We're Still Doing It (All) Wrong
Recommender Systems, Fifteen Years Later

Said, Alan; Pera, Maria Soledad; Ekstrand, Michael D.

Publication date
2025

Document Version
Final published version

Published in
BEYOND Workshop at RecSys 2025

Citation (APA)
Said, A., Pera, M. S., & Ekstrand, M. D. (2025). We're Still Doing It (All) Wrong: Recommender Systems, Fifteen Years Later. In E. Zangerle, A. Said, & C. Bauer (Eds.), *BEYOND Workshop at RecSys 2025* (CEUR Workshop Proceedings; Vol. 4063).

We're Still Doing It (All) Wrong: Recommender Systems, Fifteen Years Later

Alan Said[1], Maria Soledad Pera[2] and Michael D. Ekstrand[3]

[1]University of Gothenburg, Gothenburg, Sweden
[2]TU Delft, Delft, Netherlands
[3]Drexel University, Philadelphia, USA

Abstract

In 2011, Xavier Amatriain sounded the alarm: recommender systems research was "doing it all wrong" [1]. His critique, rooted in statistical misinterpretation and methodological shortcuts, remains as relevant today as it was then. But rather than correcting course, we added new layers of sophistication on top of the same broken foundations. This paper revisits Amatriain's diagnosis and argues that many of the conceptual, epistemological, and infrastructural failures he identified still persist, in more subtle or systemic forms. Drawing on recent work in reproducibility, evaluation methodology, environmental impact, and participatory design, we showcase how the field's accelerating complexity has outpaced its introspection. We highlight ongoing community-led initiatives that attempt to shift the paradigm, including workshops, evaluation frameworks, and calls for value-sensitive and participatory research. At the same time, we contend that meaningful change will require not only new metrics or better tooling, but a fundamental reframing of what recommender systems research is for, who it serves, and how knowledge is produced and validated. Our call is not just for technical reform, but for a recommender systems research agenda grounded in epistemic humility, human impact, and sustainable practice.

Keywords

Reflection, evaluation, call for action

1. Still Doing It Wrong

In 2011, Xavier Amatriain wrote a blog post that many remember but few acted on. Ratings, like Likert scales, are ordinal. However, recommender systems routinely treat them as interval, applying Pearson correlation, computing RMSE, and optimizing linear models based on assumptions users do not make. Around the same time, Konstan and Riedl [2] observed that despite the community's emphasis on prediction accuracy (in light of the Netflix Prize in the mid-00s), small changes in error metrics were not the path to meaningful improvements in the user experience (so foundational to recommender systems). And yet, it cannot be denied that most of the community still today continues to chase minuscule performance gains and treats it as a win. Amatriain called for a "drastic change in the way we approach these issues" [1]. While some of the specific technical details have changed, i.e., moving from predicting ratings to generating top-N lists or considering new scenarios like session-based recommendation, the more fundamental change in assumptions never came. The field has made enormous progress in algorithmic innovation, industrial adoption, and academic visibility. Still, we have to ask: *is this simply more sophisticated machinery built on the same unstable ground?*

We revisit that moment, almost fifteen years later, not out of nostalgia but necessity. Despite powerful tools, reproducibility frameworks, and expansive literature, many of the field's underlying assumptions remain untouched or have been altered in mostly superficial ways. The metrics change, but optimization on historical data still dominates. We argue that recommender systems research continues to prioritize optimization over understanding, evaluation over reflection, and abstraction over accountability. We do not offer new metrics or architectures. Instead, we offer a provocation, a reflection informed by a decade and a half of scholarship, community experience, and frustrated attempts at reform. Our goal is to name

Beyond Algorithms: Reclaiming the Interdisciplinary Roots of Recommender Systems Workshop (BEYOND 2025), September 26th, 2025, co-located with the 19th ACM Recommender Systems Conference, Prague, Czech Republic.
alansaid@acm.org (A. Said); m.s.pera@tudelft.nl (M. S. Pera); mdekstrand@drexel.edu (M. D. Ekstrand)
0000-0002-2929-0529 (A. Said); 0000-0002-2008-9204 (M. S. Pera); 0000-0003-2467-0108 (M. D. Ekstrand)

persistent epistemological failures and to point toward the structural and conceptual shifts needed to address them. If recommendation is to serve people, not just datasets, then RecSys researchers and practitioners driving advances in recommender systems research must ask better questions.

2. Same Fallacies, More Layers

Modern recommenders rarely expose raw ratings. They operate on embeddings, graph convolutions, and attention networks. Yet many still optimize for average error or ranking metrics based on the same flawed assumptions. Behavior is treated as truth, preference as (stable) ground truth, and interaction logs as stable evidence. Recent reproducibility analyses show that even simple algorithms, such as MostPop and ItemKNN, yield inconsistent results across frameworks due to differing defaults, metrics, and evaluation logic [3]. Even on the same framework, poor parameter tuning can lead to incorrect assumptions about which algorithms outperform others in specific contexts [4]. Ranking-based methods such as BPR or LightGCN still assume that a small set of clicks reveals true preference ordering. Many models trained on implicit data are validated with metrics grounded in explicit semantics.

Advancements in the field can sometimes bring new incorrect assumptions. For instance, sequential recommendation [5] is based on the sound premise that sequence and time matter for modeling user preference and generating effective recommendations. Naïve sequential problem formulations like next-click prediction, however, undermine the value of that advance. Training and evaluating a model based on its ability to predict the next item the user clicked inherently ties it to past user trajectories and the internal and external factors that influenced them. This approach prioritizes replicating historical patterns over building models that help users discover items they need *more* efficiently or effectively. This issue is compounded by the long-standing limitation of dataset-driven evaluation: models are typically evaluated on their ability to find the items the user found somehow in the past, not on their ability to *improve* the user experience, e.g., identifying better items [6, 7, 8]. Off-policy evaluation [9, 10] seeks to adjust for the influence of the previous recommender system in product exposure, but cannot correctly evaluate recommendations of relevant items to which the user was never exposed, which is the core issue. Studies like those of Chaney et al. [11] quantify the impact of previous exposure policies on user experience and item exposure distribution, but do not address the fundamental evaluation problem. Targeted data collection from users [7] or experts [12] are promising directions and grapple with the fundamental problem, but are not yet widely used and are likely difficult to scale.

The need for reform is no longer a fringe position. A 2024 Dagstuhl Seminar [13] dedicated to recommender systems evaluation outlined major deficits in theoretical justification, fairness treatment, reproducibility practices, and long-term evaluation. Despite these community diagnoses, standard practice has changed little.

At their best, recommender systems as a field and RecSys as a community represent a broad, interdisciplinary investigation into how to effectively match users with information, entertainment, products, artists, creators, romance partners, etc., that align with their needs and interests and promote mutually-beneficial confluences of informational, artistic, personal, or commercial interests. Early research [14, 15, 16] was holistic, considering data, models, user experience, and user response together in a single work. As the field has matured, research has inevitably and necessarily become more specialized. Maintaining balance and visibility for the range of approaches and perspectives needed for effective recommendation and recommender system evaluation is an ongoing concern and debate for the field, with the relative balance of machine learning and human-oriented research (including human-computer interaction, psychology, marketing, economics, etc.) shifting over time. Based on the balance of topics in published papers and the venues in which different types of work appear [17], as well as reviewer comments on submitted manuscripts, machine learning and algorithmic approaches often dominate, even to the point of reviewers objecting to human-focused or evaluative research because it does not show algorithmic improvements (the authors have multiple examples of such comments).

The result is a field that, despite the ongoing discussions, is often more concerned with recommendation as a machine learning benchmark than a human problem—or at least it appears to be the case given

the vast majority of published works in this area. Delivering meaningful, impactful recommendations that real users find useful in their lives, and assessing those recommendations, is a much bigger and much more difficult problem than improving a standard metric on a standard dataset.

3. The Cult of Evaluation

Most research up to this point has focused on improving the "accuracy" of recommender systems. We believe that this narrow focus has been misguided and even detrimental to the field. The recommendations that are most accurate according to the standard metrics are sometimes not the recommendations that are most useful to users [18]. RMSE has been replaced by nDCG, Precision, and Recall, but the fundamental logic of dataset optimization remains. A systematic review of evaluation-focused recommender system research from 2017 to 2022 found that the vast majority of work uses only a few metrics and relies heavily on offline experiments [19]. Beyond-accuracy metrics such as diversity, novelty, and fairness remain rare [20]. Most papers evaluate on one or two standard datasets, often MovieLens 1M or an Amazon review subset. Within evaluation-focused research, 32% of papers use a single metric; over 70% use three or fewer [19]. Many include RMSE or MAE, despite consensus that they are inadequate for assessing user-centric quality. The result is a precise but shallow evaluation culture, disconnected from actual outcomes.

Evaluation is also not the same thing as knowledge. Reflecting on TREC[1] and its role in information retrieval research, Ian Soboroff tweeted in 2021[2] that 'The datasets were not built to be solved. They were built as tools to understand the problem and the systems we build to "solve" them.'

Recommender systems datasets can play a similar role: they can serve as tools to help us scientists improve our understanding of user behavior, recommendation problems, and the relative strengths and weaknesses of different approaches to recommendation. Such knowledge is more likely to be generalizable into new problem settings and applications than improved metric performance. To generate such knowledge, however, experiments need to be structured to produce knowledge, not just assess performance. A simple example that is thankfully increasing in prevalence is an ablation study, that seeks to understand *which* components of a proposed solution are driving observed metrics. Similarly, disaggregated [22] and distributional [23] evaluations aim to understand how a system's performance may vary for different types of users and items. Detailed understandings of the context of recommender system behavior and performance will also make it easier to determine whether and when proposed new advances can be applied in production. Experimentation for knowledge needs more design like this: metrics, analyses, and runs that look beyond leaderboard benchmark performance to understand *when*, *why*, *how*, and *for who* a new idea is delivering beneficial results.

There are ongoing efforts to improve this landscape. The FEVR framework offers a structured approach to organizing recommender systems evaluation goals, metrics, and methodologies [24]. The CAFE framework [25], focused on conversational recommender (and search) systems, propose evaluation methodologies that consider stakeholder goals, user tasks, user characteristics, diverse assessment criteria, methodology, and quantitative measures. Frameworks such as these aim to make evaluation more transparent, purpose-driven, and context-aware. However, such frameworks remain underutilized, and their influence on standard practice has so far been limited.

4. Reproducibility ≠ Reliability

Improved tooling and reproducibility tracks have increased transparency, but not comparability. A small set of datasets (MovieLens, Amazon) dominates, yet papers rarely disclose preprocessing details, timestamp filtering, or train/test leakage risks [26, 27]. Nearly half of papers use a single dataset; two-thirds rely on datasets used only once [19].

[1]The Text REtrieval Conference (TREC) is a series of workshops that promote research in information retrieval, offering test collections, standardized evaluation methods, and a platform for result comparison [21].

[2]https://web.archive.org/web/20210815134713/https://twitter.com/ian_soboroff/status/1426901262369439751

Offline evaluation continues to serve as the default setting. Even in dedicated evaluation studies, temporal dynamics are often ignored, and random splits are still widely employed [28]. Empirical evidence shows that evaluation frameworks such as Elliot, RecBole, and Cornac can produce significantly different outcomes for the same algorithm, even when configurations are aligned [29]. This inconsistency is not limited to complex or novel models. Performance variance has been observed even for simple baselines such as ItemKNN or MostPop.

Many experimental defaults, including relevance thresholding, sampling strategies, cutoff choices, and metric implementations, differ across frameworks and publications. These differences are rarely disclosed or justified. For example, the choice of random seed can meaningfully alter results, yet most studies report only a single run without intervals or variance estimates. In some cases, the variation caused by these defaults exceeds the claimed improvements over established baselines.

In this environment, reproducibility often amounts to rerunning fragile configurations rather than building confidence in robust insights. As long as evaluation is treated as procedural rather than interpretive, the field risks reinforcing artifacts of implementation instead of uncovering meaningful patterns. Reliability requires not only access to code and data, but also critical reflection on design decisions and their consequences.

5. New Sins for a New Era

As a community, we have not only failed to correct past mistakes, but we have invented new ones:

- **Environmental neglect**: Recommenders increasingly rely on resource-intensive architectures. Few papers disclose compute costs or carbon impact. Training one deep recommender system model can consume orders of magnitude more energy than traditional approaches [30, 31, 32].
- **Unchecked need**: A significant chunk of recommender system research has embraced Large Language Models (LLMs) as a panacea, making them core to the process. This has been done without regard to the extent to which LLMs can memorize common datasets; an effect that is unsurprisingly tied to improved performance [33]. Further, in some cases, these models perform on par with less resource-intensive alternatives in standard offline evaluation environments [33, 30]. Perhaps it is better to verify that such models are necessary and will deliver benefits commensurate with their costs before deploying them.
- **Ethical fragility**: Algorithmic fairness and user autonomy are now part of the discourse, but rarely built into model or system design or evaluation. Metrics are post hoc [34].
- **User disempowerment**: Transparency, control, and interaction remain secondary concerns. Recommenders "push" rather than "negotiate" [35]. Recommendation lacks reciprocity with the system designers [36] or, in generative recommendation, with the artists, authors, and other creators of recommended work [37]. Belkin and Robertson [38] sounded the alarm about push-oriented and sender-focused information science in 1976. Their warnings are even more apropos today as generative modalities converge with disinformation and influence campaigns within the recommender-mediated platforms that shape much of how people understand the world.

6. What Would Doing It Right Look Like?

Doing it right does not mean optimizing harder. It means asking better questions and reflecting on the answers. There is no need for a single metric or framework; reducing improvements to standardized procedures is more likely to hinder the necessary diversity of research methods and lines of inquiry needed to keep recommender systems on a course of human welfare. However, recommender systems research needs: (i) Diverse datasets, evaluated across varied contexts; (ii) Human-aligned evaluation: not only offline precision, but meaningful outcomes; (iii) Transparent reporting: preprocessing, hyperparameters, compute, and code; (iv) Sustainable practices: energy as a first-class consideration; (v) Epistemic humility: acknowledging noise, preference volatility, and modeling limits; and (vi) Normative

and human grounding: explicit articulation of the individual or human social goals of the system or evaluation, and connection of the technical work to those goals.

Recent work has emphasized the societal obligations of recommender system researchers. In the context of RecSys for Social Good, it has been argued that recommender systems research must be responsive not only to users' preferences, but also to broader social outcomes such as justice, inclusion, well-being, and sustainability development goals [39, 40, 41, 42]. Doing it right, then, requires not just methodological reform, but a reframing of what we consider valuable progress in the field.

Community-led initiatives have emerged to create space for alternative voices and perspectives. The AltRecSys workshop (RecSys'24) explicitly foregrounded speculative, critical, and unconventional ideas in recommendation [43]. Rather than showcasing polished technical contributions, the workshop invited participants to question foundational assumptions, share negative results, and propose revisiting what counts as valuable research. This venue demonstrated that the community has become increasingly aware that many of the field's core challenges are epistemological and institutional as much as they are technical. The NORMalize [44] and RecSoGood [45] workshops (at RecSys) encouraged the community to rethink evaluation practices, highlighting the importance of considering norms and values driving recommenders, along with their potential to guide users towards more sustainable choices and behaviors, supporting broader environmental and social goals. At the risk of tooting our own horn, ongoing work on the LensKit Codex (https://codex.lenskit.org) aims to improve consistency and sustainability of baseline comparisons by providing standardized evaluation results and hyperparameter tunings for common models on a range of public datasets, allowing researchers to compare against well-tuned baseline models without incurring the environmental cost (and risk of error) of re-tuning the baselines themselves, and to invite community collaboration on tuning methods and standardized results.

Recent work has proposed reimagining recommender system development as a participatory process, where users, providers, and other stakeholders act not only as research subjects, but as co-designers and co-evaluators of the systems that shape their digital experiences [46]. This approach challenges the prevailing notion that designers alone should define the goals and metrics. Instead, it advocates for a redistribution of design authority, grounded in democratic values and the lived experiences of those affected. Doing it right, in this view, means sharing power, not just optimizing performance. Burke and Sylvester [37] argue for recommender systems research to make a *relational* turn (or re-turn, as key early research had a decidedly relational bent [16]). This sentiment is also prevalent in recent works focused on multistakeholder-focused evaluation [47], arguing to move beyond accuracy measures that simply capture the overall utility of a single stakeholder, to consider the complex, domain-dependent, and multifaceted task of probing the impact of recommender systems on all of their stakeholders.

Holistic, human-centered recommender systems research is not *easy*. It often takes more time to plan and execute, and may require resources beyond those needed by algorithmic experiments on published data. Several projects are working to make grounded, in-vivo evaluation more accessible, such as Informfully [48] and POPROX[3] [49, co-developed by the last author], as well as the earlier CLEF NewsReel challenge organized with Plista [50]. Some RecSys Challenges, like the one in 2024 [51], focused not only on accuracy but also on beyond-accuracy metrics to account for the normative complexities inherent in news recommendations; the 2021 one [52] aimed for multi-goal optimization, prompting participants to predict user engagement probabilities for tweets while ensuring recommendations were both accurate and fair. These platforms and challenges provide infrastructure that can support more holistic research, as well as working examples for the development of further resources. However, the community has yet to consistently adopt them, let alone sustain their use in the long term.

Echoing Olteanu et al. [53], rigorous research on recommender systems demands going beyond technical and statistical notions of rigor to include normative, epistemic, and other forms of rigor. Encouraging the community towards a well-rounded evaluation framework, one that considers multiple angles to probe a recommender and convey if it is 'good' (in its architecture, performance, but also for whom, from which perspective, and in which context), is already supported in the literature [e.g., 54, 55, 56, 57]. We "just" need to embrace it and take it from theory to practice, and make closing those

[3] https://poprox.ai

Table 1
Shifting recommender systems practice

Dimension	Status Quo	Better Alternative
Assumptions about data	Preferences are stable; ordinal = interval	Preferences are noisy, contextual
Evaluation focus	RMSE, nDCG, Precision	Human goals, mixed methods
System goal	Maximize clicks or accuracy	Support reflection, fairness, trust
Model validation	One-off benchmarks	Transparent, replicable, contextual
User role	Target of optimization	Stakeholder, co-designer

gaps a priority for the field and community. Table 1 summarizes several persistent structural defaults in recommender systems research and highlights what a more human-centered, reflective, and accountable paradigm might look like. These contrasts are not exhaustive, but capture the epistemic shift we argue is necessary: from narrow optimization toward broader alignment with user and societal values.

7. Conclusion: Enough Already

Reflecting in 2004 about evaluation of collaborative filtering recommender systems, Herlocker et al. [58] wrote that "Effective and meaningful evaluation of recommender systems is challenging. To date, there has been no published attempt to synthesize what is known [...], nor to systematically understand the implications [...] for different tasks and different contexts." Today, there are more surveys and frameworks, but a thorough and systematic assessment for collaborative filtering or the many model families extant in the field does not seem much closer now than it did 20 years ago.

Amatriain's critique remains true. Recommender systems research is still doing it wrong. It is doing so at scale, with better tools, more complex models, and increased compute. This is not a problem of ignorance. The challenges are well-documented: from the ordinal nature of ratings or the inadequacy of top-k metrics and implicit feedback as proxies for user experience and ground truth, to the costs of large models. Researchers, developers, and reviewers are aware of these issues. Yet the publication and reward structures in the field continue to favor narrow performance gains and apparent (often mathematical) sophistication over conceptual clarity, ecological awareness, or social relevance [59].

It is time to stop pretending that stronger baselines, larger datasets, new benchmarks, or even more rigorous experimental protocols will fix foundational misalignments. Recommender systems research must shift from engineering models to understanding influence and impact. Fixing this requires recommender systems research to adopt an entirely different posture, one of epistemic humility. We must acknowledge the limitations of our methods and data, and treat evaluation not as a procedural step, but as a site of inquiry. We must ask what kinds of knowledge recommender systems research should produce, and for whom. The point is not to stop working on models and algorithms, but to do better and more useful research on models and algorithms: to contextualize and evaluate them in terms of how well they — and the research and engineering methods behind them — actually drive meaningful improvements in the human experience of recommender systems. There are many gaps between current practice, capabilities, and knowledge and the kinds of human impact the field looks to have, providing important problems and settings for future research.

Recommender systems are not just algorithms; they are sociotechnical interventions [60, 55, 61, 62]. They shape what people see, believe, and desire. If we do not take this responsibility seriously, we will continue to develop systems that are efficient and elegant, but ultimately misaligned with the people they are meant to serve.

Let us stop solving the wrong problem better. Let us define a better problem.

Declaration on Generative AI

During the preparation of this work, the author(s) used X-GPT-4 and Grammarly for Grammar and spelling check and rephrasing. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] X. Amatriain, Recommender Systems: We're doing it (all) wrong, AI, software, tech, and people. Not in that order. By X (2011).

[2] J. A. Konstan, J. Riedl, Recommender systems: from algorithms to user experience, User modeling and user-adapted interaction 22 (2012) 101–123.

[3] A. Akhadam, O. Kbibchi, L. Mekouar, Y. Iraqi, A comparative evaluation of recommender systems tools, IEEE Access (2025).

[4] F. Shehzad, D. Jannach, Everyone'sa winner! on hyperparameter tuning of recommendation models, in: Proceedings of the 17th ACM Conference on Recommender Systems, 2023, pp. 652–657.

[5] M. Quadrana, P. Cremonesi, D. Jannach, Sequence-Aware Recommender Systems, ACM Comput. Surv. 51 (2018) 1–36. doi:10.1145/3190616.

[6] M. D. Ekstrand, V. Mahant, Sturgeon and the Cool Kids: Problems with Top-N Recommender Evaluation, in: Proceedings of the 30th Florida Artificial Intelligence Research Society Conference, FLAIRS 30, AAAI Press, 2017. URL: https://aaai.org/papers/639-flairs-2017-15534/.

[7] M. D. Smucker, H. Chamani, Extending MovieLens-32M to Provide New Evaluation Objectives, 2025. doi:10.48550/arXiv.2504.01863. arXiv:2504.01863.

[8] R. Cañamares, P. Castells, Should I Follow the Crowd?: A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18, ACM, 2018, pp. 415–424. doi:10.1145/3209978.3210014.

[9] O. Jeunen, Revisiting offline evaluation for implicit-feedback recommender systems, in: Proceedings of the 13th ACM Conference on Recommender Systems, RecSys '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 596–600. doi:10.1145/3298689.3347069.

[10] Y. Saito, T. Udagawa, H. Kiyohara, K. Mogi, Y. Narita, K. Tateno, Evaluating the Robustness of Off-Policy Evaluation, in: Proceedings of the 15th ACM Conference on Recommender Systems, RecSys '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 114–123. doi:10.1145/3460231.3474245.

[11] A. J. B. Chaney, B. M. Stewart, B. E. Engelhardt, How algorithmic confounding in recommendation systems increases homogeneity and decreases utility, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18, ACM, 2018, pp. 224–232. doi:10.1145/3240323.3240370.

[12] P. Kouki, I. Fountalis, N. Vasiloglou, X. Cui, E. Liberty, K. Al Jadda, From the Lab to Production: A Case Study of Session-Based Recommendations in the Home-Improvement Domain, in: Proceedings of the Fourteenth ACM Conference on Recommender Systems, RecSys '20, ACM, 2020, pp. 140–149. doi:10.1145/3383313.3412235.

[13] C. Bauer, A. Said, E. Zangerle, Evaluation perspectives of recommender systems: Driving research and education (dagstuhl seminar 24211), Dagstuhl Reports 14 (2024) 58–172. doi:10.4230/DagRep.14.5.58.

[14] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, GroupLens: An Open Architecture for Collaborative Filtering of Netnews, in: CSCW '94, ACM, New York, NY, USA, 1994, pp. 175–186. doi:10.1145/192844.192905.

[15] D. Billsus, M. J. Pazzani, A hybrid user model for news story classification, in: Proceedings of the Seventh International Conference on User Modeling, volume 407 of *CISM International Centre for Mechanical Sciences*, Springer, 1999, pp. 99–108. doi:10.1007/978-3-7091-2490-1_10.

[16] W. Hill, L. Stead, M. Rosenstein, G. Furnas, Recommending and evaluating choices in a virtual community of use, in: CHI '95: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 1995, pp. 194–201. doi:10.1145/223904.223929.

[17] B. Smyth, People who liked this also liked... a publication analysis of three decades of recommender systems research, ACM Transactions on Recommender Systems (2025).

[18] S. M. McNee, J. Riedl, J. A. Konstan, Being accurate is not enough: how accuracy metrics have hurt recommender systems, in: CHI'06 extended abstracts on Human factors in computing systems,

2006, pp. 1097–1101.
[19] C. Bauer, E. Zangerle, A. Said, Exploring the Landscape of Recommender Systems Evaluation: Practices and Perspectives, ACM Trans. Recomm. Syst. 2 (2024) 11:1–11:31. doi:10.1145/3629170.
[20] M. Kaminskas, D. Bridge, Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems, ACM Transactions on Interactive Intelligent Systems (TiiS) 7 (2016) 1–42.
[21] E. Voorhees, D. Harman, Overview of the sixth text retrieval conference (TREC-6), Nist Special Publication Sp (1998) 1–24.
[22] S. Barocas, A. Guo, E. Kamar, J. Krones, M. R. Morris, J. W. Vaughan, W. D. Wadsworth, H. Wallach, Designing disaggregated evaluations of AI systems: Choices, considerations, and tradeoffs, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, 2021, pp. 368–378. doi:10.1145/3461702.3462610.
[23] M. D. Ekstrand, B. Carterette, F. Diaz, Distributionally-informed recommender system evaluation, ACM Transactions on Recommender Systems 2 (2024) 6:1–27. doi:10.1145/3613455.
[24] E. Zangerle, C. Bauer, Evaluating Recommender Systems: Survey and Framework, ACM Comput. Surv. 55 (2022) 170:1–170:38. doi:10.1145/3556536.
[25] C. Bauer, L. Chen, N. Ferro, N. Fuhr, Conversational agents: A framework for evaluation (cafe)(dagstuhl perspectives workshop 24352), Dagstuhl Reports 14 (2025) 53–58.
[26] Z. Sun, D. Yu, H. Fang, J. Yang, X. Qu, J. Zhang, C. Geng, Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison, in: RecSys '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 23–32. doi:10.1145/3383313.3412489.
[27] B. Hidasi, Á. T. Czapp, Widespread Flaws in Offline Evaluation of Recommender Systems, in: Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 848–855. doi:10.1145/3604915.3608839.
[28] Z. Meng, R. McCreadie, C. Macdonald, I. Ounis, Exploring Data Splitting Strategies for the Evaluation of Recommendation Models, in: Fourteenth ACM Conference on Recommender Systems, Association for Computing Machinery, New York, NY, USA, 2020, pp. 681–686. doi:10.1145/3383313.3418479.
[29] M. Schmidt, J. Nitschke, T. Prinz, Evaluating the performance-deviation of itemKNN in RecBole and LensKit, 2024. doi:10.48550/arXiv.2407.13531. arXiv:2407.13531.
[30] T. Vente, L. Wegmeth, A. Said, J. Beel, From Clicks to Carbon: The Environmental Toll of Recommender Systems, in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 580–590. doi:10.1145/3640457.3688074.
[31] G. Spillo, A. De Filippo, C. Musto, M. Milano, G. Semeraro, Comparing data reduction strategies for energy-efficient green recommender systems, Journal of Intelligent Information Systems (2025). doi:10.1007/s10844-025-00965-1.
[32] G. Spillo, A. G. Valerio, F. Franchini, A. De Filippo, C. Musto, M. Milano, G. Semeraro, RecSys CarbonAtor: Predicting Carbon Footprint of Recommendation System Models, in: L. Boratto, A. De Filippo, E. Lex, F. Ricci (Eds.), Recommender Systems for Sustainability and Social Good, Springer Nature Switzerland, Cham, 2025, pp. 98–110. doi:10.1007/978-3-031-87654-7_10.
[33] D. Di Palma, F. A. Merra, M. Sfilio, V. W. Anelli, F. Narducci, T. Di Noia, Do llms memorize recommendation datasets? a preliminary study on movielens-1m, in: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2025, pp. 2582–2586.
[34] S. Milano, M. Taddeo, L. Floridi, Recommender systems and their ethical challenges, AI & SOCIETY 35 (2020) 957–967. doi:10.1007/s00146-020-00950-y.
[35] S. Wang, X. Zhang, Y. Wang, F. Ricci, Trustworthy Recommender Systems, ACM Trans. Intell. Syst. Technol. 15 (2024) 84:1–84:20. doi:10.1145/3627826.
[36] M. D. Ekstrand, M. C. Willemsen, Behaviorism is not enough: Better recommendations through listening to users, in: Proceedings of the 10th ACM Conference on Recommender Systems, 2016,

pp. 221–224. doi:10.1145/2959100.2959179.
[37] R. Burke, M. Sylvester, Post-userist recommender systems: A manifesto, 2024. doi:10.48550/arXiv.2410.11870. arXiv:2410.11870.
[38] N. J. Belkin, S. E. Robertson, Some ethical and political implications of theoretical research in information science, in: Proceedings of the ASIS Annual Meeting, 1976. URL: https://www.researchgate.net/publication/255563562.
[39] A. Said, Recommender Systems for Social Good:The Role of Accountability and Sustainability, in: Workshop on Recommender Systems for Social Good, volume Proceeding of the 2024 workshop on Recommender Systems for Social Good of *RecSoGood*, Springer, 2024, pp. 1–4.
[40] D. Jannach, A. Said, M. Tkalcic, M. Zanker, Recommender Systems for Good (RS4Good): Survey of Use Cases and a Call to Action for Research that Matters, ACM Trans. Recomm. Syst. (2025). doi:10.1145/3746648.
[41] A. Felfernig, M. Wundara, T. N. T. Tran, S. Polat-Erdeniz, S. Lubos, M. El Mansi, D. Garber, V.-M. Le, Recommender systems for sustainability: overview and research issues, Frontiers in big Data 6 (2023) 1284511.
[42] G. K. Patro, A. Chakraborty, A. Banerjee, N. Ganguly, Towards safety and sustainability: designing local recommendations for post-pandemic world, in: Proceedings of the 14th ACM Conference on Recommender Systems, 2020, pp. 358–367.
[43] M. Ekstrand, M. S. Pera, A. Said, AltRecSys: A Workshop on Alternative, Unexpected, and Critical Ideas in Recommendation, in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 1216–1218. doi:10.1145/3640457.3687104.
[44] S. Vrijenhoek, L. Michiels, J. Kruse, A. Starke, N. Tintarev, J. Viader Guerrero, Normalize: The first workshop on normative design and evaluation of recommender systems, in: Proceedings of the 17th ACM Conference on Recommender Systems, 2023, pp. 1252–1254.
[45] L. Boratto, A. De Filippo, E. Lex, F. Ricci, First international workshop on recommender systems for sustainability and social good (recsogood 2024), in: Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 1239–1241.
[46] M. D. Ekstrand, A. Razi, A. Sarcevic, M. S. Pera, R. Burke, K. L. Wright, Recommending with, not for: Co-designing recommender systems for social good, ACM Transactions on Recommender Systems Just Accepted (2025). doi:10.1145/3759261.
[47] R. Burke, G. Adomavicius, T. Bogers, T. Di Noia, D. Kowald, J. Neidhardt, Özlem Özgöbek, M. S. Pera, N. Tintarev, J. Ziegler, De-centering the (traditional) user: Multistakeholder evaluation of recommender systems, International Journal of Human-Computer Studies 203 (2025) 103560. URL: https://www.sciencedirect.com/science/article/pii/S107158192500117X. doi:https://doi.org/10.1016/j.ijhcs.2025.103560.
[48] L. Heitz, J. A. Croci, M. Sachdeva, A. Bernstein, Informfully - Research Platform for Reproducible User Studies, in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 660–669. doi:10.1145/3640457.3688066.
[49] R. Burke, M. Ekstrand, Conducting Recommender Systems User Studies Using POPROX, in: Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct '25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 1–2. doi:10.1145/3708319.3727558.
[50] T. Brodt, F. Hopfgartner, Shedding light on a living lab: The CLEF NEWSREEL open recommendation platform, in: Proceedings of the 5th Information Interaction in Context Symposium, ACM, Regensburg Germany, 2014, pp. 223–226. doi:10.1145/2637002.2637028.
[51] RecSys Challenge 2024, https://www.recsyschallenge.com/2024/, 2024. [Accessed 02-09-2025].
[52] RecSys Challenge 2021, https://www.recsyschallenge.com/2021/, 2021. [Accessed 02-09-2025].
[53] A. Olteanu, S. L. Blodgett, A. Balayn, A. Wang, F. Diaz, F. d. P. Calmon, M. Mitchell, M. Ekstrand, R. Binns, S. Barocas, Rigor in AI: Doing Rigorous AI Work Requires a Broader, Responsible AI-Informed Conception of Rigor, 2025. doi:10.48550/arXiv.2506.14652. arXiv:2506.14652.

[54] C. Bauer, C. Bagchi, O. A. Hundogan, K. van Es, Where are the values? a systematic literature review on news recommender systems, ACM Transactions on Recommender Systems 2 (2024) 1–40.
[55] J. Stray, A. Halevy, P. Assar, D. Hadfield-Menell, C. Boutilier, A. Ashar, C. Bakalar, L. Beattie, M. Ekstrand, C. Leibowicz, et al., Building human values into recommender systems: An interdisciplinary synthesis, ACM Transactions on Recommender Systems 2 (2024) 1–57.
[56] T. V. Rampisela, T. Ruotsalo, M. Maistro, C. Lioma, Can we trust recommender system fairness evaluation? the role of fairness and relevance, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 271–281.
[57] R. Ungruh, M. Al Nahadi, M. S. Pera, Mirror, mirror: Exploring stereotype presence among top-n recommendations that may reach children, ACM Transactions on Recommender Systems (2025).
[58] J. L. Herlocker, J. A. Konstan, L. G. Terveen, J. T. Riedl, Evaluating collaborative filtering recommender systems, ACM Transactions on Information Systems (TOIS) 22 (2004) 5–53.
[59] Z. C. Lipton, J. Steinhardt, Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research., Queue 17 (2019) Pages 80:45–Pages 80:77. doi:10.1145/3317287.3328534.
[60] N. Seaver, Algorithms as culture: Some tactics for the ethnography of algorithmic systems, Big Data & Society 4 (2017) 2053951717738104. doi:10.1177/2053951717738104.
[61] E. Lucherini, M. Sun, A. Winecoff, A. Narayanan, T-RECS: A simulation tool to study the societal impact of recommender systems, 2021. URL: http://arxiv.org/abs/2107.08959.
[62] B. Mitra, Search and Society: Reimagining Information Access for Radical Futures, 2024. doi:10.48550/arXiv.2403.17901. arXiv:2403.17901.