

# Normative Alignment of Recommender Systems via Internal Label Shift

Johannes Kruse  
JP/Politikens Media Group  
Copenhagen, Denmark  
Applied Mathematics and Computer  
Science  
Technical University of Denmark  
Kongens Lyngby, Denmark  
Johannes.Kruse@jppol.dk

Kasper Lindskow  
JP/Politikens Media Group  
Copenhagen, Denmark  
Copenhagen Business School  
Frederiksberg, Denmark  
kasper.lindskow@jppol.dk

Michael Riis Andersen  
Applied Mathematics and Computer  
Science  
Technical University of Denmark  
Kongens Lyngby, Denmark  
miri@dtu.dk

Ryotaro Shimizu  
Computer Science  
University of California San Diego  
La Jolla, California, USA  
ZOZO Research  
Tokyo, Japan  
r2shimizu@ucsd.edu

Julian McAuley  
Computer Science  
University of California San Diego  
La Jolla, California, USA  
jmcauley@ucsd.edu

Pierre-Alexandre Mattei  
Inria, Université Côte d'Azur  
Nice, France  
pierre-alexandre.mattei@inria.fr

Jes Frellsen  
Applied Mathematics and Computer  
Science  
Technical University of Denmark  
Kongens Lyngby, Denmark  
Pioneer Centre for Artificial  
Intelligence  
Copenhagen, Denmark  
jefr@dtu.dk

## Abstract

Recommender systems optimized solely for user engagement often fail to meet broader normative objectives such as fairness, diversity, or editorial values. We introduce NAILS (Normative Alignment of recommender systems via Internal Label Shift), a simple and scalable method for aligning recommendation outputs with target distributions over item-level attributes (e.g., categories). NAILS modifies the user-conditional item distribution to induce a specified marginal distribution over attributes, leveraging existing user-item preferences without retraining the model. To achieve this, we recast the problem as a form of label shift applied internally within a hierarchical classification framework. Adopting a stakeholder-centric perspective, NAILS enables alignment with global normative goals. Empirically, we show that NAILS consistently improves attribute-level alignment with minimal impact on user engagement,

providing a practical mechanism for value-driven recommendation. Our code is available at <https://github.com/johanneskruse/nails>.

## CCS Concepts

• **Information systems** → **Recommender systems**.

## Keywords

Recommender Systems; Aligned Recommendation; Normative Design; Relevance Prioritized Reranking

## ACM Reference Format:

Johannes Kruse, Kasper Lindskow, Michael Riis Andersen, Ryotaro Shimizu, Julian McAuley, Pierre-Alexandre Mattei, and Jes Frellsen. 2025. Normative Alignment of Recommender Systems via Internal Label Shift. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25)*, September 22–26, 2025, Prague, Czech Republic. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3705328.3759309>

## 1 Introduction

Recommender systems play a central role in content delivery across a wide range of domains, including news, entertainment, and e-commerce [4, 9, 14, 24, 25]. While these systems are typically optimized for user engagement, such optimization may come at the expense of alignment with broader organizational or societal

---

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

RecSys '25, Prague, Czech Republic

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1364-4/25/09

<https://doi.org/10.1145/3705328.3759309>

goals [7, 31]. In domains like news, for example, purely behavior-driven personalization can reinforce filter bubbles, reduce content diversity, and underrepresent important but less popular content [20, 30]. This growing concern has motivated research into the normative design of recommender systems—designing recommender systems that are not only effective but also align with normative objectives, such as fairness, diversity, or editorial values [6, 11, 16, 18, 28, 29].

In this paper, we propose Normative Alignment of recommender systems via Internal Label Shift (NAILS), a lightweight framework that enables alignment of recommendations toward normative objectives. While much of the research on alignment of recommendation has adopted a user-centric perspective [1, 10, 27], i.e., focusing on calibrating recommendations based on individual user interactions, our work takes a stakeholder-centric view, empowering platforms, such as mission-driven news publishers, to align recommendations with global normative objectives while preserving engagement and personalization. Our contributions in this paper are:

- We introduce NAILS, a simple and efficient method for aligning recommendations with global normative objectives across users.
- We evaluate NAILS on the EB-NeRD dataset [15] and demonstrate that it effectively aligns recommendation outputs with diverse normative target distributions.
- We motivate future directions for normative alignment, including live deployment, long-term effects, and editorial and policy-driven recommendation systems.

## 2 Related Work

Our method is closely related to calibration in recommender systems, a concept introduced by Steck [27], who argued that recommendation lists should reflect the distribution of content categories in a user’s historical preferences. For example, if a user has consumed 80% romance and 20% action content, a calibrated system should recommend items in similar proportions. To achieve this, Steck [27] proposed CaliRec, a greedy post-processing algorithm that adjusts recommendation lists to prevent users’ minority interests from being overwhelmed by dominant ones—a common side effect of optimizing purely for ranking accuracy.

Subsequent work has further advanced this idea. Seymen et al. [23] proposed a non-greedy approach by formulating calibration as a constrained optimization problem solved via mixed integer programming. Chen et al. [3] introduced an end-to-end framework that decouples accuracy and calibration into separate encoders and modifies output distributions to improve diversity and balance. Abdollahpouri et al. [1] formulated the task as a minimum cost flow problem, yielding an efficient and exact solution for calibrated ranking. Jeon et al. [10] proposed LeapRec, a two-phase method combining calibration-disentangled learning during training with a relevance-prioritized re-ranking step for sequential recommendations. These works frame calibration from a user-centric perspective, aligning recommendations with individual users’ historical preferences.

In contrast, we adopt a stakeholder-centric perspective, focusing on alignment with broader normative objectives across the user

population. While this perspective has been overlooked, a few recent works have explored similar ideas. For example, Wang et al. [32] propose Personalized Calibration Targets, aiming to balance user-level interest alignment with a system-level target exposure distribution. Similarly, Zhao et al. [36] introduce a target customer re-ranking algorithm to adjust the population distribution and composition in the top- $K$  target customers of an item, while preserving recommendation quality. Our approach contributes to this emerging line of work by offering a novel formulation of stakeholder-level calibration as a label shift problem—a perspective that enables a scalable, principled, and model-agnostic mechanism for aligning recommendation distributions with external normative targets.

## 3 Normative Alignment via Internal Label Shift

Let  $p_\theta(i | u)$  denote the probability that user  $u \in \mathcal{U}$  interacts with item  $i \in \mathcal{I}$  in a recommender system parameterized by  $\theta$ . Here,  $\mathcal{I}$  denotes the set of candidate items to be ranked, and  $\mathcal{U}$  denotes the set of users or their representations, which may include, for example, contextual information and session-specific features. The parameters  $\theta$  are typically estimated from historical user–item interaction data. Additionally, consider a probabilistic mapping from items  $i \in \mathcal{I}$  to a set of attributes  $\mathcal{C}$ , represented by  $p(c | i)$ . We assume the joint model

$$p(c, i, u) = p(c | i)p_\theta(i | u)p(u), \quad (1)$$

where  $p(u)$  is the marginal distribution of user representations, which could be estimated, e.g., from the empirical distribution of users in our dataset. An important point is that we never actually have to compute it in practice, as we shall see in Equation (5). The attribute  $c \in \mathcal{C}$  could denote, for example, the category of an item. If each item belongs to exactly one category, the mapping becomes deterministic. In this case, if each category  $c$  is represented by the set of items belonging to it, then  $\mathcal{C}$  forms a partition of  $\mathcal{I}$ , and the mapping reduces to the indicator function  $p(c | i) = \mathbf{1}_c(i)$ . We aim to solve a normative alignment problem by replacing the marginal attribute distribution  $p(c) = \sum_{i \in \mathcal{I}, u \in \mathcal{U}} p(c, i, u)$  with an alternative target distribution  $\tilde{p}(c)$ . Here,  $p(c)$  represents the frequency of user interactions with attribute  $c$  (e.g., categories), and can be estimated from data or computed via marginalization of the joint model  $p(c, i, u)$ . For example, in news recommendations, an editor might wish to enforce a desired distribution of news categories that better reflects the newspaper’s editorial values. This problem can be naturally framed as one of adjusting class proportions of the attribute, for which the label shift literature offers principled correction methods [5, 22]. However, our model, c.f. Equation (1), more closely resembles hierarchical classification [13, 26], where labels are structured in an attribute–item hierarchy. Standard label shift methods assume a non-hierarchical label space and apply the shift at the leaf level—corresponding to individual items—whereas we seek to impose the shift at an internal level, corresponding to attributes.

To solve the normative alignment problem, we consider a new model explicitly constructed to have a normative marginal distribution  $\tilde{p}(c)$ , defined as

$$\tilde{p}(c, i, u) = \tilde{p}(i, u | c)\tilde{p}(c). \quad (2)$$

Following Saerens et al. [22], we assume that the conditional distribution of users and items given the attribute remains unchanged, i.e.,  $\forall c \in \mathcal{C} : p(i, u | c) = \tilde{p}(i, u | c)$ . Under this assumption, the normatively aligned model can be expressed as

$$\tilde{p}(c, i, u) = p(i, u | c)\tilde{p}(c) = \frac{p(c | i)p_{\theta}(i | u)p(u)}{p(c)}\tilde{p}(c). \quad (3)$$

The resulting probability that a user interacts with an item under this normative aligned model is then given by

$$\tilde{p}(i | u) = \frac{\sum_{c \in \mathcal{C}} \tilde{p}(c, i, u)}{\tilde{p}(u)} = \frac{p(u)}{\tilde{p}(u)} \left( \sum_{c \in \mathcal{C}} \frac{\tilde{p}(c)}{p(c)} p(c | i) \right) p_{\theta}(i | u), \quad (4)$$

where the user-dependent normalization constant can be computed as

$$\frac{\tilde{p}(u)}{p(u)} = \sum_{i \in \mathcal{I}} \left( \sum_{c \in \mathcal{C}} \frac{\tilde{p}(c)}{p(c)} p(c | i) \right) p_{\theta}(i | u). \quad (5)$$

This correction resembles the ratio proposed by Saerens et al. [22], with the additional summation over categories. In practice, it can be implemented by adding the weight  $\log \sum_{c \in \mathcal{C}} \frac{\tilde{p}(c)}{p(c)} p(c | i)$  to the log probabilities of the model, and passing the result through a softmax layer. This weight can be viewed as a reweighting term that adjusts the recommendations to align with the normative marginal distribution  $\tilde{p}(c)$ .

Label shift in our setting involves modifying the attribute distribution  $p(c)$ , which theoretically implies a corresponding shift in the marginal user distribution from  $p(u)$  to  $\tilde{p}(u)$ . However, in practice, we apply the correction while keeping  $p(u)$  fixed. As a result, the corrected model does not strictly match the target marginal  $\tilde{p}(c)$ , but instead offers an approximate solution that preserves the intended directional shift. Despite this mismatch, we find the method to be effective empirically.

To control the influence of normative alignment, we introduce a tunable hyperparameter  $\lambda \in [0, 1]$  that mixes between the original and normative aligned recommenders, i.e.,

$$\tilde{p}_{\lambda}(i | u) := \lambda \tilde{p}(i | u) + (1 - \lambda) p_{\theta}(i | u). \quad (6)$$

## 4 Experimental Setup

We evaluate NAILS in the context of news recommendations. In this setup, the items are news articles, and the attribute of interest is the category of each article. We assume each article belongs to exactly one category; therefore,  $p(c | i)$  is an indicator function.

*Dataset.* We use the Ekstra Bladet News Recommendation Dataset (EB-NeRD), a large-scale benchmark dataset for news recommendation [15]. It contains over 37 million impression logs collected from more than 1 million unique users interacting with over 125,000 distinct news articles. The articles are labeled with one of eight editorial categories: *entertainment* (23%), *news* (22%), *crime* (18%), *sports* (15%), *miscellaneous* (9%), *lifestyle* (7%), *erotic* (3%), and *opinion* (3%). All results are reported on the hidden test set. We use the 13,336,710 impression logs to evaluate ranking performance, and 200,000 beyond-accuracy samples to assess distributional calibration. We remove the auto category (auto-generated articles) from the candidate list (mainly found in the beyond-accuracy samples), resulting in 155 candidate articles per impression.

*Baselines.* We use NRMS [34] as our base recommendation model  $p_{\theta}(i | u)$ . It is a widely adopted neural news recommender based on multi-head self-attention, commonly used in recent news recommendation research [8, 16, 17, 21, 35].

To establish a baseline, we include CaliRec [27], a post-hoc reranking method that greedily constructs a top- $K$  list by balancing personalization with alignment to a target distribution over item attributes (e.g., categories). Similar to NAILS, as shown in Equation (6), CaliRec includes a tunable hyperparameter  $\lambda \in [0, 1]$  that interpolates between the original recommender and the normatively aligned objective. We follow the original implementation and apply additive smoothing to prevent zero-probability issues.

*Hyperparameter Tuning.* We use Optuna [2] with the Tree-Structured Parzen Estimator [33] to tune the NRMS model for AUC performance on the EB-NeRD small validation set. The search space is explored over 25 trials. Each model is trained on the training portion (232,887 samples), with the final 24 hours as hold-out set for early stopping. The validation set contains 244,647 samples.

Following Wu et al. [34], we fix the negative sampling ratio to  $K = 4$  and use a batch size of 32. News encoders are initialized with the dataset’s open-sourced Word2Vec embeddings [19] which outperformed randomly initialized embeddings in preliminary experiments. The final model uses the best hyperparameters found during tuning: a learning rate of  $10^{-4}$  with the Adam optimizer [12],  $\ell^2$ -regularization ( $\lambda_{\ell^2} = 10^{-4}$ ), 20% dropout, attention query dimensionality of 100, and 24 attention heads each producing a 24-dimensional output.

After tuning, we merge the training and validation sets and reserve the final day of the validation split for early stopping. All experiments are conducted on Amazon EC2 instances using g5.xlarge machines with NVIDIA A10 GPUs.

*Evaluation Metrics.* We report both ranking and calibration-aware metrics. To evaluate ranking quality, we use Area Under the Curve (AUC). To access the calibration quality with normative objectives, we compute Kullback–Leibler (KL) divergence at top- $K$  ranks (noted as @ $K$ ) between the distribution of recommended content and the desired target distribution [1, 3, 10, 23, 27]. Lower KL values indicate better alignment with the normative target. Finally, we look at coverage of the candidate list, i.e., we aggregate all recommendations across users and compute the fraction of unique selected articles relative to the full candidate list.

*Model Distribution.* To estimate the marginal distribution over categories for the original model,  $p(c)$ , we aggregate predicted probabilities across all users’ candidate lists, yielding a global estimate of how the model allocates probability mass across categories.

*Normative Distributions.* We define two target distributions,  $\tilde{p}(c)$ , which serve as normative objectives for calibration:

- *Editorial distribution:* Computed by aggregating category frequencies across the union of users’ candidate lists. Since candidate lists are curated by editors, the resulting distribution approximates editorial intent.
- *Uniform distribution:* A distribution that promotes equal representation across all categories, defined over the set of unique categories present in the candidate lists.

## 5 Results and Discussions

### 5.1 Normative Distributions

We first evaluate how well the methods align the distribution of recommended articles with a given normative distribution. This analysis is conducted on the top-10 articles selected per user from the beyond-accuracy test set (see Section 4). We consider two variants of NAILS: deterministic (NAILS-det) selecting articles based on highest-probability scores, and a stochastic (NAILS-stoch) sampling articles according to their probability scores.

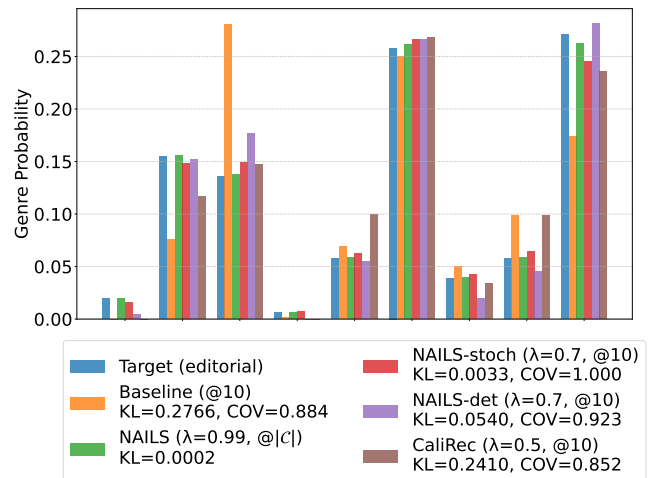
*Alignment Effectiveness.* Table 1 reports KL-divergence and coverage for each method across varying values of  $\lambda$ . As  $\lambda$  increases, KL-divergence consistently decreases for all methods, indicating improved alignment with the target normative distribution. However, for the editorial distribution, KL-divergence begins to increase again at higher  $\lambda$  values, suggesting that  $\lambda$  serves as a tunable parameter that balances alignment strength and ranking quality. CaliRec achieves the lowest KL@10 overall, which is expected given its greedy optimization procedure that explicitly minimizes KL-divergence. In contrast, NAILS nudges the distribution toward the target but does not directly optimize for KL, and therefore does not guarantee or force exact alignment.

As expected, NAILS-stoch achieves full coverage of the candidate pool, as articles are sampled probabilistically based on their aligned scores. While CaliRec’s coverage steadily decreases, NAILS-det increased as  $\lambda$  increases. This suggests that, with appropriate tuning, NAILS-det can better balance distributional alignment and content diversity compared to strictly optimization-based approaches.

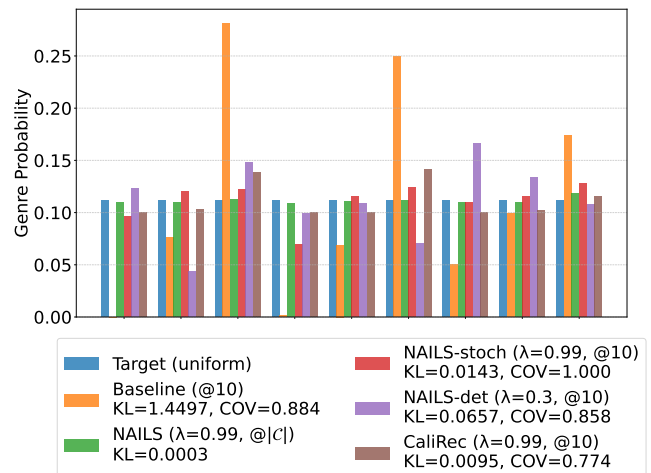
*Distributional Comparison.* Figure 1 visualizes the aggregated category distributions produced by the best-performing configuration of each method (selected from Table 1), shown alongside the target distribution, the uncalibrated baseline, and NAILS at top-C. These results show that NAILS effectively aligns with the stakeholder-defined distribution at the system level when aggregating recommendations across users. The comparison also highlights that the effectiveness of a given method depends on the alignment objective. For example, NAILS performs better under the editorial distribution, while CaliRec achieves closer alignment under the Uniform distribution.

### 5.2 Ranking Performance

We now turn to the effect of alignment on ranking quality, as measured by AUC. This analysis is conducted on the hidden test set used for evaluating ranking performance. We focus exclusively on deterministic methods, as our goal is to assess the effect of alignment in a stable and reproducible setting. Figure 2 illustrates how different alignment strengths ( $\lambda$ ) influence AUC under different target distributions. We observe that the choice of normative target distribution significantly affects ranking performance. In particular, certain distributions can even lead to improvements. For example, under the editorial distribution, applying alignment improves AUC for both CaliRec and NAILS, with NAILS achieving slightly higher performance overall. In contrast, alignment to the Uniform distribution generally reduces AUC, reflecting the challenge of balancing strict fairness with user relevance. These results highlight



(a) Editorial distribution



(b) Uniform distribution

**Figure 1: Category distribution among the top-10 recommended articles for alignment toward the Editorial distribution (a) and the Uniform distribution (b). We report KL-divergence (KL, ↓), aggregated across all users. KL@|C| refers to the predictive distribution over the full candidate list.**

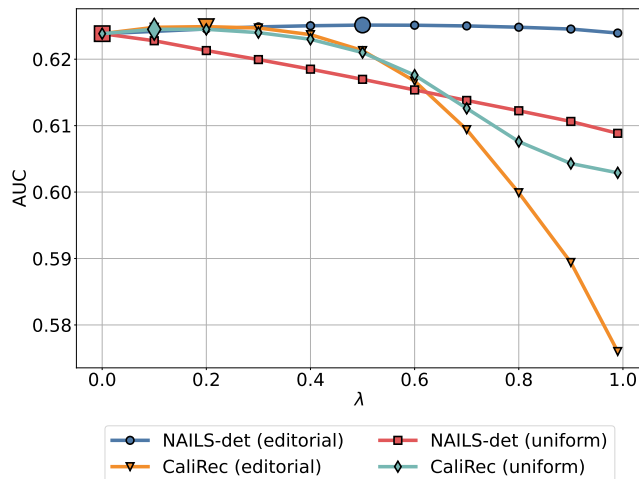
an important trade-off between ranking performance and alignment with normative goals. They also underscore the importance of carefully designing target distributions. Depending on the application, alternative targets could be constructed to encourage specific properties, such as greater content diversity, while minimizing the loss in engagement metrics.

## 6 Conclusion and Future Work

This paper proposed Normative Alignment of Recommender Systems via Internal Label Shift (NAILS), a novel method for aligning recommendation distributions with normative objectives. Our experiments demonstrate that NAILS effectively aligns the category distribution of recommendations. Moreover, we find that the choice

**Table 1: KL-divergence (KL, ↓) and coverage (COV) for the top-10 recommended articles under different alignment methods. We report KL as the mean KL-divergence computed per user.**

$\lambda$	Editorial distribution						Uniform distribution					
	NAILS-stoch		NAILS-det		CaliRec		NAILS-stoch		NAILS-det		CaliRec	
	KL@10	COV@10	KL@10	COV@10	KL@10	COV@10	KL@10	COV@10	KL@10	COV@10	KL@10	COV@10
0.0	2.060	100	3.802	88.4	3.802	88.4	3.889	100	4.872	88.4	4.872	88.4
0.1	2.052	100	3.707	89.7	1.630	90.3	3.670	100	3.905	88.4	2.774	88.4
0.2	2.046	100	3.626	92.3	0.735	88.4	3.473	100	3.318	87.1	1.442	86.5
0.3	2.036	100	3.555	92.3	0.574	85.8	3.290	100	<b>3.160</b>	85.8	0.481	83.2
0.4	2.030	100	3.499	92.9	0.512	83.9	3.129	100	3.253	84.5	0.084	80.6
0.5	2.028	100	3.460	92.9	<b>0.496</b>	85.2	2.983	100	3.414	80.6	0.035	78.7
0.6	2.026	100	3.434	92.9	0.502	84.5	2.858	100	3.602	80.0	0.029	77.4
0.7	<b>2.025</b>	100	<b>3.422</b>	92.3	0.527	83.9	2.745	100	3.783	79.4	0.029	77.4
0.8	2.026	100	3.425	92.3	0.562	83.9	2.650	100	3.916	76.8	0.028	77.4
0.9	2.027	100	3.441	92.3	0.572	85.2	2.564	100	3.994	74.2	0.028	77.4
0.99	2.031	100	3.464	91.6	0.571	81.9	<b>2.508</b>	100	4.034	73.5	<b>0.028</b>	77.4

**Figure 2: AUC (↑) across varying calibration strengths ( $\lambda$ ). Larger markers indicate the best performance.**

of target distribution can significantly influence both ranking performance and calibration outcomes. Compared to the established calibration method CaliRec, NAILS offers a lightweight and highly efficient alternative that avoids solving a greedy optimization problem during inference. More broadly, we do not claim universal superiority for any one method. The notion of optimality is inherently context-dependent and depends on the specific deployment goals, whether prioritizing ranking metrics like AUC, KL-divergence, or coverage. Our intention is to highlight the trade-offs embodied by different approaches. While the results are promising, future work is needed to explore the long-term effects of applying normative objectives in live environments, including their impact on user engagement, diversity, and system dynamics over time. Furthermore, although most prior alignment work has centered on a user-centric perspective, we argue that extending calibration to align with global normative goals across users presents a promising

and necessary direction for future research, particularly in editorial and policy-driven recommendation systems. Ultimately, incorporating normative alignment frameworks like NAILS could play a key role in building responsible recommendation systems that are aligned with broader normative objectives, such as societal values.

## Acknowledgments

We would like to extend our gratitude to and acknowledge our employers and funding bodies, including Ekstra Bladet, JP/Politikens Media Group, Technical University of Denmark, Copenhagen Business School, Innovation Foundation Denmark (grant number 1044-00058B), and the Platform Intelligence in News-Project (grant number 0175-00014B).

## References

- [1] Himan Abdollahpouri, Zahra Nazari, Alex Gain, Clay Gibson, Maria Dimakopoulou, Jesse Anderton, Benjamin Carterette, Mounia Lalmas, and Tony Jebara. 2023. Calibrated Recommendations as a Minimum-Cost Flow Problem. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (Singapore, Singapore) (WSDM '23). Association for Computing Machinery, New York, NY, USA, 571–579. <https://doi.org/10.1145/3539597.3570402>
- [2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) (KDD '19). Association for Computing Machinery, New York, NY, USA, 2623–2631. <https://doi.org/10.1145/3292500.3330701>
- [3] Jiayi Chen, Wen Wu, Liye Shi, Yu Ji, Wenxin Hu, Xi Chen, Wei Zheng, and Liang He. 2022. DACSR: Decoupled-Aggregated End-to-End Calibrated Sequential Recommendation. *Applied Sciences* 12, 22 (2022). <https://doi.org/10.3390/app122211765>
- [4] Yashar Deldjoo, Fatemeh Nazari, Arnau Ramisa, Julian McAuley, Giovanni Pellegrini, Alejandro Bellogin, and Tommaso Di Noia. 2023. A Review of Modern Fashion Recommender Systems. *ACM Comput. Surv.* 56, 4, Article 87 (Oct. 2023), 37 pages. <https://doi.org/10.1145/3624733>
- [5] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. 2020. A unified view of label shift estimation. *Advances in Neural Information Processing Systems* 33 (2020), 3290–3300.
- [6] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (Barcelona, Spain) (RecSys '10). Association for Computing Machinery, New York, NY, USA, 257–260. <https://doi.org/10.1145/1864708.1864761>
- [7] Natali Helberger. 2019. On the Democratic Role of News Recommenders. *Digital Journalism* 7, 8 (2019), 993–1012.

- [8] Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2024. Train Once, Use Flexibly: A Modular Framework for Multi-Aspect Neural News Recommendation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9555–9571. <https://doi.org/10.18653/v1/2024.findings-emnlp.558>
- [9] Dietmar Jannach and Michael Jugovac. 2019. Measuring the Business Value of Recommender Systems. *ACM Trans. Manage. Inf. Syst.* 10, 4, Article 16 (dec 2019), 23 pages. <https://doi.org/10.1145/3370082>
- [10] Hyunsik Jeon, Se-eun Yoon, and Julian McAuley. 2024. Calibration-Disentangled Learning and Relevance-Prioritized Reranking for Calibrated Sequential Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (Boise, ID, USA) (CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 973–982. <https://doi.org/10.1145/3627673.3679728>
- [11] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 7, 1, Article 2 (dec 2016), 42 pages.
- [12] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG] <https://arxiv.org/abs/1412.6980>
- [13] Daphne Koller and Mehran Sahami. 1997. Hierarchically Classifying Documents Using Very Few Words. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 170–178.
- [14] Johannes Kruse, Kasper Lindskow, Michael Riis Andersen, and Jes Frellsen. 2023. Creating the next generation of news experience on ekstrabladet.dk with recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems (Singapore, Singapore) (RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1067–1070. <https://doi.org/10.1145/3604915.3610248>
- [15] Johannes Kruse, Kasper Lindskow, Saikishore Kalloori, Marco Polignano, Claudio Pomo, Abhishek Srivastava, Anshuk Uppal, Michael Riis Andersen, and Jes Frellsen. 2024. EB-NeRD a large-scale dataset for news recommendation. In *Proceedings of the Recommender Systems Challenge 2024 (Bari, Italy) (RecSysChallenge '24)*. Association for Computing Machinery, New York, NY, USA, 1–11. <https://doi.org/10.1145/3687151.3687152>
- [16] Johannes Kruse, Kasper Lindskow, Saikishore Kalloori, Marco Polignano, Claudio Pomo, Abhishek Srivastava, Anshuk Uppal, Michael Riis Andersen, and Jes Frellsen. 2024. RecSys Challenge 2024: Balancing Accuracy and Editorial Values in News Recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems (Bari, Italy) (RecSys '24)*. Association for Computing Machinery, New York, NY, USA, 1195–1199. <https://doi.org/10.1145/3640457.3687164>
- [17] Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. 2022. MINER: Multi-Interest Matching Network for News Recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 343–352. <https://doi.org/10.18653/v1/2022.findings-acl.29>
- [18] Feng Lu, Anca Dumitrache, and David Graus. 2020. Beyond Optimizing for Clicks: Incorporating Editorial Values in News Recommendation. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (Genoa, Italy) (UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 145–153. <https://doi.org/10.1145/3340631.3394864>
- [19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL] <https://arxiv.org/abs/1301.3781>
- [20] Efrat Nechushtai and Seth C. Lewis. 2019. What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior* 90 (2019), 298–307. <https://doi.org/10.1016/j.chb.2018.07.043>
- [21] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. News Recommendation with Candidate-aware User Modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (Madrid, Spain) (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 1917–1921. <https://doi.org/10.1145/3477495.3531778>
- [22] Marco Saerens, Patrice Latine, and Christine Decaestecker. 2002. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation* 14, 1 (2002), 21–41.
- [23] Sinan Seymen, Himan Abdollahpour, and Edward C. Malthouse. 2021. A Constrained Optimization Approach for Calibrated Recommendations. In *Proceedings of the 15th ACM Conference on Recommender Systems (Amsterdam, Netherlands) (RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 607–612. <https://doi.org/10.1145/3460231.3478857>
- [24] Ryotaro Shimizu, Takashi Wada, Yu Wang, Johannes Kruse, Sean O'Brien, Sai HtaungKham, Linxin Song, Yuya Yoshikawa, Yuki Saito, Fugee Tsung, Masayuki Goto, and Julian McAuley. 2025. Disentangling Likes and Dislikes in Personalized Generative Explainable Recommendation. In *Proceedings of the ACM on Web Conference 2025 (Sydney NSW, Australia) (WWW '25)*. Association for Computing Machinery, New York, NY, USA, 4793–4809. <https://doi.org/10.1145/3696410.3714583>
- [25] Ryotaro Shimizu, Yu Wang, Masanari Kimura, Yuki Hirakawa, Takashi Wada, Yuki Saito, and Julian McAuley. 2024. A Fashion Item Recommendation Model in Hyperbolic Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 8377–8383.
- [26] Carlos N. Silla and Alex A. Freitas. 2011. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery* 22, 1 (2011), 31–72.
- [27] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems (Vancouver, British Columbia, Canada) (RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 154–162. <https://doi.org/10.1145/3240323.3240372>
- [28] Jonathan Stray, Ivan Vendrov, Jeremy Nixon, Steven Adler, and Dylan Hadfield-Menell. 2021. What are you optimizing for? Aligning Recommender Systems with Human Values. arXiv:2107.10939 [cs.LR] <https://arxiv.org/abs/2107.10939>
- [29] Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten De Rijke. 2022. RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems (Seattle, WA, USA) (RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 208–219. <https://doi.org/10.1145/3523227.3546780>
- [30] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a Mission: Assessing Diversity in News Recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval (Canberra ACT, Australia) (CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 173–183. <https://doi.org/10.1145/3406522.3446019>
- [31] Sanne Vrijenhoek, Lien Michiels, Johannes Kruse, Alain Starke, Nava Tintarev, and Jordi Viader Guerrero. 2023. NORMalize: The First Workshop on Normative Design and Evaluation of Recommender Systems. In *Proceedings of the 17th ACM Conference on Recommender Systems (Singapore, Singapore) (RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1252–1254. <https://doi.org/10.1145/3604915.3608757>
- [32] Chenyang Wang, Yankai Liu, Yuanqing Yu, Weizhi Ma, Min Zhang, Yiqun Liu, Haitao Zeng, Junlan Feng, and Chao Deng. 2023. Two-sided Calibration for Quality-aware Responsible Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems (Singapore, Singapore) (RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 223–233. <https://doi.org/10.1145/3604915.3608799>
- [33] Shuhei Watanabe. 2023. Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance. arXiv:2304.11127 [cs.LG] <https://arxiv.org/abs/2304.11127>
- [34] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Multi-Head Self-Attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 6389–6394. <https://doi.org/10.18653/v1/D19-1671>
- [35] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 3597–3606.
- [36] Xing Zhao, Ziwei Zhu, Majid Alfiifi, and James Caverlee. 2020. Addressing the Target Customer Distortion Problem in Recommender Systems. In *Proceedings of The Web Conference 2020 (Taipei, Taiwan) (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 2969–2975. <https://doi.org/10.1145/3366423.3380065>