Counterfactual Metrics for Auditing Black-Box Recommender Systems for Ethical Concerns

Nil-Jana Akpinar [*1] **Liu Leqi** [*1] **Dylan Hadfield-Menell** [2] **Zachary Lipton** [1]

Abstract

Recommender systems can shape peoples' online experience in powerful ways which makes close scrutiny of ethical implications imperative. Most existing work in this area attempts to measure induced harm exclusively based on observed recommendations under a set policy. This neglects potential dependencies on other quantities and can lead to misleading conclusions about the behavior of the algorithm. Instead, we propose counterfactual metrics for auditing recommender systems for ethical concerns. By asking how recommendations would change if users behaved differently or if the training data was different, we are able to isolate the effects of the recommendation algorithm from components like user preference and information. We discuss the ethical context of the suggested metrics and propose directions for future work.

1. Introduction

Recommender systems are socio-technical systems that play an active role in shaping peoples' online experience by moderating which news, social media content, and products are most readily available to them. They influence preferences, beliefs, and choices on an individual level and hold the power to sway public opinion which can lead to undesired consequences for users and society at large (e.g. Rafailidis & Nanopoulos, 2016; Burki, 2019; Milano et al., 2020). In recent years, government organizations and academics have called for more ethical scrutiny in evaluating recommender systems which has led to an active—but fragmented—area of research including, for example, efforts to measure and mitigate (demographic) bias (Chen et al., 2020; Patro et al., 2022) and efforts to quantify user agency (e.g., notions of reacheability (Dean et al., 2020; Curmei et al., 2021).

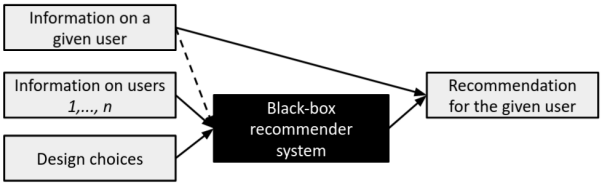

Figure 1. Dependency graph of the recommendation for a given user. The information of the given user may or may not be used to train a recommendation policy.

Auditing recommender systems for ethical concerns has shown to be a difficult task that comes with a complex set of normative and technical questions. We focus on the technical side and use 'recommender system' as a collective term to describe both the algorithm that devises a recommendation policy based on training data and said recommendation policy that maps information for a given user to recommendations for the user. Most previous work implicitly assumes that a recommender system can be audited by only considering the output recommendations. Yet, output recommendations depend on a multitude of factors including users' past behavior, and algorithmic design choices which imposes a complex dependency graph (see Figure 1). Consideration of these confounding factors is crucial in order to reliably determine whether potential harm is *caused by the recommender system*.

We address this conceptual problem by proposing counterfactual metrics for harm auditing which view recommendation algorithms not just through its output recommendations. Our metrics are build around an interventional perspective on algorithm auditing asking how recommendendations would change if the information of one or several users was different. Considering potential outcomes under different training data sets and user inputs allows us to disentangle the effects of the recommendation algorithm we want to audit from the impact of preferences and information of users. Our counterfactual metrics span a wide range of ethical concerns including user agency, stability, personalisation, diversity and fairness.[1] Since real-life recommendation al-

[*]Equal contribution [1]Carnegie Mellon University, Pittsburgh, USA [2]EECS CSAIL, MIT, Cambridge, USA. Correspondence to: Nil-Jana Akpinar <nakpinar@andrew.cmu.edu>, Liu Leqi <leqil@andrew.cmu.edu>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

[1]Due to space constraint, we only focus on user agency and

gorithms are generally proprietary, we only assume access to inputs in the form of user interactions with content over time and black-box recommendation outputs.

The remainder of the paper is structured as follows. In Section 2, we discuss some traditional metrics for auditing recommender systems for harm. We argue for the need of counterfactual metrics and instantiate our general counterfactual auditing framework with two metrics in Section 3. The proposed metrics are set into context with the wider debates and taxonomy of ethical concerns in recommender systems in Section 4.

2. Observational metrics

Past work has proposed a multitude of metrics for ethical concerns in the recommendation setting including measures of fairness (Patro et al., 2022; Chen et al., 2020), moral appropriateness of content (Tang & Winoto, 2015), stability (Adomavicius & Zhang, 2012), and diversity (Nguyen et al., 2014; Silveira et al., 2019; Parapar & Radlinski, 2021). The majority of these metrics exclusively relies on observational quantities under a set recommendation policy and thereby neglects potential effects caused by the users' behaviors and preferences. As an example, we consider the problem of diversity in recommendations. In many recommendation settings, suggesting content with a variety of different topics is regarded as desirable as it honors users' multi-faceted interests and avoids algorithmic profiling (Milano et al., 2020). This content diversity is generally measured by calculating some sort of inverse similarity in recommendation slates or in recommendations over time (e.g. Silveira et al., 2019; Nguyen et al., 2014). Repeatedly neglecting diversity in recommended content has been observed to lead to increasingly narrow (and sometimes niche) recommendations over time (WSJ, 2021) which can facilitate a problem of filter bubbles (Pariser, 2011) and is often measured in terms of observational quantities such as the Jaccard indices between recommended items (e.g. Chaney et al., 2018; Lunardi et al., 2020).

Underlying virtually all measures of diversity (and criticisms of lack-of-diversity) in recommendations is the assumption that missing diversity necessarily points to a flawed recommendation algorithm. While this is true in many cases, we argue that there exist settings in which a user's preference with regard to a given set of items or content is truly narrow and instead of unwanted algorithmic profiling, narrow recommendations in fact affirm a user's self-identity as member of a social group. For example, picture an Instagram user who is exclusively watching dog videos even if other videos are suggested, or a researcher who uses Twitter only for academic purposes deciding to only follow other researchers. In order to differentiate scenarios like this from cases in which recommendation algorithms inflict harm on users, we need to control for user preferences and other external factors. This requires causal reasoning and counterfactual metrics for the ethical concerns posed by recommender systems.

stability in this paper.

3. Counterfactual metrics

We consider an auditing setup where a dataset is used to train a recommendation policy. After deployment, the recommendation policy is used to recommend an item from a finite set based on the user's information. The given user may or may not belong to the training data set. Our goal is to audit the recommendation algorithm through this one-step recommendation procedure after training. More specifically, we assume a training dataset $\mathcal{D} = \{\tau_1, \dots, \tau_n\}$ that contains information $\tau_i \in \mathcal{D}$ (e.g., a collection that contains ratings, demographics, reviews, etc.) for multiple users $i \in [n]$. For any given user with information τ (τ may or may not belong to $\mathcal{D}$), a (possibly random) recommender system $\mathcal{A}$ outputs the individual's next-step recommendation $\mathcal{A}(\mathcal{D}, \tau) \in \mathbb{C}$ where $\mathbb{C}$ is a finite set of all possible recommendations. Under this notation, $\mathcal{A}(\cdot, \cdot)$ is referred as the recommender system (algorithm), and $\mathcal{A}(\mathcal{D}, \cdot)$ is the recommendation policy under dataset $\mathcal{D}$.

Observational v.s. counterfactual metrics As we have discussed above, the vast majority of metrics defined for recommender systems are observational—they are defined upon the current $\mathcal{D}$, τ and $\mathcal{A}(\mathcal{D}, \tau)$. The output recommendation $\mathcal{A}(\mathcal{D}, \tau)$ is a consequence of (i) the training data $\mathcal{D}$, (ii) the user's own information τ, and (iii) the algorithm $\mathcal{A}$ itself (Figure 1). Metrics defined solely upon $\mathcal{A}(\mathcal{D}, \tau)$ cannot be used to assess properties of $\mathcal{A}$. The goal of counterfactual metrics is to inspect properties of the algorithm $\mathcal{A}(\cdot, \cdot)$ by considering output of $\mathcal{A}$ under training data and user information that deviate from $\mathcal{D}$ and τ. Doing so allows us to understand the behavior of the algorithm and audit the system by teasing apart different causes of harmful recommendations.

3.1. A general counterfactual audit procedure

Counterfactuals are answers to *what-if* questions. For example, in the recommender system setting, one may be interested in what the change in recommendation for a user would be if they were of a different age group or if they had rated a movie differently. Counterfactual metrics are defined upon these *what-if* scenarios. Procedurally, there are three steps to obtain a counterfactual metric:

1. Decide the treatment space $\mathcal{W}$ that contains permissable new training datasets and/or new information of the indi-

vidual of interest. For example, the treatment space could contain removing ratings in the information τ of the user or it may contain updated training data by deleting some entries in the original $\mathcal{D}$.

2. Define the outcome of interest which we select to be the next-step recommendation denoted by Y. We formally assume a random treatment denoted by W which, in some settings, is fixed at $W = w$ for a $w \in \mathcal{W}$. We use Y^w to refer to the potential outcome[2] under treatment $W = w \in \mathcal{W}$. For example, if $\mathcal{W}$ contains permissible new training datasets, then $Y^w = \mathcal{A}(w, \tau)$ denotes the recommendation that the user with information τ would have obtained if the recommender were to be trained using dataset w. One key assumption we rely on is *consistency*: $Y = Y^w$ when $W = w$. That is, if the treatment w is applied, the observed outcome Y is the potential outcome Y^w.

3. Define the counterfactual metric upon the potential outcome Y^w. In the following, we present individual-level counterfactual metrics (Definition 3.1, 3.2).

In general, it is hard to obtain counterfactuals (Y^w), since we cannot go to a parallel world where the treatment is applied to observe the outcome. In our case, we are given black-box access to the algorithm $\mathcal{A}$ to audit the recommender system. That is, we do not require to know the inner workings of $\mathcal{A}$ but only need to be able to query $\mathcal{A}$. In such settings, we can *simulate* the parallel world under which the training data or the information of the given user is replaced. It is worth noting that when the assignment of a treatment $W = w$ is *independent* to what the potential outcome Y^w would be, we can translate the potential outcome Y^w to a conditional that we can estimate:

$$Y^w \;=\; \{Y^w | W = w\} \;=\; \{Y | W = w\}.$$

Here, the first equality follows from the independence between Y^w and W and the second follows from consistency. For the rest of this paper, W is deterministic (or purely random), which ensures independence of Y^w and W.

3.2. Individual-level metrics

Under our framework, we present some individual-level metrics. That is, we are interested in answering counterfactual questions for a particular user with information τ. Throughout, we will both see how some existing metrics are in fact examples of counterfactual metrics and propose new metrics. The proposed metrics have a natural connection to ethical concerns on user autonomy and personal identity in recommender systems which we discuss in Section 4.

[2]Note that, although we borrow the language from the potential outcome framework, our goal is not to estimate a treatment effect but to inspect new recommendations under treatments.

For the first metric, we assume a setting in which user information comprises ratings for a subset of items from $\mathbb{C}$.

Definition 3.1 (Individual-level Reacheability)**.** As proposed by Dean et al. (2020); Curmei et al. (2021), we say an item is *reachable* by a user if there is an allowable modification to their rating history causes the item to be recommended. In this case, τ is a vector of length $|\mathbb{C}|$ with entries $\tau_j[k]$ denoting j's rating for item k (with n/a if not available). Under our framework, if we want to audit whether item k is *reacheable* by user j with $\tau_j \in \mathcal{D}$, we have the following:

- Treatment space $\mathcal{W} = \{\tau' : \sum_{t \in [|\mathbb{C}|]} \mathbf{1}\{\tau_j[t] \neq \tau'[t]\} \leq B\}$. That is, $\mathcal{W}$ contains new user information that deviates from τ_j by at most B entries, where $B \in \mathbb{N}$ is a pre-specified budget.
- The outcome of interest Y is the next-step recommendation for user j: $Y^w = \mathcal{A}(\mathcal{D}', w)$ where $w \in \mathcal{W}$ and $\mathcal{D}' = \{\tau_1, \ldots, \tau_{j-1}, w, \tau_{j+1}, \ldots, \tau_n\}$.
- The reacheability metric is defined to be

$$\max_{w \in \mathcal{W}} \mathbb{P}_{\mathcal{A}} \left(Y^w = k\right), \tag{1}$$

  which gives the maximal probability for user j to *reach* item k by modifying their own information τ_j. We note that $\mathbb{P}_{\mathcal{A}}$ is used to indicate stochasticity in $\mathcal{A}$.

Intuitively, reacheability is a measure of user agency under recommendation policy $\mathcal{A}$. For example, consider a job recommendation setting in which a job seeker pivots towards a new job type or industry in their search. If the desired job postings are available but not recommended to the user despite several changes to their search history, profile, etc., we say that the postings are not *reachable* by the user. Under certain circumstances (e.g. if the user has all the required qualifications for the job), missing reachability suggests that the algorithm does not grant sufficient agency over recommendations to the user.

Dean et al. (2020); Curmei et al. (2021) study ways of efficiently computing the exact value of individual-level reacheability (1) for a class of score-based recommender systems. Under our general framework, it is clear that the reacheability metric can have several extensions that resemble the practice more closely: (i) Instead of the current treatment space $\mathcal{W}$, one may define it by accounting for the fact that certain changes to τ_j may not be feasible and users may be more likely to edit certain information. (ii) When facing a more sophisticated recommender system, instead of the *maximal* reacheable probability $\max_{w \in \mathcal{W}} \mathbb{P}\left(Y^w = k\right)$, one may care about the *average* reacheable probability given by $\mathbb{E}_W \mathbb{P}\left(Y^W = k\right)$ where W is purely random.

We now propose a new individual-level counterfactual metric, termed as *stability*.

Definition 3.2 (Individual-level Stability)**.** We propose this metric to capture how stable a user's recommendations are to other users' behaviors. In settings in which we are interested in user j's recommendation (where $\tau_j \in \mathcal{D}$), we specify the following:

- Treatment space $\mathcal{W} = \{\mathcal{D}' : \mathcal{D}'$ differs from $\mathcal{D}$ for at most B users and τ_j remains unchanged$\}$, where $B \in \mathbb{N}$ is a pre-specified budget. In other words, $\mathcal{D}'$ may differ from $\mathcal{D}$ by changing, adding or deleting at most B entries (except τ_j).
- The outcome of interest Y is the next-step recommendation for user j under new training data w: $Y^w = \mathcal{A}(w, \tau_j)$ where $w \in \mathcal{W}$.
- The stability metric is defined to be

$$\max_{w \in \mathcal{W}} d\left(\mathcal{A}(\mathcal{D}, \tau_j), Y^w\right), \tag{2}$$

  where $d : \mathbb{C} \times \mathbb{C} \to \mathbb{R}_+$ is a pre-speficied measure of distance between two recommendations.

Individual-level stability measures how much a user's recommendation changes with updates in the information on other users. The distance measure d between two recommendations is subjective and captures what constitutes as "big" changes in recommendations. One may decide that changing a Husky video to an Alaskan video does not count as a big change compared to changing it to a political video. Similar as above, one may define many variants of (2). For example, instead of finding the maximal discrepency in recommendations, one may replace $\max_{w \in \mathcal{W}}$ by $\mathbb{E}_W$ where W is purely random.

Related work Some prior workon inspecting recommender systems can be conceptualized as counterfactual auditing. For example, Yao et al. (2021) trainrecommendation policies using simulated users with different behavioral models and analyze how the recommendations differ under different simulated users. This is a counterfactual audit where training data under various behavioral models has been chosen as the treatment space. Many social media platforms (e.g., Facebook) have provided a functionality for users to understand why certain recommendations have been made to them. The nature of this functionality is of counterfactual flavor—would the user still be recommended the same post if they had behaved differently?

4. Ethical considerations

Several lines of work have aimed at identifying a taxonomy of ethical concerns in algorithms (Mittelstadt et al., 2016; Jobin et al., 2019; Tsamados et al., 2022; Milano et al., 2021). In the recommendation context, Milano et al. (2020) identify six general areas of concern including inappropriate content, privacy, autonomy & personal identity, opacity, fairness, and wider social effects. Reachability and stability touch on several of these categories.

As argued in Section 3, we interpret reachability as a measure of users' agency over their recommendations which can be regarded as a matter of autonomy and personal identity. Recommendation algorithms often explicitly or implicitly assign users categories that do not necessarily align with recognizable social attributes that the users would identify themselves (Milano et al., 2020). This 'algorithmic profiling' can lead to negative user experience as it clashes with the users' perception of personal identity (de Vries, 2010; Leese, 2014). A lack of reachability of content outside one's assigned categories may point towards a system that over-categorized users in this way.

Stability has a natural connection to user autonomy and non-comparative fairness. Consider a scenario in which addition or removal of new user information to the training data set drastically changes the recommendations for an unrelated user. We argue that in this case, the algorithm does not sufficiently value the user's autonomy over their recommendations. The change in recommendations does not follow consistent criteria and can be regarded as somewhat arbitrary which ties the concept of stability to ideas around leave-one-out unfairness (Black & Fredrikson, 2021).

While missing reachability can directly harm users' experience and perceived utility, a lack of stability can also be understood as an indirect potential for harm. Missing stability renders a recommender system vulnerable to manipulation by a small group of users which is able to influence what others are recommended through high levels of interaction with the respective content (Howard et al., 2019). We note that stability is a property of the recommendation algorithm and does not make normative claims on the content promoted in these settings which, dependent on context, could be judged as desirable.

5. Future Work

There is a rich line of future work we aim to pursue. The current counterfactual metrics are at the individual-level. Most of the time, auditing recommender system requires an inspectation of the system across a population of users. We plan to design group/population-level counterfactual metrics that capture the effect of the algorithm across a set of individuals. Another important extension to our work is to account for the following two dependencies: (i) Dependency among time steps in τ: For a single trajectory, a treatment that applies to a particular past time step may influence future time steps; and (ii) Dependency among trajectories: When a treatment is applied to the history of one user (that belongs to the training dataset), it may influence other users' trajectories.

References

Adomavicius, G. and Zhang, J. Stability of recommendation algorithms. *ACM Transactions on Information Systems*, 30(4):1–31, November 2012.

Black, E. and Fredrikson, M. Leave-one-out unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 285–295, New York, NY, USA, 2021. Association for Computing Machinery.

Burki, T. Vaccine misinformation and social media. *The Lancet Digital Health*, 1(6):e258–e259, October 2019.

Chaney, A. J. B., Stewart, B. M., and Engelhardt, B. E. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, September 2018.

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., and He, X. Bias and debias in recommender system: A survey and future directions. Oct 2020. URL http://arxiv.org/abs/2010.03240.

Curmei, M., Dean, S., and Recht, B. Quantifying availability and discovery in recommender systems via stochastic reachability. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 2265–2275. PMLR, 2021.

de Vries, K. Identity, profiling algorithms and a world of ambient intelligence. *Ethics and Information Technology*, 12(1):71–85, January 2010.

Dean, S., Rich, S., and Recht, B. Recommendations and user agency: The reachability of collaboratively-filtered information. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 436–445, New York, NY, USA, 2020. Association for Computing Machinery.

Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., and François, C. The ira, social media and political polarization in the united states, 2012-2018. 2019.

Jobin, A., Ienca, M., and Vayena, E. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1 (9):389–399, Sep 2019.

Leese, M. The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the european union. *Security Dialogue*, 45(5):494–511, September 2014.

Lunardi, G. M., Machado, G. M., Maran, V., and de Oliveira, J. P. M. A metric for filter bubble measurement in recommender algorithms considering the news domain. *Applied Soft Computing*, 97:106771, December 2020.

Milano, S., Taddeo, M., and Floridi, L. Recommender systems and their ethical challenges. *AI Soc.*, 35(4):957–967, December 2020.

Milano, S., Taddeo, M., and Floridi, L. Ethical aspects of multi-stakeholder recommendation systems. *Information Society*, 37(1):35–45, 2021.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), December 2016.

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., and Konstan, J. A. Exploring the filter bubble. In *Proceedings of the 23rd international conference on World wide web - WWW '14*. ACM Press, 2014.

Parapar, J. and Radlinski, F. Towards unified metrics for accuracy and diversity for recommender systems. In *Fifteenth ACM Conference on Recommender Systems*. ACM, September 2021.

Pariser, E. *The filter bubble*. Penguin Press, May 2011.

Patro, G. K., Porcaro, L., Mitchell, L., Zhang, Q., Zehlike, M., and Garg, N. Fair ranking: a critical review, challenges, and future directions. Jan 2022. URL http://arxiv.org/abs/2201.12662.

Rafailidis, D. and Nanopoulos, A. Modeling users preference dynamics and side information in recommender systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(6):782–792, 2016.

Silveira, T., Zhang, M., Lin, X., Liu, Y., and Ma, S. How good your recommender system is? a survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, 10(5):813–831, May 2019.

Tang, T. Y. and Winoto, P. I should not recommend it to you even if you will like it: the ethics of recommender systems. *New Review of Hypermedia and Multimedia*, 22 (1-2):111–138, July 2015.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., and Floridi, L. The ethics of algorithms: key problems and solutions. *AI & society*, 37(1):215–230, Mar 2022.

WSJ. Investigation: How tiktok's algorithm figures out your deepest desires. *The Wall Street Journal*, Jul 2021. URL https://tinyurl.com/bdftvc7y.

Yao, S., Halpern, Y., Thain, N., Wang, X., Lee, K., Prost, F., Chi, E. H., Chen, J., and Beutel, A. Measuring recommender system effects with simulated users. Jan 2021. URL https://arxiv.org/abs/2101.04526.