Chapter 2
Transparency in Fair Machine Learning: the Case of Explainable Recommender Systems

Behnoush Abdollahi and Olfa Nasraoui

Abstract Machine Learning (ML) models are increasingly being used in many sectors, ranging from health and education to justice and criminal investigation. Therefore, building a fair and transparent model which conveys the reasoning behind its predictions is of great importance. This chapter discusses the role of explanation mechanisms in building fair machine learning models and explainable ML technique. We focus on the special case of recommender systems because they are a prominent example of a ML model that interacts directly with humans. This is in contrast to many other traditional decision making systems that interact with experts (e.g. in the health-care domain). In addition, we discuss the main sources of bias that can lead to biased and unfair models. We then review the taxonomy of explanation styles for recommender systems and review models that can provide explanations for their recommendations. We conclude by reviewing evaluation metrics for assessing the power of explainability in recommender systems.

2.1 Fair Machine Learning and Transparency

2.1.1 Fairness and Explainability

ML models make predictions that affect decision making. These decisions can have an impact on humans, either individually (for a single person) or collectively for a group of people. Such an impact can be unfair if it is based on an inference that is

B. Abdollahi · O. Nasraoui (✉)
Knowledge Discovery and Web Mining Lab, Computer Engineering and Computer Science Department, University of Louisville, KY, USA
e-mail: olfa.nasraoui@louisville.edu

B. Abdollahi
e-mail: b.abdollahi@louisville.edu

J. Zhou and F. Chen (eds.), *Human and Machine Learning*, Human–Computer Interaction Series, https://doi.org/10.1007/978-3-319-90403-0_2

biased against a certain group of people. Hence fairness is an important criterion in ML. Fairness in ML is a nascent topic that has only recently attracted attention [19, 34]. How to achieve this fairness is therefore still a matter of debate and there have recently been only a few attempts to define fairness and design fair algorithms within the ML context [18, 19, 26]. In our view, fairness can be achieved in multiple ways and either completely or partially. In particular, fairness can be addressed by changing the data that models ingest, the ways (i.e. algorithms) that models are learned, or the predictions that are made by these models. Another way that fairness can be achieved is by completely transparent models and thus scrutable predictions; in other words, predictions that can be justified as to the reasons why a particular prediction has been made and scrutinized for potential biases or mistakes. This is because such a scrutiny provides a certain level of accountability. For this reason, we believe that explainability can play an important role in achieving fairness in ML. Figure 2.1 presents a diagram that shows the relation between explainability, transparency and fairness. Figures 2.2 and 2.3 show two forms of designing explainable ML systems. In Fig. 2.2, the predictions are explained to the user using a model that is different from the ML model, while in Fig. 2.3, explainability is incorporated at the design level within the ML model.

2.1.2 Fair Machine Learning

ML models are increasingly being used in many sectors ranging from health and education to justice and criminal investigation. Hence, they are starting to affect the lives of more and more human beings. Examples include risk modeling and decision making in insurance, education (admission and success prediction), credit scoring, health-care, criminal investigation and predicting recidivism, etc [19, 54]. These models are susceptible to bias that stems from the data itself (attribute or labels

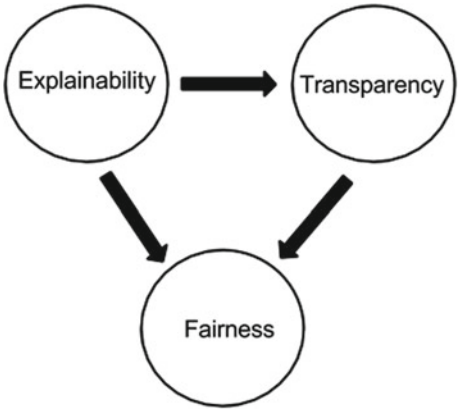

Fig. 2.1 Explainability leads to transparency and both lead to improving fairness of ML models

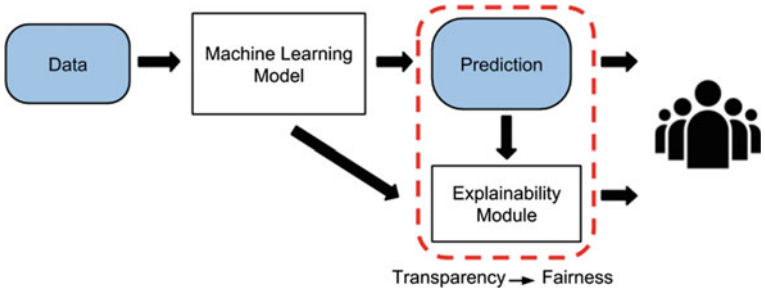

Fig. 2.2 In this form of fair ML, explainability occurs at the prediction step which results in more transparency and increasing fairness by presenting justified results to the user

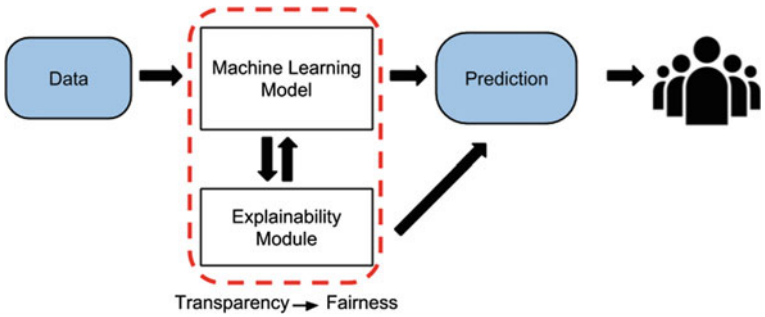

Fig. 2.3 In this form of fair ML, explainability occurs in the modeling phase which results in designing transparent ML models and consequently having more transparent and fair models

are biased) or from systemic social biases that generated the data (e.g. recidivism, arrests). As such, models that are learned from real world data can become unethical. Data can be collected and labeled in a biased way such that it is discriminative against a certain race, gender, ethnicity or age. As bias in the data can result in unfair models, ML algorithms are also susceptible to strategic manipulation [6, 24]. Therefore, they can be built such that the model creates bias against a certain group of people. The involvement of the human in all the stages of collecting data, building a model, and reporting the results, creates the setting for various types of bias to affect the process. Some of the sources of human bias in the stages of collecting and processing the data and reporting the results are [39]:

- Confirmation bias: It is a tendency to intentionally search for and include certain data and perform analysis in such a way as to make a predefined conclusion and prove a predetermined assumption.
- Selection bias: This happens when the sample is not collected randomly and because of a subjective selection technique, the data does not represent the whole population under study. Based on the elimination of samples or inclusion of certain samples, the resulting bias can be of the omission or inclusion type, respectively.

- Implicit bias: This type of bias is associated with the unconscious tendency to favor a certain group of people against others based on characteristics such as race, gender, and age.
- Over-generalization bias: This form of bias can come from making a certain conclusion based on information that is too general, especially when the sample size is small or is not specific enough.
- Automation bias: The tendency to favor decisions made from an automated system over the contradictory correct decision made without the automation.
- Reporting bias: This form of bias is the result of an error made in the reporting of the result when a certain positive finding is favored over the negative results.

Recent studies proposed techniques for building fair models by alleviating the effect of bias. Avoiding the use of sensitive features has been shown to be insufficient for eliminating bias, because of the correlation between some features which can indirectly lead to oriented and unfair data [32]. Kamishima et al. [33] formulated causes of unfairness in ML and presented a technique based on regularization by penalizing the classifier for discrimination and building discriminative probabilistic models to control the bias that resulted from prejudice. Since their solution as the prejudice remover is formulated as a regularizer, it can be used in a variety of probabilistic models such as the logistic regression classifier. Fish et al. [20] proposed a method based on shifting the decision boundary in the learning algorithm for achieving fairness and providing a trade-off between bias and accuracy.

In addition to designing fair algorithms, [32] proposed an approach for removing bias and generating fair predictions by changing the data before training the model. This method is based on modifying the dataset in order to transform the biased data into an unbiased one. The authors used a ranking function learned on the biased data to predict the class label without considering the sensitive attribute. Using this technique, they estimate the probability of the objects belonging to the target class. Their results showed that they could reduce the discrimination by changing the labels between the positive and negative class.

2.2 Explainable Machine Learning

Conventional evaluation metrics such as accuracy or precision do not account for the fairness of the model. Thus, to satisfy fairness, explainable models are required [36]. While building ethical and fair models is the ultimate goal, transparency is the minimum criterion that ML experts can directly contribute to and this could be the first step in this direction. Therefore, designing explainable intelligent systems that facilitate conveying the reasoning behind the results is of great importance in designing fair models. Note that we do not make a distinction between “explainability” and “interpretability” and use both terms interchangeably.

In the context of machine learning, interpretability means "explaining or presenting in understandable terms" [4]. In addition, interpretability and explanations can help to determine if qualities such as fairness, privacy, causality, usability and trust are met [18]. Doshi-Velez and Kim [18] presented a taxonomy of approaches for the evaluation of interpretability in ML models in general: application-grounded, human-grounded, and functionality-grounded. Application-grounded and human-grounded evaluation approaches are both user-based, while the functionality-grounded approach does not require human evaluation and uses some definition of the explainability for the evaluation. Experiments can be designed based on different factors, such as global versus local, which considers the general patterns existing in the model as global, while considering local reasoning behind the specific prediction of the model as local [18]. The global pattern is usually helpful for the designer and developer of the model when understanding or detecting bias or causality in the model. The local pattern, on the other hand, can be aimed at the end user of the systems to understand the justifications of the system decisions.

Explainability-aware ML techniques can generally be categorized into two main groups:

1. Models that explain their predictions in a way that is interpretable by the user. These types of methods usually only justify their output without providing an in depth understanding of the ML algorithm. This form of explanation is usually helpful when the user of the system is not an expert such as in the case of recommender systems. The explanation generation module can be located in a separate module relative to the predictor.
2. Models that incorporate interpretable models in the building of the automated systems. White-box models, such as decision trees where the ML model is intuitive for the user, can be categorized in this group, although, in these models the model is usually kept simple and in many cases they might not be as accurate as the more powerful black-box techniques.

Ribeiro et al. [42] proposed an explanation technique that explains the prediction of the classifiers locally, using a secondary learned white box model. Their proposed explanation conveys the relationship between the features (such as words in texts or parts in images) and the predictions; and helps in feature engineering to improve the generalization of the classifier. This can help in evaluating the model to be trusted in real world situations, in addition to using the offline accuracy evaluation metrics. Freitas [21] reviewed comprehensibility or interpretability of five classification models (decision trees, decision tables, classification rules, nearest neighbors, and Bayesian network classifiers). It is important to distinguish understanding or interpreting an entire model (which the paper does) from explaining a single prediction (which is the focus of this chapter). In addition, we note that Freitas overviews the problem from several perspectives and discusses the motivations for comprehensible classier models, which are:

1. Trusting the model: Regardless of accuracy, users are more prone to trusting a model if they can comprehend why it made the predictions that it did.
2. Legal requirements, in some cases like risk modeling, where a justification is required in case of denying credit to an applicant.
3. In certain scientific domains such as bioinformatics, new insights can be obtained from understanding the model, and can lead to new hypothesis formation and discoveries.
4. In some cases, a better understanding can help detect learned patterns in the classification model that are not really stable and inherent in the domain, but rather result from overfitting to the training data, thus they help detect the data shift problem: i.e., when the new instances deviate in their distribution from past training data; we note that concept drift (when a previously learned and accurate model no longer fits the new data because of changes in patterns of the data) can be considered as a special case of the data shift.

Understanding the logic behind the model and predictions (in other words, comprehension) can reveal to the user the fact that the (new) data has outpaced the model. The user can then realize that the model has become old and needs to be updated with a new round of learning on new data. Various interpretation methods exist depending on the family of classier models: decision trees, rule sets, decision tables, and nearest neighbors. Different studies have shown that the interpretability of entire classier models depends on the application domain and the data, with findings that sometimes contradict each other. Regardless of all the findings in interpreting models, we note that the task of interpreting an "entire classifier model" (e.g. a complete decision tree or a set of 500 rules) is different from that of one user trying to understand the rationale behind a "single prediction/recommendation" instance. That said, we find Freitas' review to be very important for transparency, fairness and explainability: first, he argues that model size alone is not sufficient to measure model interpretability, as some models' complexity is beyond mere size and small models can actually hurt the user's trust in the system (a notorious example is decision stump models [1-level trees]). Also, extremely small models would likely suffer in accuracy. Second, the work on interpreting rule-based models and nearest neighbor models can be useful to us because it is closest to the Collaborative Filtering (CF) recommendation mechanisms we study. For nearest neighbor models, Freitas [21] mentions that attribute values of nearest neighbors can help provide explanations for predictions, and that showing these values in decreasing order of relevance (based on an attribute weighting mechanism) is a sensible strategy. Another strategy is to show the nearest prototypes of training instances, for example after clustering the training instances. However, in both of these strategies, Freitas [21] was motivating interpretations of entire models rather than individual prediction explanations in the context of recommending items.

2.3 Explainability in Recommender Systems

2.3.1 Transparency and Fairness in Recommender Systems

Dascalu et al. [15] presented a survey of educational recommender systems and Thai-Nghe et al. [54] presented a recommender system for predicting student performance. Because the data in the education setting can be biased due to the underrepresentation of women in Science, Technology, Engineering, and Mathematics (STEM) topics [7, 23, 49], the predictive models resulted in an unfair system when evaluated using fairness metrics [58]. This form of bias can be dominant when the demographic profile of the user, consisting of features such as gender, race and age, is used in the model. To avoid unfairness or bias in the recommendations, the influence of specific information should be excluded from the prediction process of recommendation and for this reason CF techniques can be preferable to content-based recommender systems. While using CF models with only rating data can eliminate this bias, rating data can include another form of bias. For example in the MovieLens data [27], the ratings are obtained from the users who have rated a sufficient number of movies and the data is inherently biased towards the "successful users" [27]. This shows the serious problem of unfairness that can happen in a recommender model due to the bias in the data. This setting provides a motivation for designing transparent models and generating explainable recommendations. Sapiezynski et al. [43] studied the fairness of recommender systems used for predicting the academic performance of students. They showed that because of the gender imbalance in many data sets, the accuracy for female students was lower than male students and a different selection of features can result in a fair model.

2.3.2 Taxonomy of Explanations Styles

Recommender systems are a prominent example of a ML model that interacts directly with humans (users). This is in contrast to for instance, traditional medical decision making systems that interact with health-care providers/experts. Explanations have been shown to increase the user's trust in a recommender system in addition to providing other benefits such as scrutability, meaning the ability to verify the validity of recommendations [29]. This gap between accuracy and transparency or explainability has generated an interest in automated explanation generation methods. Explanations can be given using words related to item features or user demographic data, but these cannot be done easily in CF approaches. They vary from simple explanation formats such as: "people also viewed" in e-commerce websites [55] to the more recent social relationships and social tag based explanations [44, 57]. Bilgic and Mooney [8] showed how explaining recommendations can improve the user's estimation of the item's quality and help users make more accurate decisions (i.e. user satisfaction). Based on [8], three different approaches to explanations can be delineated:

1. Neighbor Style Explanation (NSE): this explanation format compiles a chart in CF that shows the active user's nearest CF neighbors' ratings on the recommended item. A histogram of these ratings among the nearest neighbors can be presented to the user. Figure 2.4 (1) and (3) show two different formats of the neighbor style explanation.
2. Influence Style Explanation (ISE): this explanation format presents a table of those items that had the most impact on computing the current recommendation. They can be used in both CBF and CF. An example is shown in Fig. 2.4 (2).
3. Keyword Style Explanation (KSE): this explanation format analyzes the content of recommended items and the user's profile (interests) to find matching words in CBF. An example of the KSE format which is obtained from the MovieLens benchmark dataset is shown in Fig. 2.4 (4). Figure 2.4 (3), shows an example of a neighbor style explanation (NSE) for a recommended movie based on the user's neighbors. This user-based example presents the ratings distribution of the user's neighbors on three rating levels.

Giving the user information about what type of data is used in the system encourages the user to provide more helpful data of that kind, such as preference ratings. Information about the neighbors selected as the predictors could give the user a chance to examine their ratings and to disregard the recommendations if the right neighborhood is not selected. A good explanation could also help discover weak predictions. The distribution of the ratings of the neighbors on a target item is helpful in identifying whether the prediction is based on enough data or not. Herlocker et al. [29] compared 20 other explanation systems and found histograms to perform best based on promotion only. Abdollahi and Nasraoui [3] presented an Explain-

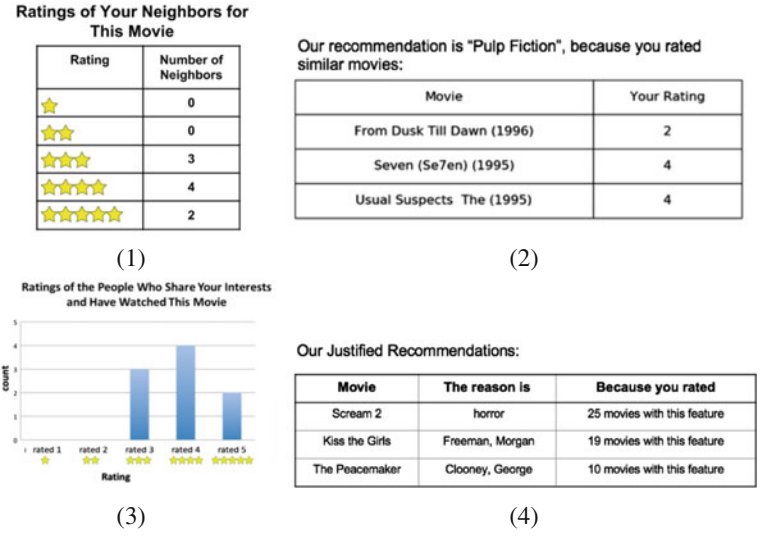

Fig. 2.4 Four different explanation style formats: (1) NSE, (2) ISE, (3) NSE, (4) KSE

able Matrix Factorization (EMF) technique that proposed a metric for evaluating the explainability of the NSE and ISE style explanations and proposed to precompute explanations in a graph format and then incorporate them in a matrix factorization-based recommender system. NSE can be formulated based on the empirical density distribution of the similar users' ratings on a recommended item. Therefore, for user u, given the set of similar users as N_u, the conditional probability of item i having rating k can be written as:

$$\mathbf{P}(r_{u,i} = k|N_u) = \frac{|N_u \cap U_{i,k}|}{|N_u|} \tag{2.1}$$

where $U_{i,k}$ is the set of users who have given rating k to item i [3]. For each explanation, the expected value of the ratings given by N_u to the recommended item i can be calculated as follows:

$$\mathbf{E}(r_{u,i}|N_u) = \sum_{k \in \kappa} k \times \mathbf{P}(r_{u,i} = k|N_u) \tag{2.2}$$

where κ is the set of rating values [3]. Higher expected values indicate higher NSE explainability of item i for user u. Similarly, ISE can be formulated based on the empirical density distribution of the ratings given by user u to the items that are similar to the recommended item i. Given the set of similar items to item i, N_i, the conditional probability of item i having rating k can be written as:

$$\mathbf{P}(r_{u,i} = k|N_i) = \frac{|N_i \cap I_{u,k}|}{|N_i|} \tag{2.3}$$

where $I_{u,k}$ is the set of items that were given rating k by user u [3]. The expected value of the ratings of user u to the items in the set N_i can be calculated as follows:

$$\mathbf{E}(r_{u,i}|N_i) = \sum_{k \in \kappa} k \times \mathbf{P}(r_{u,i} = k|N_i) \tag{2.4}$$

The expected rating of similar users or similar items, obtained using Eqs. 2.2 or 2.4 gives a reasonable and intuitive measure of goodness or strength of a neighbor-based explanation.

Abdollahi and Nasraoui [2] expanded the EMF technique to Restricted Boltzmann Machines (RBM) and presented an explainability-aware RBM for CF. Bilgic and Mooney [8] proposed a book recommendation system (LIBRA). They argued that the quality of explanation can be measured using two different approaches: the promotion approach or the satisfaction approach. The promotion approach favors the explanation that would convince the user to adopt an item, while the satisfaction approach favors an explanation that would allow the user to assess the quality of (or how much they like) an item best. The conclusion from Bilgic and Mooney is that while the NSE style explanations were top performers in Herlocker et al.'s [29] experiments from

the point of view of "promotion", KSE and next ISE explanations were found to be the top performers from a "satisfaction" perspective. Other than [8], Vig et al. [57] proposed a KSE explanation by introducing tagsplanation, which is generating explanations based on community tags. In their method, they consider a form of content-based explanation. The average of a given user's ratings of the movies with a specific tag defines how relevant a tag is for that user.

Another KSE approach was presented by McCarthy [37]. Their explanation is knowledge and utility based; that is, based on the users' needs and interests. The explanation is presented by describing the matched item for the specified requirements from the user. Zhang et al. [59] proposed an Explicit Factor Model (LFM) to generate explainable recommendations. They extracted explicit product features and user opinions using sentiment analysis. Ardissono et al. [5] built a recommendation system that suggests places to visit based on the travelers' type (e.g. children, impaired). In this case, the explanation comes in the form of the presentation of the recommendation to the user. The demographic information of the user is utilized to group users, and the explanation is focused on the most meaningful types of information for each group.

Billsus and Pazzani [9] presented a keyword style and influence style explanation approach for their news recommendation system which synthesizes speech to read stories to the users. The system generates explanations and adapts its recommendation to the user's interests based on the user's preferences and interests. They ask for a feedback from the user on how interesting the story had been to the user or if the user needs more information. The explanation is then constructed from the retrieved headlines that are closest to the user's interests. An example of their explanation is: "This story received a [high | low] relevance score, because you told me earlier that you were [not] interested in [closest headline]."

Symeonidis et al. [53] proposed a recommendation system that was based on the Feature-Weighted Nearest Bi-cluster (FWNB) algorithm, and they measured the accuracy of the recommendation using precision and recall. Their recommendation is based on finding bi-clusters containing item content features that have strong partial similarity with the test user. The item content features can later be used for justifying the recommendations. Their survey-based user study measured the user satisfaction against KSE, ISE and their own style, called KISE. They designed a user study with 42 pre- and post-graduate students of Aristotle University, who filled out an online survey. Each target user was asked to provide ratings for at least five movies that exist in the MovieLens data set. They then recommended a movie to each target user, justifying their recommendation by using the three justification styles (a different style each time). Finally, target users were asked to rate (in 1–5 rating scale) each explanation style separately to explicitly express their actual preference among the three styles. Subsequent analysis of the mean and standard deviation of the users' ratings for each explanation style, found KISE to outperform all other styles.

Paired t-tests also concluded that the difference between KISE and KSE and ISE was statistically significant at p-value $= 0.01$ level. Although the findings in [8, 53] did not compare with NSE, their study and experiments were similar to those of Bilgic and Mooney [8] who previously found KSE to be the top performer, followed

closely by ISE (then by a margin, NSE). However it is worth mentioning that the data sets in the two studies were different: MovieLens for [53] versus books for [8]. Thus, their item content features are different (genre, keywords, directors, actors collected from the Internet and Movie Database (IMDb) for movies versus keywords in the author, title, description, subject, related authors, related titles, that are crawled from Amazon for books). It is easy to see that the content features for the books in LIBRA draw significantly more on Human Expert knowledge (subject, related authors and book titles) compared to the IMDB-sourced content features of movies in Symeonidis (no related movie titles or related producers).

Regardless of the type of explanation used for CF approaches, most explanation generation techniques reported in the literature are designed for transparent, or white-box methods, such as classical neighborhood-based CF. The prediction is performed as the process of aggregation of the ratings of the neighbor. This process could end up giving weak recommendations which might be discovered with good explanations. Other explanation methods, designed for opaque models such as latent factor models, assume some form of content data or an additional data source for explanations. Therefore, their explanation module is a separate approach from the recommender module which does not reflect the algorithm behind the suggestion made. Therefore, the explanation may, or may not reflect the underlying algorithm used by the system.

Thus it is of great interest to propose explainable CF techniques that computes the top-n recommendation list from items that are explainable in the latent space. To generate latent factors, some well-known latent factor models can be used such as: Matrix Factorization (MF) and Restricted Boltzmann Machines (RBM) methods [1–3].

2.4 Evaluation Metrics for Explainability in Recommender Systems

Evaluation of explanations in recommender systems require user-based metrics to evaluate the perceived quality of the explanation and the efficiency of the justification of the recommendation provided to the user by the explanation. Pu et al. [41] proposed a method that consists of 60 questions to assess the perceived quality of the recommendations such as usefulness, user satisfaction, influence on the users' intention to purchase the recommended product, and so on. However, this questionnaire was designed for user-based evaluation of the recommender system and not the explanation. Herlocker et al. [29] provided some initial explorations into measuring how explanations can improve the filtering performance of users, but their study was more focused on different aspects of the explanation generation than their evaluation. The user-based experiments in the two studies are different in two perspectives: Symeonidis et al. [53] used both (i) a quantitative (objective) metric for justification (coverage ratio) which is based on the amount of influence from content features in the justified recommendation list, and (ii) direct user's 1–5 scale ratings

about how satisfied they are with each explanation style (KSE, ISE or KISE), while Bilgic and Mooney [8] collected the user's satisfaction via analysis of their ratings of the explanations before and after examining the recommended item in question. Furthermore [8] collected the user satisfaction without showing them which explanation method was used and most importantly, they collected the user satisfaction by providing an explanation of why the item was recommended before being shown and examining the item, thus allowing measurement of the user's satisfaction with the explanation itself and not merely the recommendation. Bilgic and Mooney's measure of the quality of an explanation is based on how similar the user's ratings of the recommendation are before and after examining the recommended item, thus measuring the power of the explanation to convey the true nature of the recommended item, even in cases where the recommended item was rated low by the user, and not merely a promotion-based explanation (which accounts only for highly rated recommended items). Despite the apparent limitation of [53], it remains easier to implement because it does not require the user to examine the item being recommended, and because it also computes an objective quantitative measure (based on total contribution of the influence of recommended items' dominant content features relative to the dominant user profile features). These can be computed directly from the ratings data, recommended lists, and explanations, none of which require actual user-based tests.

2.5 Conclusion

Machine learning models are increasingly reliant on data that is being generated at a fast pace. In particular, more and more of this data is related to humans or generated by human activity, and this in turn makes the data susceptible to various forms of human bias. Bias that can originate from the data or the design of the algorithm itself can result in building unfair machine learning models. Therefore, it is important to study and recognize the source of the bias before designing ML models. One way to determine if a model is fair is by incorporating explainability which results in transparency. Prominent examples of ML models are recommender system models that interact directly with humans and whose outputs are consumed directly by humans. Designing explainable recommender system models and explaining recommendations can help enhance the scrutability of the learned models and help detect potential biases, in addition to offering, as additional output, the reasoning behind the predictions. In this chapter, we presented our definition of fairness and transparency in ML models in addition to the main sources of bias that can lead to unfair models. We further reviewed the taxonomy of explanation styles for recommender systems, and reviewed existing models that can provide explanations for their recommendations. We concluded by reviewing several evaluation metrics for assessing the power of explainability in recommender systems.

Acknowledgements This research was partially supported by KSEF Award KSEF-3113-RDE-017 and NSF Grant NSF IIS-1549981.

References

1. Abdollahi, B., Nasraoui, O.: Explainable Matrix Factorization for Collaborative Filtering. In: Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee (2016)
2. Abdollahi, B., Nasraoui, O.: Explainable Restricted Boltzmann Machines for Collaborative Filtering (2016). arXiv preprint arXiv:1606.07129
3. Abdollahi, B., Nasraoui, O.: Using Explainability for Constrained Matrix Factorization. In: Proceedings of the Eleventh ACM Conference on Recommender Systems, pp. 79–83. ACM (2017)
4. Antunes, P., Herskovic, V., Ochoa, S.F., Pino, J.A.: Structuring dimensions for collaborative systems evaluation. ACM Comput. Surv. (CSUR) **44**(2), 8 (2012)
5. Ardissono, L., Goy, A., Petrone, G., Segnan, M., Torasso, P.: Intrigue: personalized recommendation of tourist attractions for desktop and hand held devices. Appl. Artif. Intell. **17**(8–9), 687–714 (2003)
6. Baeza-Yates, R.: Data and algorithmic bias in the web. In: Proceedings of the 8th ACM Conference on Web Science, pp. 1–1. ACM (2016)
7. Beede, D.N., Julian, T.A., Langdon, D., McKittrick, G., Khan, B., Doms, M.E.: Women in STEM: A Gender Gap to Innovation (2011)
8. Bilgic, M., Mooney, R.J.: Explaining recommendations: satisfaction versus promotion. In: Beyond Personalization Workshop, IUI, vol. 5, 153 p. (2005)
9. Billsus, D., Pazzani, M. J.: A personal news agent that talks, learns and explains. In: Proceedings of the Third Annual Conference on Autonomous Agents, pp. 268–275. ACM (1999)
10. Brown, B., Aaron, M.: The politics of nature. In: Smith, J. (ed.) The Rise of Modern Genomics, 3rd edn. Wiley, New York (2001)
11. Broy, M.: Software engineering – from auxiliary to key technologies. In: Broy, M., Dener, E. (eds.). Software Pioneers, pp. 10–13. Springer, Berlin (2002)
12. Calfee, R.C., Valencia, R.R.: APA guide to preparing manuscripts for journal publication. American Psychological Association, Washington, DC (1991)
13. Cameron, D.: Feminism and Linguistic Theory. St. Martin's Press, New York (1985)
14. Cameron, D.: Theoretical debates in feminist linguistics: questions of sex and gender. In: Wodak, R. (ed.) Gender and Discourse, pp. 99–119. Sage Publications, London (1997)
15. Dascalu, M.I., Bodea, C.N., Mihailescu, M.N., Tanase, E.A., Ordoez de Pablos, P.: Educational recommender systems and their application in lifelong learning. Behav. Inf. Technol. **35**(4), 290–297 (2016)
16. Dod, J.: Effective substances. In: The Dictionary of Substances and Their Effects. Royal Society of Chemistry (1999) Available via DIALOG. http://www.rsc.org/dose/titleofsubordinatedocument. Cited 15 Jan 1999
17. Dod, J.: Effective substances. In: The dictionary of substances and their effects. Royal Society of Chemistry (1999). Available via DIALOG. http://www.rsc.org/dose/Effectivesubstances. Cited 15 Jan 1999
18. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017)
19. Dwork, C.: What's Fair?. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1–1. ACM (2017)
20. Fish, B., Kun, J., Lelkes, D.: A confidence-based approach for balancing fairness and accuracy. In: Proceedings of the 2016 SIAM International Conference on Data Mining, pp. 144–152. Society for Industrial and Applied Mathematics (2016)

21. Freitas, A.A.: Comprehensible classification models: a position paper. ACM SIGKDD Explor. Newsl. **15**(1), 1–10 (2014)
22. Geddes, K.O., Czapor, S.R., Labahn, G.: Algorithms for Computer Algebra. Kluwer, Boston (1992)
23. Griffith, A.L.: Persistence of women and minorities in STEM field majors: is it the school that matters? Econ. Educ. Rev. **29**(6), 911–922 (2010)
24. Hajian, S., Bonchi, F., Castillo, C.: Algorithmic bias: from discrimination discovery to fairness-aware data mining. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2125–2126. ACM (2016)
25. Hamburger, C.: Quasimonotonicity, regularity and duality for nonlinear systems of partial differential equations. Ann. Mat. Pura. Appl. **169**, 321–354 (1995)
26. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems, pp. 3315–3323 (2016)
27. Harper, F.M., Konstan, J.A.: The movielens datasets: history and context. ACM Trans. Interact. Intell. Syst. (TiiS) **5**(4), 19 (2016)
28. Harris, M., Karper, E., Stacks, G., Hoffman, D., DeNiro, R., Cruz, P., et al.: Writing labs and the Hollywood connection. J Film Writing **44**(3), 213–245 (2001)
29. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, pp. 241–250. ACM (2000)
30. Ibach, H., Lüth, H.: Solid-State Physics, 2nd edn, pp. 45–56. Springer, New York (1996)
31. John, Alber, O'Connell, Daniel C., Kowal, Sabine: Personal perspective in TV interviews. Pragmatics **12**, 257–271 (2002)
32. Kamiran, F., Calders, T.: Classifying without discriminating. In: 2nd International Conference on Computer, Control and Communication, 2009. IC4 2009, pp. 1-6). IEEE, New York (2009)
33. Kamishima T., Akaho, S., Sakuma, J.: Fairness-aware learning through regularization approach. In: 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW), pp. 643–650. IEEE, New York (2011)
34. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores (2016). arXiv preprint arXiv:1609.05807
35. Kreger, M., Brindis, C.D., Manuel, D.M., Sassoubre, L. (2007). Lessons learned in systems change initiatives: benchmarks and indicators. Am. J. Commun. Psychol. https://doi.org/10.1007/s10464-007-9108-14
36. Lipton, Z.C.: The Mythos of Model Interpretability (2016). arXiv preprint arXiv:1606.03490
37. McCarthy, K., Reilly, J., McGinty, L., Smyth, B.: Thinking positively-explanatory feedback for conversational recommender systems. In: Proceedings of the European Conference on Case-Based Reasoning (ECCBR-04) Explanation Workshop, pp. 115–124 (2004)
38. O'Neil, J.M., Egan, J.: Men's and women's gender role journeys: metaphor for healing, transition, and transformation. In: Wainrig, B.R. (ed.) Gender Issues Across the Life Cycle, pp. 107–123. Springer, New York (1992)
39. Pohl, R. (ed.).: Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory. Psychology Press (2004)
40. S. Preuss, A. Demchuk Jr., M. Stuke, Appl. Phys. A **61** (1995)
41. Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems, pp. 157–164. ACM, Chicago (2011)
42. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144. ACM (2016)
43. Sapiezynski, P., Kassarnig, V., Wilson, C.: Academic performance prediction in a gender-imbalanced environment (2017)
44. Sharma, A., Cosley, D.: Do social explanations work?: studying and modeling the effects of social explanations in recommender systems. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1133–1144. ACM (2013)

45. Slifka, M.K., Whitton, J.L.: Clinical implications of dysregulated cytokine production. J. Mol. Med. (2000). https://doi.org/10.1007/s001090000086
46. Slifka, M.K., Whitton, J.L.: Clinical implications of dysregulated cytokine production. J. Mol. Med. (2000). https://doi.org/10.1007/s001090000086
47. M.K. Slifka, J.L. Whitton, J. Mol. Med. https://doi.org/10.1007/s001090000086
48. S.E. Smith, in *Neuromuscular Junction*, ed. by E. Zaimis. Handbook of Experimental Pharmacology, vol 42 (Springer, Heidelberg, 1976), p. 593
49. Smith, E.: Women into science and engineering? gendered participation in higher education STEM subjects. Br. Educ. Res. J. **37**(6), 993–1014 (2011)
50. Smith, J., Jones Jr., M., Houghton, L., et al.: Future of health insurance. N. Eng. J. Med. **965**, 325–329 (1999)
51. South, J., Blass, B.: The Future of Modern Genomics. Blackwell, London (2001)
52. Suleiman, C., O'Connell, D.C., Kowal, S.: If you and I, if we, in this later day, lose that sacred fire...': Perspective in political interviews. J. Psycholinguist. Res. (2002). https://doi.org/10.1023/A:1015592129296
53. Symeonidis, P., Nanopoulos, A., Manolopoulos, Y.: Justified recommendations based on content and rating data. In: WebKDD Workshop on Web Mining and Web Usage Analysis (2008)
54. Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., Schmidt-Thieme, L.: Recommender system for predicting student performance. Procedia Comput. Sci. **1**(2), 2811–2819 (2010)
55. Tintarev, N., Masthoff, J.: Designing and evaluating explanations for recommender systems. In: Recommender Systems Handbook, pp. 479–510 (2011)
56. Tintarev, N., Masthoff, J.: Evaluating the effectiveness of explanations for recommender systems. User Model. User-Adap. Inter. **22**(4), 399–439 (2012)
57. Vig, J., Sen, S., Riedl, J.: Tagsplanations: explaining recommendations using tags. In: Proceedings of the 14th International Conference on Intelligent User Interfaces, pp. 47–56. ACM (2009)
58. Yao, S., Huang, B.: New Fairness Metrics for Recommendation that Embrace Differences (2017). arXiv preprint arXiv:1706.09838
59. Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., Ma, S.: Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 83–92. ACM (2014)