, K. 2022a. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, 1–18. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9157-3.
- Kawakami, A.; Sivaraman, V.; Stapleton, L.; Cheng, H.-F.; Perer, A.; Wu, Z. S.; Zhu, H.; and Holstein, K. 2022b. “Why Do I Care What’s Similar?” Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference*, 454–470. Virtual Event Australia: ACM. ISBN 978-1-4503-9358-4.
- Kennard, N. N.; O’Gorman, T.; Das, R.; Sharma, A.; Bagchi, C.; Clinton, M.; Yelugam, P. K.; Zamani, H.; and McCallum, A. 2022. DISAPERE: A Dataset for Discourse Structure in Peer Review Discussions. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1234–1249. Seattle, United States: Association for Computational Linguistics.
- Kim, S.; Eun, J.; Oh, C.; Suh, B.; and Lee, J. 2020. Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–13. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6708-0.
- Kim, S.; Eun, J.; Seering, J.; and Lee, J. 2021. Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–26.
- Kuang, E.; Li, M.; Fan, M.; and Shinohara, K. 2024. Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–16. Honolulu HI USA: ACM. ISBN 979-8-4007-0330-0.
- Lee, S.-C.; Song, J.; Ko, E.-Y.; Park, S.; Kim, J.; and Kim, J. 2020a. SolutionChat: Real-time Moderator Support for Chat-based Structured Discussion. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–12. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6708-0.
- Lee, W.; Kwon, M.; Hyun, Y.; Lee, J.; Gwon, J.; and Jung, H. 2020b. Uncovering CHI Reviewers Needs and Barriers. In *Proceedings of the 2020 Symposium on Emerging Research from Asia and on Asian Contexts and Cultures*, 57–60. Honolulu HI USA: ACM. ISBN 978-1-4503-8768-2.
- Li, M.; Hovy, E.; and Lau, J. 2023. Summarizing Multiple Documents with Conversational Structure for Meta-Review Generation. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 7089–7112. Singapore: Association for Computational Linguistics.
- Li, M.; Lau, J. H.; and Hovy, E. 2024. A Sentiment Consolidation Framework for Meta-Review Generation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10158–10177. Bangkok, Thailand: Association for Computational Linguistics.
- Ma, S.; Lei, Y.; Wang, X.; Zheng, C.; Shi, C.; Yin, M.; and Ma, X. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 1–19. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9421-5.
- Meng, X.; Wang, S.; Shu, K.; Li, J.; Chen, B.; Liu, H.; and Zhang, Y. 2018. Personalized Privacy-Preserving Social Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Miley, Jill. 2025. 13th Annual International Conference on Learning Representations (ICLR) 2025 Fact Sheet. <https://>

//media.iclr.cc/Conferences/ICLR2025/ICLR2025_Fact_Sheet.pdf. Accessed: 2026-05-12.

Ngoon, T. J.; Fraser, C. A.; Weingarten, A. S.; Dontcheva, M.; and Klemmer, S. 2018. Interactive Guidance Techniques for Improving Creative Feedback. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–11. Montreal QC Canada: ACM. ISBN 978-1-4503-5620-6.

Norkute, M.; Herger, N.; Michalak, L.; Mulder, A.; and Gao, S. 2021. Towards Explainable AI: Assessing the Usefulness and Impact of Added Explainability Features in Legal Document Summarization. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7. Yokohama Japan: ACM. ISBN 978-1-4503-8095-9.

Pu, K.; Feng, K. J. K.; Grossman, T.; Hope, T.; Dalvi Mishra, B.; Latzke, M.; Bragg, J.; Chang, J. C.; and Siangliulue, P. 2025. IdeaSynth: Iterative Research Idea Development Through Evolving and Composing Idea Facets with Literature-Grounded Feedback. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–31. Yokohama Japan: ACM. ISBN 979-8-4007-1394-1.

Reinhart, M.; and Schendzielorz, C. 2024. Peer-review procedures as practice, decision, and governance—the road to theories of peer review. *Science and Public Policy*, 51(3): 543–552.

Resnik, D. B.; and Elmore, S. A. 2016. Ensuring the Quality, Fairness, and Integrity of Journal Peer Review: A Possible Role of Editors. *Science and Engineering Ethics*, 22(1): 169–188.

Rezk, A. M. 2023. “Way too good and way beyond comfort”: The trade-off between user perception of benefits and comfort in media personalisation. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’23, 996–998. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0231-0.

Roth, W.-M. 2006. Editorial: On Editing and Being an Editor. *Cultural Studies of Science Education*, 1(2): 209–217.

Salimzadeh, S.; He, G.; and Gadiraju, U. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP ’23, 215–227. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9932-6.

Solomon, J. 2014. Customization bias in decision support systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3065–3074. Toronto Ontario Canada: ACM. ISBN 9781450324731.

Stefanidi, E.; Bentvelzen, M.; Woźniak, P. W.; Kosch, T.; Woźniak, M. P.; Mildner, T.; Schneegass, S.; Müller, H.; and Niess, J. 2023. Literature Reviews in HCI: A Review of Reviews. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23, 1–24. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9421-5.

Sun, L.; Chan, A.; Chang, Y. S.; and Dow, S. P. 2024a. ReviewFlow: Intelligent Scaffolding to Support Academic Peer

Reviewing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 120–137. Greenville SC USA: ACM. ISBN 979-8-4007-0508-3.

Sun, L.; Tao, S.; Hu, J.; and Dow, S. P. 2024b. MetaWriter: Exploring the Potential and Perils of AI Writing Support in Scientific Peer Review. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–32.

Tan, S.; Wu, J.; Bei, X.; and Xu, H. 2021. Least Square Calibration for Peer Reviews. In *Advances in Neural Information Processing Systems*, volume 34, 27069–27080. Curran Associates, Inc.

Tankelevitch, L.; Kewenig, V.; Simkute, A.; Scott, A. E.; Sarkar, A.; Sellen, A.; and Rintel, S. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–24. Honolulu HI USA: ACM. ISBN 979-8-4007-0330-0.

Tran, D.; Valtchanov, A.; Ganapathy, K.; Feng, R.; Slud, E.; Goldblum, M.; and Goldstein, T. 2020. An Open Review of OpenReview: A Critical Analysis of the Machine Learning Conference Review Process. ArXiv:2010.05137 [cs].

Veale, M.; Van Kleek, M.; and Binns, R. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. Montreal QC Canada: ACM. ISBN 9781450356206.

Wang, K.; and Wan, X. 2018. Sentiment Analysis of Peer Review Texts for Scholarly Papers. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 175–184. Ann Arbor MI USA: ACM. ISBN 978-1-4503-5657-2.

Wang, R.; Harper, F. M.; and Zhu, H. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI ’20, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6708-0.

Wilson, M. L.; and Nacke, L. 2022. How to: Peer Review for CHI (and Beyond). In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA ’22, 1–4. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9156-6.

Yang, C.-L.; Uhde, A.; Yamashita, N.; and Kuzuoka, H. 2025. Understanding and Supporting Peer Review Using AI-reframed Positive Summary. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, 1–16. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1394-1.

Yang, Q.; Hao, Y.; Quan, K.; Yang, S.; Zhao, Y.; Kuleshov, V.; and Wang, F. 2023. Harnessing Biomedical Literature to Calibrate Clinicians’ Trust in AI Decision Support Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–14. Hamburg Germany: ACM. ISBN 9781450394215.

Yen, Y.-C. G.; Dow, S. P.; Gerber, E.; and Bailey, B. P. 2017. Listen to Others, Listen to Yourself: Combining Feedback Review and Reflection to Improve Iterative Design. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition, C&C '17*, 158–170. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-4403-6.

Yen, Y.-C. G.; Kim, J. O.; and Bailey, B. P. 2020. Decipher: An Interactive Visualization Tool for Interpreting Unstructured Design Feedback from Multiple Providers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20*, 1–13. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6708-0.

Yudhanto, Y.; Pryhatyanto, W. M.; and Sulandari, W. 2022. Designing and Making UI/UX Designs on The Official Website with The Design Thinking Method. In *2022 1st International Conference on Smart Technology, Applied Informatics, and Engineering (APICS)*, 165–170.

Zaidi, A.; and Karahalios, K. 2025. From Sociotechnical Gaps to Solutions: Designing AI Tools with Parents to Address Special Education Advocacy Barriers in IEP Processes. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, 2619–2636. Madeira Portugal: ACM. ISBN 979-8-4007-1485-6.

Zeng, Q.; Sidhu, M.; Blume, A.; Chan, H. P.; Wang, L.; and Ji, H. 2025. Scientific Opinion Summarization: Paper Meta-review Generation Dataset, Methods, and Evaluation. In Yin, W.; Ahn, J. J.; Zhang, R.; Huang, L.; Hadfi, R.; Ito, T.; Ohnuma, S.; and Shiramatsu, S., eds., *Artificial Intelligence for Research and Democracy*, 20–38. Singapore: Springer Nature. ISBN 978-981-97-9536-9.

Zhang, S.; Yu, J.; Xu, X.; Yin, C.; Lu, Y.; Yao, B.; Tory, M.; Padilla, L. M.; Caterino, J.; Zhang, P.; and Wang, D. 2024. Rethinking Human-AI Collaboration in Complex Medical Decision Making: A Case Study in Sepsis Diagnosis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24*, 1–18. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0330-0.

Zhang, X.; Li, J.; Chi, P.-W.; Chandrasegaran, S.; and Ma, K.-L. 2023. ConceptEVA: Concept-Based Interactive Exploration and Customization of Document Summaries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16. Hamburg Germany: ACM. ISBN 978-1-4503-9421-5.

Zhang, Y.; Bellamy, R. K.; and Kellogg, W. A. 2015. Designing Information for Remediating Cognitive Biases in Decision-Making. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2211–2220. Seoul Republic of Korea: ACM. ISBN 9781450331456.