Do AI Conferences' Ethics Reviews Steer Research Practices?

Kento Nishi
Harvard University
Cambridge, MA, USA
kentonishi@college.harvard.edu

Alec Laprevotte
Harvard University
Cambridge, MA, USA
alaprevotte@g.harvard.edu

Isaiah Bullock
Harvard University
Cambridge, MA, USA
isaiahabbey@college.harvard.edu

Mfoniso M. Andrew
Harvard University
Cambridge, MA, USA
mfonisoandrew@college.harvard.edu

Abstract

Selective conferences in artificial intelligence and machine learning now require ethics statements, impact statements, and review flags to steer research practice through peer review. When reviewers raise ethical concerns and a paper is rejected or withdrawn, authors typically revise the paper or project and submit elsewhere; in that next venue, the prior critique may lead authors to change the work, repackage it, or face a new filtering process. The later version may therefore show no response, substantive change (data, evaluation, scope, release decisions), or rhetorical change (expanded limitations and harms framing without changing the technical core). To study when ethics review does and does not steer research, we trace ICLR submissions from 2021 to 2026 that received ethics flags, examining the reviews themselves, any revisions, and whether subsequent resubmissions to later venues addressed the concerns raised. Across 47,893 submissions and 185,194 public reviews, we identify 2,498 submissions with reviewer ethics flags, including 1,857 that were rejected or withdrawn. Using retrieval through OpenAlex, we produced 6,618 candidate matches to later papers and selected 601 potential candidates for successors. In initial screening of these candidates, we judged 459 pairs likely to describe the same project or a direct descendant; within this screened set, many later papers show no observable change or mainly rhetorical change, while a smaller subset shows procedural and methodological change. Early interviews from a stratified sample of 25 cases further suggest that reviewer inattentiveness, venue fit, and filtering can make reviewer ethics critiques feel less actionable. We released our code here.

Keywords

AI ethics, peer review, ethics review, research governance, OpenReview

1 Introduction

In AI and machine learning, selective conferences are primary venues for publishing new research. Researchers accordingly aim to publish their strongest work at venues such as NeurIPS, ICML, and ICLR, with others often building their next papers on work accepted there. Because acceptance and rejection at these venues affect which projects attract attention and which research directions receive sustained investment, organizers have increasingly extended peer review to include ethical reflection alongside technical evaluation. By the early 2020s, broader impact statements, ethics statements, and reviewer ethics flags had become part of review at these venues [2, 6, 7, 9]. Under these arrangements, reviewers could raise privacy, fairness, safety, misuse, or deployment concerns alongside technical objections. In many cases, those concerns depended on specific properties of the system under review. For example, Staab et al. [10] show that large language models can infer sensitive personal attributes such as location, income, and sex from real Reddit profiles.

When reviewers criticize a paper on grounds like these, the authors usually do not resubmit their work to the same venue. Instead, they commonly revise the paper, the project, or its presentation and submit it to another conference or journal, where a different reviewer pool evaluates the paper from scratch. Between venues, authors may replace a dataset or add an evaluation, which changes the evidence offered in support of their claims. They may also narrow a claim or change whether code or model weights will be released, which changes the scope of the work and the conditions under which other researchers can use it. In other cases, they may expand a limitations section, soften claims, or elaborate possible harms, affecting how the work is presented to the next reviewers even when the underlying research process remains the same. These choices respond to criticism in different ways: revising data changes the research itself, revising harms and limitations reframes the work for new readers, while leaving a concern unaddressed preserves the original research and framing. Since reviewers usually see a paper after much of the technical work is already in place but before the authors finalize the structure of their final revision, the period between one decision and the next submission offers a direct view of what authors chose to change, what they chose to repackage, and what they chose to leave untouched.

OpenReview makes this kind of comparison possible because it publishes the submission, the reviews, the decision, and the associated ethics flags in one public archive. This archive lets us read a paper and its attached criticism concurrently, and once we trace the next public version of the same project, we evaluate how exactly earlier criticism led to change in a later submission.

Researchers have used these records to study ethics review inside a single peer review cycle. For example, Liu et al. [3] analyze 96 ethics reviews across 50 NeurIPS 2021 papers, the NeurIPS 2021 retrospective reports conditional acceptances and one rejection on ethical grounds, and ICLR 2023 program chairs report screening 190 papers whose reviews included ethics flags before sending a smaller subset for further review [2, 3, 7]. However, these studies focus on a single review cycle and typically do not follow projects into the next submission, when authors decide what a new set

of reviewers will see and whether the prior critique seems worth acting on. We follow projects first submitted to ICLR whose reviews included ethics flags into later public versions, report preliminary screening results on pairs of original and later papers, and conduct interviews with a stratified sample of 25 corresponding authors to clarify why projects changed (or did not) between venues and how authors interpreted prior review.

2 Related Work

Prunkl et al. [9] describe broader impact statements as an attempt to incorporate ethical reflection into ordinary peer review, and they also explain why that arrangement may produce limited change: reviewers may not share standards or domain knowledge, authors may face weak incentives, and ethically salient design choices may already be fixed by the time a paper reaches review. Those timing constraints also reappear in work on what authors actually wrote under these requirements. For instance, Nanayakkara et al. [5] analyze how NeurIPS authors described societal consequences, including how specific those descriptions were and how they handled uncertainty, and Ashurst et al. [1] extend that analysis at larger scale by showing that authors engaged the requirement unevenly. Taken together, these studies help us read later additions about harms, limitations, and uncertainty as changes in how authors present a paper to readers. They do not yet tell us, though, whether the authors changed the underlying research in the same way.

Liu et al. [3] are more directly aligned with our question in their use of public NeurIPS records to analyze how reviewers and authors assigned responsibility for harm and what kinds of revisions reviewers requested. Their analysis, however, remains within a single peer review cycle and does not follow projects into subsequent submissions, where authors decide what to revise, what to defer, which reviewer judgments to trust, what to reframe, and what to leave untouched for a new reviewer pool. That later stage is also shaped by practical constraints documented in studies of AI work itself. For example, Pant et al. [8] synthesize studies of applied AI practice and report repeated gaps between ethical concern and concrete revision in data practices, evaluation choices, and deployment plans, while Morley et al. [4] describe similar tensions between general ethical principles and the constraints of daily engineering work. We extend this line of work by comparing criticized submissions with the later papers their authors prepared for another venue, and by interviewing authors about what they changed and why.

3 Data and Methods

3.1 Building the Corpus

We built the corpus from ICLR because OpenReview publishes the submission, the reviews, the decision, and the associated ethics flags together. From ICLR submissions between 2021 and 2026, we harvested submissions and public replies, normalized titles, abstracts, author lists, decisions, review text, and ethics categories, and then identified papers whose reviews included ethics flags. To follow the conference's own review process rather than our own retrospective labeling, we used reviewer ethics flags instead of keyword search. We then restricted the corpus to rejected, withdrawn, or unresolved submissions. For accepted papers, the work had already passed through the review process in which reviewers raised the criticism. Rejected and withdrawn papers, by contrast, leave authors with a familiar choice in this field: revise the work and send it to another venue. Across the full corpus, we assembled 47,893 submissions and 185,194 public reviews; 2,498 papers had reviews that included ethics flags, and 1,857 of those papers entered the search for later papers. One boundary of this corpus is that the harvested OpenReview snapshot does not expose public ethics flags for ICLR 2021, so our cohort of papers with ethics flags begins in ICLR 2022.

3.2 Finding Later Versions

To find successors, we searched OpenAlex for later versions of each paper using full titles, title variants, author surnames, and salient abstract terms. Authors often changed titles, venues, coauthor lists, or project scope after rejection or withdrawal, so exact title matching would have missed many plausible later papers. We accordingly scored candidates by title similarity, author overlap, abstract overlap, and publication year, kept the candidate with the highest score for each paper, and saved the larger candidate list whenever the best match remained uncertain. The search produced 6,618 candidate matches to later papers in total and 601 top candidates (one per successfully matched paper that had been rejected or withdrawn). We therefore treat these best matches as a starting point for manual validation rather than as ground truth.

3.3 Reading Pairs and Coding Revision

Automated retrieval, however, cannot determine whether a later paper truly continues the same project, so successor validation requires reading. We conduct a preliminary screening of candidate pairs to support sampling and planning: for each matched pair, we record whether the later paper appears to be the same project or a direct descendant, and we describe the kinds of changes that are observable between versions (no change, mainly rhetorical change, procedural change, methodological change, or both). This screening will be followed with validation by human coders and checks across raters on a subset.

3.4 Interviewing Authors

Version comparison shows what changed, but we cannot reliably infer why authors changed one part of a project, left another untouched, or judged a review actionable. We therefore conduct interviews with a partly structured protocol, using corresponding authors (or designated contacts) from a stratified subset of cases. After receiving staff approval to recruit, we emailed all selected cases on April 14, 2026 and performed interviews beginning April 15. The interviews are designed to last approximately 30 minutes. Before each interview, we send the consent form and obtain verbal consent at the start of the call (with a separate prompt for permission to record audio).

For recruitment, we use publicly listed contact information associated with the successor paper (e.g., institutional webpages or arXiv metadata). Participation is voluntary, and participants may skip questions or stop at any time. To reduce risk of professional sensitivity, we store recordings and transcripts in storage with access controls, remove identifying information from transcripts for analysis, and report results in aggregate; if we quote remarks, we

anonymize them and review quotations to reduce the risk of identification. Audio recording is optional and, when enabled, is used only to support transcription and analysis.

To complement these interviews with authors, we are pursuing an interview with a former program chair who has overseen review processes related to ethics, to clarify how ethics flags are used in practice and what constraints chairs face when responding to flagged submissions and calibrating review standards. We will not ask for or report confidential deliberations about specific submissions.

To keep recruitment consistent and outreach short, we used a standardized email template that introduces the study at a high level, references the public paper pair we traced, and requests a brief call. The template has the following form (with identifying details omitted):

> Hello {First name},
> We are conducting a study on how ethics reviews at AI conferences shape papers after reviewers raise ethical concerns and a paper is rejected. We saw that your paper "{short paper title}" was rejected from ICLR {year} after reviewers raised ethical concerns, but the work is publicly available via {public venue or archive}. We would appreciate a brief video call (about 30 minutes) to learn about your rebuttal and resubmission experience, and what (if anything) is different between the ICLR submission and the public version.
> If you are willing, please choose a time at {scheduling link}. Participation is voluntary; you may decline or stop at any time.
> Sincerely, {research team}

4 Results

4.1 Corpus and Candidate Successors

The number of papers whose reviews included ethics flags rose from 107 in 2022 to 1,100 in 2026, and those reviews raised concerns about fairness and discrimination, privacy and safety, harmful applications, legal compliance, responsible research practice, and research integrity. From the 1,857 rejected or withdrawn papers in our corpus, our retrieval produced 601 top candidate successors. In initial screenings, we judged 459 of the 1,857 rejected or withdrawn cases likely to have a true successor in this set of best matches.

For interviews, we selected 25 cases using stratified quotas intended to cover distinct steering relationships and issue areas rather than to estimate population rates. Based on our preliminary screening notes, the sample includes 11 cases in which the successor appears to show especially clear, substantive changes; 6 cases in which changes relevant to ethical concerns were already present in the original submission and the successor largely reflects intrinsic research decisions; and 8 cases in which the flagged concern appears not to be taken up in the successor. We also balanced coverage across four issue areas (harmful applications: 4; fairness and discrimination: 9; responsible research practice: 6; privacy, security, and safety: 6).

4.2 What Changed in Later Versions

In our current initial screening of the 459 likely successor pairs, 217 successors show no observable change relative to the criticized submission, 153 show mainly rhetorical change (e.g., expanded limitations, harms framing, or responsible use discussion without technical revision), and a smaller subset shows substantive procedural or methodological change (57 cases with both procedural and methodological change, 14 with primarily procedural change, and 5 with primarily methodological change). We treat these labels as preliminary: they guide sampling and highlight where deeper reading, author accounts, and coding by multiple raters should focus, rather than serving as final measurements.

The substantive changes in our interview sampling set often fall into a few recurring categories. Some later papers document new dataset curation and quality control procedures (e.g., explicit leakage checks, majority voting, or volunteer annotation workflows for benchmark construction). Other cases add evaluation and defense analysis (e.g., ablations testing safety finetuning or other mitigations for backdoor and jailbreak work). Still others add documentation and detail relevant to governance (e.g., clearer descriptions of consent and IRB review in research with human subjects, dataset cleaning steps to remove identifiers in sensitive domains, or more explicit discussion of data provenance and licensing). Unaddressed cases, by contrast, often expand exposition or experiments, or tailor the presentation to a new venue, while leaving the concern flagged as ethical unaddressed in the successor text.

4.3 Case Interviews

In the interviews, we ask authors to reconstruct their rebuttal and resubmission experience, what (if anything) is different between the ICLR submission and the later public version, what feedback mattered or felt actionable, and what constrained revision before the next submission or release. We then ask directly about whether ethics review influenced technical decisions, evaluation choices, release decisions, or only the framing presented to a new reviewer pool. Together, these interviews are designed to clarify when rhetorical revisions reflect strategic reframing versus when they reflect constraints on changing systems that had already been implemented or distrust in the review process.

So far, 3 interviews (P1, P2, and P3) have been conducted, with 2 more interviews planned. Across these interviews, participants describe an initial rejection at ICLR followed by resubmission to a smaller conference where the work was eventually accepted, but they diverge in how much the paper changed and how they interpret the role of peer review.

P1 reports relatively limited substantive change between versions. He describes the later paper as "mostly similar," with revisions focused on "adding another table," incorporating "more metrics," and "fixing the major issues [...] mostly metric based." He also emphasizes frustration with reviewer attentiveness, noting that a reviewer "copied some of the weakness of the previous reviewer without checking if they were actually fixed," even referencing text that had already been revised. More broadly, he characterizes the process as primarily selective rather than constructive: "the role of peer reviewing [...] is mostly to filter out papers," and suggests that structural factors such as acceptance rate targets further reinforce this filtering dynamic.

By contrast, P2 describes more extensive revisions following rejection. Initially, P2's work was meant as an interdisciplinary introduction to highlight his field's overlap with ICLR's focus, but notes that "the paper wasn't written in a way that was clear enough" and that reviewers were "very focused on a particular aspect, [...] instead of seeing a larger picture." In response, the team "decided to just cut it out completely" (referring to the interdisciplinary section), rewrote the introduction for a CS audience, and recruited 5,000 participants for another annotation pass. Despite accepting many reviewer ethical concerns as "fair points", he emphasizes that "the criticism is really about my presentation" and venue fit, and that "I don't think [the] review process can actually help the research... it's really about an editorial process." At the same time, he highlights how this process shapes which contributions are visible: reviewers "have so much power shaping what stories get seen" and in his case, this meant that "it's not as interdisciplinary as you'd like it to be, because the CS reviewers chose to filter it", showing how rejection after reviewers raise ethical concerns can filter a contribution from major venues and influence which kinds of contributions can enter them.

Similarly, P3 describes revisions as mostly framing, with the introduction and abstract rewritten while the methods and results stayed essentially unchanged, and characterizes review as closer to a "coin flip" than a source of research direction.

4.4 Themes from Rebuttal Analysis

Alongside the interviews, we read reviews and author rebuttals for the 25 stratified cases in detail. Four themes emerge that extend the interview findings and surface dynamics not visible in the document-pair screening alone.

Firstly, engagement with ethics across panels is often uneven and limited. In most cases, only one reviewer on a panel of three to five raised ethical concerns. The remaining reviewers typically engaged the paper on technical grounds without acknowledging the flagged issue, even when it had been raised in a public comment. This pattern indicates that ethics critiques often enter the review process as dissenting opinions rather than a shared evaluative frame, which has consequences for how authors weigh it against other feedback in their revisions.

In other cases, domain mismatches and ethic critiques are conflated or confused. Several rebuttals reveal that what reviewers flagged as ethical concerns were entangled with disagreements over disciplinary norms and what counts as required technical framing. Authors often responded with lengthy explanations of foundational concepts they had deliberately omitted, arguing those details were peripheral to the contribution. It seemed that some ethics flags function partly as proxies for unmet expectations about completeness or scope, complicating the interpretation of the flag itself.

Authors also routinely opened responses with formal gratitude and submitted multiple detailed replies per reviewer comment, including partial concessions to critiques their later submissions did not in fact incorporate. Because reviewers control the outcome, rebuttals at times appear oriented toward demonstrating responsiveness rather than registering disagreement, which therefore does not reliably predict substantive revision in the next version of the paper.

Finally, in at least one case, reviewers who did not flag ethical concerns assigned scores in the 4-5 range across categories while a reviewer who emphasized ethical considerations assigned a 6, indicating that ethical framing can shift evaluation in either direction and that reviewers apply inconsistent standards when ethics enters the assessment. We treat this as illustrative rather than representative pending broader coding.

5 Discussion

Our preliminary screening suggests that rejected or withdrawn projects whose submissions had ethics flags often reappear with limited observable change, and that when changes do appear they are frequently concentrated in framing, limitations, and responsible use discussion rather than in core methods. Accordingly, our study design combines manual pair validation with author interviews to characterize which kinds of ethics concerns lead to substantive changes versus rhetorical changes, and what constraints and review experiences authors report when they choose not to revise the research itself. A plausible hypothesis motivating this design is that reviewer ethics critiques are most impactful when they target aspects of a project that authors can still change before the work goes to a different reviewer pool (e.g., scope claims, release plans, or targeted evaluation additions), while changes that require rebuilding a data pipeline or restarting large experimental programs may be less feasible under time and resource constraints or when critique does not appear specific, competent, and actionable.

5.1 Reviewer Attentiveness and Filtering

The interviews also shift the analysis toward reviewer inattentiveness during the reviewing process. Initially, the project was focused on rhetoric versus practical change in researchers' work, and through the interviews, that question expanded to include whether reviewers gave authors guidance they could trust. For example, when asked about the reviewing process when submitting to UAI after being declined by ICLR, P1 shared with us that:

> "When I resubmitted this paper, I fixed many things, and the reviewer basically found the ICLR submission, and copied some of the weakness of the previous reviewer without checking if they were actually fixed in the current submission"

P1 supported this by sharing that the UAI reviewer flagged the submission for the same reason the ICLR reviewer flagged, but the sentence that the reviewer criticized was not present in the newer submission where P1 made the necessary adjustments. As a result, P1 interpreted some flags as filtering signals rather than reliable ethical guidance.

That interpretation matters because authors are less likely to treat ethics criticism as steering when the review appears inattentive or mismatched to the paper.

6 Limitations

Our evidence covers only publicly available papers. Some rejected or withdrawn papers may have changed substantially without leaving a public trace that we could connect to the original submission,

and some projects may have split across multiple papers in ways that resist clean reconstruction. Authors may also incorporate feedback from multiple venues, internal lab discussion, legal review, or product constraints into later papers, so even a careful pairwise comparison cannot assign every revision to a single source of criticism. To address part of that uncertainty, we used interviews, but we still rely on retrospective accounts from people who agreed to speak with us, who may be especially critical of the review process, and who reconstructed decisions made under time pressure months or years earlier. At the same time, we reduce complicated revision histories to a small set of categories so that we can compare cases systematically. That simplification necessarily drops some detail. For these reasons, we treat our findings as evidence about common patterns in later revision after ethics criticism rather than as definitive causal accounts of every individual paper.

7 Conclusion

Selective AI conferences ask reviewers to evaluate ethical concern alongside technical merit, but the consequences of that criticism are only partially visible within a single review cycle. Our approach therefore follows rejected or withdrawn ICLR projects whose reviews included ethics flags into later public versions and uses both document comparisons and author interviews to understand what changed and why. In preliminary screening, many successors show no observable change or mostly rhetorical change, while a smaller subset show procedural and methodological revisions. Ongoing interviews are intended to clarify when those patterns reflect genuine constraints on revision, strategic choices about what to change and what to present to new reviewers, or distrust in review as ethical guidance.

The steering power of ethics review therefore depends not only on authors' willingness to revise, but also on whether the review system produces criticism that appears specific, competent, and actionable.

References

[1] Carolyn Ashurst, Emmie Hine, Paul Sedille, and Alexis Carlier. 2022. AI Ethics Statements: Analysis and Lessons Learnt from NeurIPS Broader Impact Statements. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2301–2313. doi:10.1145/3531146.3533231

[2] International Conference on Learning Representations. 2023. Ethics Review Process for ICLR 2023. https://blog.iclr.cc/2023/04/12/ethics-review-process-for-iclr-2023/. Accessed 2026-03-08.

[3] David Liu, Priyanka Nanayakkara, Sarah Ariyan Sakha, Grace Abuhamad, Su Lin Blodgett, Nicholas Diakopoulos, Jessica R. Hullman, and Tina Eliassi-Rad. 2022. Examining Responsibility and Deliberation in AI Impact Statements and Ethics Reviews. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 424–435. doi:10.1145/3514094.3534155

[4] Jessica Morley, Liz Kinsey, Ali Elhalal, et al. 2023. Operationalising AI Ethics: Barriers, Enablers and Next Steps. *AI & Society* 38 (2023), 411–423. doi:10.1007/s00146-021-01308-8

[5] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the Expressed Consequences of AI Research in Broader Impact Statements. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 795–806. doi:10.1145/3461702.3462608

[6] Neural Information Processing Systems Foundation. 2021. NeurIPS 2021 Call for Papers. https://neurips.cc/Conferences/2021/CallForPapers. Accessed 2026-03-08.

[7] NeurIPS. 2021. A Retrospective on the NeurIPS 2021 Ethics Review Process. https://blog.neurips.cc/2021/12/03/a-retrospective-on-the-neurips-2021-ethics-review-process/. Accessed 2026-03-08.

[8] A. Pant, R. Hoda, C. Tantithamthavorn, et al. 2024. Ethics in AI through the Practitioner's View: A Grounded Theory Literature Review. *Empirical Software Engineering* 29, 67 (2024). doi:10.1007/s10664-024-10465-5

[9] Carina E. A. Prunkl, Carolyn Ashurst, Markus Anderljung, Helena Webb, Jan Leike, and Allan Dafoe. 2021. Institutionalizing Ethics in AI through Broader Impact Requirements. *Nature Machine Intelligence* 3, 2 (2021), 104–110. doi:10.1038/s42256-020-00298-3

[10] Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=kmn0BhQk7p