Under the hood of the ethical recommender system: ensuring fairness and bias reduction

Hadrien Sevel

December 2022

Abstract

Recommender systems, which use machine learning algorithms to personalize recommendations for products, services, or content, have the potential to enhance the user experience significantly. However, these systems can also perpetuate and amplify existing biases if they are not designed and implemented carefully. This can lead to the creation of "filter bubbles" that isolate individuals by only showing them content or products that align with their existing beliefs or preferences, leading to a lack of exposure to diverse perspectives and reinforcing existing biases. In this paper, I discuss how bias can manifest in recommender systems and explore strategies for ensuring that these systems are fair and do not reinforce existing inequalities. I propose several best practices for designing and implementing ethical recommender systems, including using diverse training data, adopting fairness metrics and breaking out filter bubbles by showing users content or products that may challenge their existing beliefs or by promoting a more comprehensive range of options. With this set of general guidelines, I aim to create a more ethical system that is fair and does not disadvantage or discriminate against specific individuals or groups.

Contents

1 Introduction

If you have ever surfed the web, you have most likely encountered them without even knowing it: the recommender systems. Online shopping, streaming platforms, and social networks, among others, use these systems to enhance the user's online experience by providing them with relevant content recommendations that best suit their needs, often with machine learning algorithms. Making their appearance in the 1990s, they have been abundantly used by companies offering their services on the web since the 2000s, thanks to the evolution of the Internet and the availability of many data that have made possible the development of very sophisticated systems.

Although these systems can significantly enhance the user's experience while surfing the web, they raise many ethical issues that can negatively impact the user. We can mention for example the lack of transparency (it is impossible for the user to know how the algorithm made the decision to give him a certain recommendation), the exposure of the user's personal data to privacy breaches, or the collection of data without their consent (Milano et al., 2020, pp. 960–964), the creation of "filter bubbles" (lack of exposure to information that challenges existing beliefs) (Bozdag, 2013, p. 218) or the perpetuation of bias from the data used by the algorithm (Müller, 2021, p. 8). These problems become increasingly important as these systems become more efficient. There is, therefore, a need to define a more ethical recommender system. But how can we design and ensure that such a system is fair and does not perpetuate or amplify existing biases? In this paper, I will focus on the concepts of bias and unfairness, and the consequence of "filter bubbles".

To address this question, I will first give a definition and a brief technical overview of recommender systems. Second, I will show how these systems can lead to the specific ethical issues of bias and unfairness and address the concrete case of "filter bubbles". Finally, I will discuss some specific design guidelines that can help eliminate them.

2 Technical overview

The definition of recommender systems is broad and encompasses many machine-learning-based systems of today. A definition given by Burke in the early 2000s is "any system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options" (Burke, 2002, p. 331). From this, two fundamental principles can be extracted (Burke et al., 2011, p. 14): first, a recommender system is designed to produce an output specific to the user in order to optimize its experience, and second, it aims to help the user choose among a set of discrete options. Burke et al. (2011) also emphasize the difference between a search engine (that also gives the user a selection of options) and a recommender system. The key aspect is personalization: a search engine will provide the same results for the same query and is, therefore, not dependent on the user. However, many modern search engines like Google now implement personalization to tailor the search results to the user by filtering irrelevant information (Bozdag, 2013, p. 211). Thus, this difference becomes less and less clear as recommender systems are implemented nowadays in many applications and become just components of more extensive systems.

Two main approaches are used to provide user recommendations: the collaborative and the content approaches. In the following paragraphs, I give a brief technical overview of each of them because it is essential to know how they work in order to understand where specific ethical questions may arise.

2.1 Collaborative approach

The collaborative approach uses users' past interactions with items (a search result, an article, a post on a social network, an ad, etc.) to provide new recommendations (Rocca & Rocca, 2019). The recommender systems using this approach aggregate engagements of a user with items (for instance, clicks, views, ratings, time spent viewing the item, etc.) and can thus create a profile of this user that can be compared to other profiles (Burke, 2002, pp. 332–333) to find similarities and create recommendations based of these. Intuitively, two users having very similar profiles are likely to have common interests, therefore an item that user 1 engaged positively with could be recommended to user 2.

This approach is straightforward as the profiles can be embedded in an interactions matrix, with the profiles as rows and the items as columns. Finding the missing values of the sparse matrix then enables the production of recommendations (Burke et al., 2011, p. 16). Multiple machine-learning methods can be used to solve this problem.

Such a recommender system doesn't need much data (only user interactions) and can be very simple and effective in producing recommendations. However, it requires many interactions to be accurate and thus doesn't work well with new users or items. This problem is known as *cold start.*

2.2 Content approach

While collaborative systems are agnostic of the content of the items to be recommended, content-based recommender systems are built upon the combination of user personal information and item features (Rocca & Rocca, 2019). The features of an item can be constructed in several ways, depending on the item type. For instance, the features of a movie might be information about the actors and the story, while the features of a song could be the genre, the authors, and the year. In realistic scenarios, the features will come from complex embedding mechanisms that rely on advanced machine-learning techniques. However, the accuracy of the recommendations depends heavily on the quality of the features (Burke et al., 2011, p. 16), this approach is thus less straightforward that the collaborative one.

Once a history of user preferences is collected, content-based recommendations can be done by suggesting items with features similar to those in the user's preferences. More complex models can also be used in this framework, which might be able to recommend specific items based on users' characteristics (i.e. the age of the user might be a good predictor for a music genre).

The content approach is more complex than the collaborative one but has advantages: it takes into account a lot more data to produce more accurate recommendations and doesn't suffer much from cold starts.

However, nowadays, two or more approaches are often combined to form *hybrid recommender systems* (Burke, 2002, p. 339) to improve the recommendations' performances and accuracy while reducing the drawbacks of the individual approaches.

3 Bias and unfairness in recommender systems

In the following paragraphs, I will first define the ethical concepts involved and then show how today's recommender systems raise ethical issues concerning bias and unfairness. I will finally address the immediate consequence of "filter bubbles".

3.1 Definitions

We first have to define the concept of ethics. In this paper, I treat ethics as normative ethics, in the sense that it is not a description of norms (which is descriptive ethics). According to Bartneck, normative ethics can be defined as "the analysis of human actions from the perspective of 'good' and 'evil,' or of 'morally correct' and 'morally wrong.' " (Bartneck et al., 2021, p. 19) We can deduce from this definition that ethics intervenes when there is a "human action" involved, i.e. when an individual has to make a decision.

This is not without consequences for engineers making recommender systems. Indeed, they have to make numerous decisions during the design process of the systems, which raises normative questions: the engineers should be able to justify choosing one option over others; this is the crucial part of the ethical question. This justification is closely linked to the values that engineers embody. (Rochel & Evéquoz, 2021, p. 612).

I define bias as systematic prejudiced results and fairness as the absence of discrimination in the results produced by recommender systems. As the recommendations have to be personalized for each user since it is the goal of the systems, there exists a difference in treatment between users. However, I will consider in this paper discrimination not as a difference in treatment between individuals, but rather as a "systematic disadvantage on one a social group relative to others" (Barocas et al., 2019, Chapter 4).

3.2 Bias and unfairness in modern recommender systems

One of the main ethical concerns surrounding recommender systems is the potential for bias and unfairness. Bias can manifest in various ways in these systems, including using biased data, the selection of certain algorithms or metrics that may favor certain groups over others, and the lack of accountability and transparency in the decision-making process.

One common source of bias in recommender systems is the data used to train the algorithms (Gebru, 2020, p. 256). If the data is not representative of the population or if it contains inherent biases, such as stereotypes or prejudices, the algorithms will also reflect these biases. This can lead to unequal treatment of different groups, as the recommendations may be based on inaccurate or biased assumptions about their preferences or needs. Moreover, training algorithms on biased data create runaway feedback loops: biased results based on biased data will further amplify the bias by creating new biased data. This phenomenon happened with an automatic hiring tool at Amazon (giving recommendations on who to hire) that discriminated against women because it was trained on biased data and was therefore unfair.

Another source of bias is the selection of certain algorithms or metrics that may favor certain groups over others (Coeckelbergh, 2020, p. 128). For example, the content approach for recommender systems may prioritize certain types of content over others, or certain metrics may give more weight to certain interactions of the users with items, leading to unequal treatment of different groups and thus to unfairness.

Additionally, the lack of transparency and accountability in the decision-making process of recommender systems can also contribute to bias (Milano et al., 2020, pp. 962–963). It is difficult to identify and address potential biases without understanding how the algorithms make their recommendations. Moreover, the lack of diversity in teams (Gebru, 2020, pp. 264–267) is also a factor amplifying unfairness

3.3 Filter bubbles

The consequences of these biases can be severe, as they can reinforce and amplify existing inequalities and discrimination. This can lead to "filter bubbles" (Milano et al., 2020, p 964), where individuals are only shown content or products that align with their existing beliefs or preferences, leading to a lack of exposure to diverse perspectives and reinforcing existing biases. This greatly impacts society and democracy (Bozdag, 2013, p. 254) because it restricts individual liberty and increases unawareness.

In addition, "filter bubbles" can contribute to the polarization of society (Stray, 2021, p. 2), as individuals are only exposed to information that supports their existing beliefs and may not be exposed to alternative viewpoints. This can lead to a lack of understanding and empathy towards other groups and can lead to further social divisions. The impact of these "filter bubbles" can be especially harmful for marginalized groups, who may already have limited access to diverse perspectives and may be further isolated by these systems. Therefore, it is crucial that steps are taken to ensure that recommender systems do not reinforce existing inequalities and biases, but rather promote a more inclusive and diverse society.

4 Designing a more ethical recommender system

We saw in the paragraphs above how recommender systems can be biased and unfair, potentially leading to the creation of filter bubbles that have a negative impact on the user, and more broadly on democracy and society. In this section, I give some guidelines that can be followed to design more ethical recommender systems.

Team diversity

For Gebru, the first step towards a less biased system is the diversity in the team. "Ethical AI is not an abstract concept but one that is in dire need of a holistic approach. It starts from who is at the table, who is creating the technology, and who is framing the goals and values of AI." (Gebru, 2020, p. 264). As I wrote above, much of the justification for engineering decision-making comes from the values they embody. Thus, a diverse team implies a broader spectrum of values that can be implemented in the recommender system. According to the IEEE, values should be translated into norms (obligations and prohibitions) to be embedded into the AI systems (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019, p. 169). However, these norms may differ in different communities and nations, making it challenging, if not impossible, to define a set of universal norms; thus, a diverse team with different backgrounds and cultures may help choose a set of norms carefully.

Training data

We saw that the data plays a crucial role in the bias of the recommender systems: biased data is going to amplify the bias of the recommendations. Therefore, the engineers have to select it carefully. But in order to do that, they need to have a deep understanding of the objectives of the AI (Rochel & Evéquoz, 2021, p. 614). Indeed, this represents a big responsibility in the design of the recommender system and can significantly influence the achievement of the objectives. The preparation of the data also plays a part, as it is important to clean and preprocess the data to ensure that it is representative and accurate. This can involve removing any irrelevant or redundant data, correcting any errors or inconsistencies, and standardizing the data format. Additionally, removing any sensitive or personal information from the data may be necessary to protect users' privacy and prevent it from leaking through data breaches (Milano et al., 2020, p. 961).

Fairness metrics

Fairness metrics can be used to measure and assess the fairness of recommendation systems. These metrics can help identify any potential biases or inequalities in the system and allow engineers to take corrective action to address them.

Several different fairness metrics can be used, depending on the specific goals and objectives of the recommendation system (Barocas et al., 2019). Some common fairness metrics include:

- Equal opportunity: This metric measures whether the system provides equal opportunities to different groups, regardless of their characteristics. For example, it could be used to ensure that a recommendation system for job openings does not discriminate against certain groups based on factors such as race or gender.
- Equal odds: This metric measures whether the system provides equal outcomes for different groups, regardless of their characteristics. For example, it could be used to ensure that a recommendation system for credit applications does not discriminate against certain groups based on factors such as income or credit history.
- Equal treatment: This metric measures whether the system treats different groups equally, regardless of their characteristics. For example, it could be used to ensure that a recommendation system for healthcare services does not discriminate against certain groups based on factors such as age or disability.

Breaking out filter bubbles

The big issue with filter bubbles is the polarization of the creation of echo chambers since the recommender system only recommends items that the user is likely to agree with. One way to break out filter bubbles is to show users content or products that may challenge their beliefs or expose them to new perspectives. This can be done by introducing various options or promoting content or products outside a user's specific interests or preferences (Bozdag & Hoven, 2015). For example, a recommendation system for a streaming platform could suggest documentaries or movies that explore different cultures or viewpoints rather than just showing users content that aligns with their current interests. This can help users learn and grow by exposing them to new and diverse perspectives.

Another way to break out filter bubbles is to design recommendation systems to promote a more comprehensive range of options. Instead of just showing users content or products that align with their existing beliefs or preferences, these systems can be designed to present a more balanced and diverse set of options. This can help prevent users from becoming isolated in their own echo chambers and promote more diverse and inclusive exposure to different viewpoints and ideas.

5 Conclusion

In conclusion, recommender systems have the potential to greatly enhance the user experience by providing personalized recommendations for products, services, or content. They are now a preponderant part of the web as they are becoming increasingly sophisticated and implemented in many more extensive AI systems, like search engines or social networks. However, these systems can also perpetuate and amplify existing biases as well as unfairness if they are not designed and implemented carefully, for instance through biased data or lack of diversity in the teams creating them. This can lead to the problematic creation of "filter bubbles" that isolate individuals by only showing them content or products that align with their existing beliefs or preferences, leading to a lack of exposure to diverse perspectives and reinforcing existing biases. This is an issue for the users, as well as for society and democracy in general.

Several guidelines can be used to design more ethical recommender systems, and we saw in this paper the importance of having diverse teams of engineers formed of people with diverse backgrounds and cultures, using diverse training data that is correctly selected and prepared, adopting fairness metrics, incorporating transparency and accountability measures, and breaking out filter bubbles by showing users content or products that may challenge their existing beliefs or by promoting a more comprehensive range of options. By following these best practices, it is possible to create recommender systems that tend to be fairer and do not disadvantage or discriminate against specific individuals or groups.

References

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. fairmlbook.org.

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). What is ethics? In *An introduction to ethics in robotics and AI* (pp. 17–26). Springer International Publishing. https://doi.org/10.1007/978-3-030-51110-4_3

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, *15*(3), 209–227. https://doi.org/10.1007/s10676-013-9321-6

Bozdag, E., & Hoven, J. van den. (2015). Breaking the filter bubble: Democracy and design. *Ethics and Information Technology*, *17*(4), 249–265. https://doi.org/10.1007/s10676-015-9380-y

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, *12*(4), 331–370. https://doi.org/10.1023/A:1021240730564

Burke, R., Felfernig, A., & Göker, M. H. (2011). Recommender systems: An overview. *AI Magazine*, *32*(3), 13–18. https://doi.org/10.1609/aimag.v32i3.2361

Coeckelbergh, M. (2020). *AI Ethics*. The MIT Press. https://doi.org/10.7551/mitpress/12549.001.0001

Gebru, T. (2020). Race and Gender. In *The Oxford Handbook of Ethics of AI*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.16

Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & SOCIETY*, *35*(4), 957–967. https://doi.org/10.1007/s00146-020-00950-y

Müller, V. C. (2021). Ethics of Artificial Intelligence and Robotics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2021). https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/; Metaphysics Research Lab, Stanford University.

Rocca, B., & Rocca, J. (2019). *Introduction to recommender systems*. Towards Data Science. https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada

Rochel, J., & Evéquoz, F. (2021). Getting into the engine room: A blueprint to investigate the shadowy steps of AI ethics. *AI & SOCIETY*, *36*(2), 609–622. https://doi.org/10.1007/s00146-020-01069-w

Stray, J. (2021). *Designing recommender systems to depolarize*. arXiv. https://doi.org/10.48550/ARXIV.2107.04953

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (IEEE, Ed.; 1st ed.). https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html