The Bias of Recommender Systems and Impact on Social Culture

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Richard J Park

Spring 2021

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Sean M. Ferguson, Department of Engineering and Society

The Bias of Recommender Systems and its Impact on Social Culture

Recommender algorithms rank the order in which information is presented to us, guaranteeing the most relevant and accurate results. These systems are revolutionary to the age of information as they help billions of users evaluate the endless stream of online decisions. Recommender systems provide an easier decision-making process to a user's digital experience, however with the algorithmic biases in these systems, users are presented a recommendation that infringes on the autonomy of online interactions. Using explainable recommendations, I present a possible solution to algorithmic bias that increases trust of users and improves user autonomy within the online decision-making process by bridging the gap between users and the recommendation algorithms. Using a combination of frameworks for analysis, I will first identify the bias in recommender systems, then proceed to show how explainable recommendations can limit the bias's effect on user autonomy at each level of the framework. It is the goal of an online service to provide a consumer with the most relevant and matching results, but due to some of the algorithmic biases, these recommendations produce a systemic effect on social culture.

Recommender systems are important for our digital experience and are necessary for the success of online services, but these recommendation systems hold such power to shape the ways in which audiences discover information. With this power, the recommendation algorithms are designed to act as an agent in computationally shaping social culture (Morris, 2015). By using the actor network theory and previous works in algorithmic bias theories as a framework for analysis, we can identify the service providers, recommender systems, algorithmic bias, and users all as agents that affect the culture of our society.

Framework for evaluating bias in RS

Ideas dealing with the bias in computer systems has begun decades ago with researchers like Batya Friedman (1996). For decades now, the growing research in machine learning and neural networks has increased the literature for algorithmic bias, leading to major companies like Facebook, Google, and Microsoft devoting teams specifically to research algorithmic bias. In Friedman's work, there are categories for bias in computer systems: preexisting bias, technical bias, and emergent bias. By using this similar framework to classify bias in computer systems, we can evaluate the bias in recommender systems that ultimately encroaches on our autonomy of choice.

In the context of recommendations and algorithmic fairness, bias refers to the negative impact of the system's model that disproportionately favors or disfavors particular groups. With this interpretation, all recommender systems have characteristics of the data or system itself that are intended or unintended to be biased. All data sets and all algorithms have some type of bias, therefore RS are vulnerable to biased inputs. This is inevitable as data itself is curated by sampling a specific population. Researchers at Microsoft have examined this bias and differentiated it into two levels of bias: data gathering and data usage. Using the Friedman taxonomy of computer bias and Olteanu (2019) taxonomy of data bias, we can label the bias observed in data collection as preexisting bias since this is a result of the practices of data collection.

Social Impact of RS bias

Using these multiple frameworks for identifying bias in computer systems and big data, they can be summarized into 3 levels of entry for bias within recommender systems: data, algorithmic, and outcome (Cramer et al., 2018). The rapid flow of data further brings the need for recommender systems to better predict user preferences. Engineers in this field are currently working on cutting-edge technologies to improve the recommendation algorithms, but there are no industry standards to tend the unintended biases of these algorithms. These biases may create biased choices in online decision-making, which ultimately infringes on the autonomy of online users. In practice, throughout each step in the algorithmic process of recommendations: the data collection, creation of social models, and resulting recommendations, there exists some form of bias that must be handled, in which I will pose explainable recommendations to be a solution.

Personalized recommendations are based on two main strategies of content filtering and collaborative filtering. Content filtering involves a method of creating profiles for each user that contains characteristics to classify the nature of the user. Content filtering requires the need for external information that may not always be available, but can be gained by questionnaires, etc. A successful example of this is the Music Genome Project used by the internet radio service, Pandora (Koren et al.,2009).

When user profiles are created during the collaborative filtering algorithms, they may unintentionally produce biases as an effect of creating profiles out of data collected by user activity. This concept is referred to as a type of algorithmic profiling (Milano et al., 2020). When an algorithm constructs models of users that reproduce social categories, they unintentionally introduce biases if the produced social categories do not align with our recognizable social categories. Our online identities are reflected by the algorithmic categorization and promote the idea of personal identity that is dynamically changing with every action we take online. Since the recommender system's model is continuously changing, this type of labelling often does not match with our self-identifying labels. The system is also limited to the input data users give. The RS may consider different attributes more significant than others simply based on the goal of the services. Our goal as users is much different than the corporate service, therefore it is likely that the suggestions given are not a reflection of ourself, but rather a suggestion that would most benefit the recommender system and service provider.

Through each step of the algorithmic profiling process, we can use the 3 categories of bias entry (Cramer et al., 2018) to identify where a certain bias may be introduced. Since we have already established that any inputted data in an RS can be considered data bias, the next step is to identify algorithmic bias in RS. These biases are present at the level of the programmers and teams that create the RS.

The interest of online services is to make their items available and have recommendations that lure consumers to their products. The goal of the recommender system is to make the recommendations based on the given input data and other metadata that the company may have access to. The recommendation is then presented to the users, who act upon these recommendations and have an interest in receiving the most relevant recommendations. Once the recommendation is received, there exists a feedback loop in content-based recommendations where the user can interact back with the system to provide better recommendations in the future. Within these feedback loops, since we have already determined that bias feature may exist in data collection and algorithmically, these loops create bias in the outcome by manipulating recommended content (Baeza-Yates, 2016). This feature of recommendations poses ethical questions of manipulation that can ultimately affect our culture.

RS Manipulation

Recommender systems are vulnerable to unforeseen groups whose goal is to manipulate the feedback cycle between the user and system (Milano et al., 2020). For example, if a group of active users were to interact with the recommender system and drive-up positive feedback for certain items/services, it is likely that the item will be recommended for others. This can be the case for social networks, streaming platforms, and news systems. The nature of content-based filtering isolates users into a bubble of self-reinforcing ideologies that limits exposure to contrasting viewpoints since these contrasting ideas do not result in more user retention. This social effect is damaging to society and harms the function of public debate and democratic institutions (Milano et al., 2020). By this method, recommender systems are vulnerable to propaganda attacks in the circulation of media, ultimately effecting the ways in which information is presented to us.

In many cases of RS, the social effect caused by the self-reinforcing ideologies is overlooked by major companies, since the interests of online services, RS, and users all differ. We can use the actor network theory to observe how service providers, recommender systems, and consumers/users are actors within the same network, but hold different interests. It is in the best interest of online services to retain their consumers and have their services be chosen over others. On the other hand, recommender systems are designed by programmers who have the verdict in which metrics to maximize in order to give a relevant result. The interest of the RS is to maximize relevant results and provide recommendations based on the service's needs (usually to maximize user retention). The interests of consumers/users of an RS are to get the most accurate recommendations, then give feedback to reinforce the RS to give better recommendations next time. Since all of the actors have different interests, there are manipulative traps that must be made aware to avoid possible social effects. The results from an RS may be measured through other mathematical error measures, but online services measure the results from recommender systems based on click/view-through rates and user satisfaction in production (Beer, 2009). This poses the question of whether metrics for the recommender systems are defined by the engineers or the business model of the company.

Recommender systems may appear as “sticky traps” in which their purpose is to entice and hook user into long time usage of their services (Seaver, 2018). In the long term, this causes biases in the recommender systems that encroach on the autonomy of users. By providing explainable recommendations, this helps guard against these biases and help users make decisions that allow them to use recommendations as aids instead of traps. Simultaneously, certain recommendations may generate a self-reinforcing pattern in which the recommended item will continue to be recommended if it was successfully identified as a good suggestion amongst other users, fostering a feedback loop rooted in algorithmic bias.

The loss of autonomy in users of these RS can be seen through the traps, feedback cycles, manipulations, and biases. With billions of users interacting with these systems every day, the system is created to improve the online decision-making process, but is manipulative in its biased nature. The autonomy of users is infringed upon by simply being exposed to these systems, as every day decisions in buying products, listening to music, watching movies, and gaining information/news is all affected. The comparison of the algorithmic profile to a real-life profile was made earlier, and by showing the bias present in each layer of the algorithmic profile, there is no possibility that an algorithmic profile can align perfectly with our real-life social profiles. Therefore, the suggestions made by these algorithms are not fully representative of the decisions made by humans, but give us the best algorithmic glimpse of who we are.

Infringement on User Autonomy

The RS used by Instagram promises to show users a feed that would be “ordered to show the moment we believe you will care about the most”. The RS they implement involves a type of algorithmic ranking in which the feed you see is personalized to you, but the ways in which your feed is ranked is filled with bias at each level of data collection, algorithmic, and outcome, similarly to how we used the framework for other RS examples. There are many users that utilize the feed ranking to soar to the top of the Instagram visibility leaderboard by “beating the algorithm”, but this creates an impact on other users as the more likes they get, the more their profile will gain rank. The more rank they gain in the algorithm, the more likely it is for them to show up in Instagram’s promised “moments you will care about most” feed.

Facebook’s news feed ranking algorithm is a particular RS that has gained lots of attention for the relevancy of news feed articles, by becoming a news source in which users are presented information in which bias algorithms control what a user sees (Cotter, 2018). The mentioned manipulations of RS can also be found in these social news sites, which influences the choice that users have on which stories appear on their news feed. The social impact of these manipulations can continuously create online traps for users to have no choice but to cede to the algorithm’s preference of “most relevant feed”. There exist many times when the algorithmic suggestion and user preference align, but I will focus on the negative externalities as a whole, since most users are unaware of possible biases.

These possible types of manipulative attacks on RS present bias at each stage during the recommendation process. Due to all of the different biases, there is no doubt that algorithmic bias exists in RS. This becomes an ethical challenge as each recommendation made only furthers the bias to ultimately infringe on the choices online users have. Unknowingly, users may be presented with bias suggestions, in which I pose explainable recommendations as a solution.

Explainable Recommendations to limit the Social Impact

One of the key aspects of this research is to evaluate the current existing implementations of recommender systems in order to gain a knowledge of the black-box nature (Bottando, 2012) of these systems. An explainable recommendation is one where a recommendation is not given simply from an input, but rather through provided explanations of the system and how it came to this conclusion. This is quickly gaining more attention as the user base for online services grow, but the recommender systems are still black-boxed due to privacy issues.

Explainable recommendations offer explanations of why items are recommended and bridges the relationship between RS and users. Good explanations can increase trust in the RS and eliminating some of the biases present in the systems (Wang et al, 2018). Explanations can also serve to restore the user autonomy of RS and prevent manipulation by giving a type of transparency to users of the system. These explainable recommendations are becoming a hot topic in the field of Information Retrieval as sentence-forming natural language processing and neural networks are being used to create explanations for a variety of recommendations.

The idea of fairness in machine learning models that make predictions affecting decision making is crucial to the growth of RS. An explainable recommendation leads to a more transparent RS, and both of those lead to improving fairness (Abdollahi, 2018). By providing explanations to improve algorithmic fairness, I argue that this solution will help to eliminate algorithmic bias in RS. By eliminating as much algorithmic bias as possible, we are restoring trust in the user base of RS and increasing the efficacy of RS.

Since we have already proven the bias in the RS algorithms, some researchers suggest that explainable recommendations can also improve the troubleshooting of RS as well as the future for them since they can easily be modified once a bias is found. An RS should explain their predictions in such a way that users will be able to understand how the system came up with the prediction. For example, an RS showing products because “10 of your friends also bought this product” is not sufficient, there must be an explanation for how the model came up with the prediction even before “10 friends also bought this product”. The way to achieve this would be to create full transparency with these black-box algorithms, but since many privacy laws protect this type of transparency, there is a need for a method to restore trust in RS consumers.

Discussion

The current works and research being done in the field of explainable recommendations is a start to a more transparent future for recommender systems. The biases at each observed level of the RS: data bias, algorithm bias, outcome bias, along with possible manipulation of RS leads to an infringement of choice for online consumers. Consumers that are not made aware of these biases and manipulations may find themselves victim to restricted online choices, content-bubbles, and part of algorithmic profiling.

Explainable recommendations can improve data bias through the transparency of dataset values. The explanations for this would involve listing which attributes of our algorithmic profile had the largest affects in the predictions made by RS. This would in turn eliminate some of the algorithmic biases since this will show what the algorithm also thinks is important for a relevant recommendation, and when a recommendation is not relevant, the programmers will be able to quickly identify and troubleshoot the algorithmic biases. In the end, outcome biases of the recommendation will also decrease since users are aware of possible manipulations due to the explainable recommendations. By improving every level of the framework for identifying RS bias, I conclude that explainable recommendations are a solution to increase trust of users and improves user autonomy within the online decision-making process by bridging the gap between users and the recommendation algorithms.

References

Abdollahi, B., & Nasraoui, O. (2018). Transparency in fair machine learning: The case of explainable recommender systems. In *Human and Machine Learning* (pp. 21–35). Springer International Publishing.https://doi.org/10.1007/978-3-319-90403-0_2

Baeza-Yates, R. (2016, May 22). Data and algorithmic bias in the web. Proceedings of the 8th ACM Conference on Web Science. WebSci ’16: ACM Web Science Conference.https://doi.org/10.1145/2908131.2908135

Baltrunas, L., Kaminskas, M., Ludwig, B., Moling, O., Ricci, F., Aydin, A., Lüke, K.-H., & Schwaiger, R. (2011). InCarMusic: Context-Aware Music Recommendations in a Car. In C. Huemer & T. Setzer (Eds.), *E-Commerce and Web Technologies* (pp. 89–100). Springer Berlin Heidelberg.

Beer, D. (2009). Power through the algorithm? Participatory web cultures and the technological unconscious. *New Media & Society*, *11*(6), 985–1002.https://doi.org/10.1177/1461444809336551

Burr, C., Cristianini, N., & Ladyman, J. (2018). An analysis of the interaction between intelligent software agents and human users. *Minds and Machines*, *28*(4), 735–774.https://doi.org/10.1007/s11023-018-9479-0

Cramer, H., Garcia, H., Springer, A., & Reddy, S. (2018). Assessing And Addressing Algorithmic Bias in Practice. https://doi.org/10.1145/3278156

Cotter, K. (2019). Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media & Society*, *21*(4), 895–913.https://doi.org/10.1177/1461444818815684

Friedman, B. & Nissenbaum, H. (1996). Bias in computer systems. *ACM Trans. Inf. Syst. 14*, 3 , 330--347

Hallinan, B., & Striphas, T. (2016). Recommended for you: The Netflix Prize and the production of algorithmic culture. New Media & Society, 18(1), 117–137.https://doi.org/10.1177/1461444814538646

Hargittai, E. (2007). The Social, Political, Economic, and Cultural Dimensions of Search Engines: An Introduction. *Journal of Computer-Mediated Communication*, *12*(3), 769–777. https://doi.org/10.1111/j.1083-6101.2007.00349.x

Maguire, J., & Matthews, J. (2012). Are we all Cultural Intermediaries Now? *European Journal of Cultural Studies*. https://doi.org/10.1177/1367549412445762

Morris, J. W. (2015). Curation by code: Infomediaries and the data mining of taste. *European Journal of Cultural Studies*, *18*(4–5), 446–463.https://doi.org/10.1177/1367549415577387

Milano, S., Taddeo, M. & Floridi, L. Recommender systems and their ethical challenges. *AI & Soc* **35,** 957–967 (2020). https://doi.org/10.1007/s00146-020-00950-y

Mann, M., & Matzner, T. (2019). Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. *Big Data & Society*. https://doi.org/10.1177/2053951719895805

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, *2*, 13. https://doi.org/10.3389/fdata.2019.00013