A Related Work

Machine Learning (ML) is being increasingly used to automate decisions. Some of the applications where ML is being used are highly critical and directly affect humans, for example, loan approval [45], criminal justice [48], and hiring [41]. The nascent field of trustworthy ML aims to detect bias in ML models (and counteract it), understand the factors that the ML model is using in making predictions, ensure the models respect privacy and security, and frame policies and regulations that the ML models should abide by [5, 6, 58]. Research has established a few dimensions of trustworthy ML, like, fairness [6, 12, 20, 23, 54, 61, 62], interpretability [4, 15, 18, 30, 34, 35, 39, 40], robustness [11, 13, 19, 27, 36, 44, 59, 63]. In this work, we focus on interpretability and refer the readers to work by Barocas et al. [6] and Varshney [52] for a broad discussion of trustworthy machine learning.

A.1 Interpretability in ML

Interpretability is the branch of trustworthy ML that aims to provide human consumable explanations for the predictions made by ML models that are used for tasks such as classification, regression, and recommendation. Most of the research in this area has focused on the interpretability of classification models. Interpretability for classification models can be achieved by either developing inherently interpretable ML models (e.g., logistic regression, shallow decision trees) or a post-hoc explanation of complex ML models (e.g., random forests, neural networks). Post-hoc explanations can be further bifurcated into generating feature attributions using techniques like SHAP [32] or generating counterfactual explanation-based recourses [55]. Feature attribution explanations highlight the features that might have been important in making a prediction. On the other hand, counterfactual explanations provide a counterfactual situation that would have led to a different prediction from the ML model. Miller [33] in a social science study remarked that when people ask 'Why P?' questions, they are typically asking 'Why P rather than Q?', where Q is implicit in the context of the application. An example of this case is the question a person whose loan request has been rejected would ask: 'Why has my loan request been rejected?', which actually means: 'Why has my loan request been rejected instead of being accepted?'. And counterfactual explanations are a way to answer this question. They would respond, for example, by saying that 'had your income been $3000 higher, you would have gotten the loan'. This simultaneously also provides a recourse to the affected individual, who now knows that they can get the loan if they can increase their income by $3000.

A.2 Interpretability in Recommender Systems

Literature in interpretability for recommender systems has focused on highlighting the factors that might have contributed to a recommendation. This is similar to feature attribution based explanations for classification models. Interpretability research for recommender systems can be categorized into user-based, item-based, and feature-based explanations. In user-based explanations, a high rating for the item provided by a group of users similar to the user is given as an explanation for the recommended item. In item-based explanations, the recommended item is explained by its similarity to the items that the user has liked or purchased in the past. Feature-based explanations highlight the features of the recommended item that the user has shown interest in the past, for example, the cast for movie recommendations. The approaches that generate these explanations can either be model-specific or model agnostic. We refer the readers to a survey on the explainability of recommender systems for a more comprehensive discussion [64] on this topic. Similar to feature attribution explanation for classification models, a noticeable characteristic of the aforementioned explanations for recommender systems is that they are not actionable.

RecRec, on the other hand, generates recourses targeted towards the content providers of the recommender systems. This is the main contribution of this work. It provides counterfactual features that would lead to a different ranking of a specific item in the recommendation list for a target group of users.

A.2.1 Previous studies on need for recourse for content providers In this subsection we continue to discuss related work that has highlighted the need for recourse for content providers. Jhaver et al. [25] did a study with several Airbnb hosts to understand their perspectives. They clearly expressed the need for transparency and recourse on the platform. One of the hosts said: *"I feel less motivated because I don't think that it's clear what I need to do, and I think that it's frustrating seeing the search: lots of listings that are worse than mine are in higher positions."* Several hosts performed A/B testing with different factors like pricing adjustments, calendar updates, location, type of room, amenities like free parking, changing descriptions of the property, allowing dogs, allowing short-term vs. long-term guests, etc., to understand which factors can help improve their ranking. Rahman [37] interviewed freelancers working on Upwork. Freelancers also struggled in understanding what factors go into the ranking and how they can influence it. One of the freelancers said: *"all I can think about is figuring out how to raise my score"*. Similar to Airbnb hosts, freelancers tried and tested changing different attributes of their profile, like taking technical tests provided by Upwork, opening and closing contracts, having shorter projects, and inflating the hourly working rate to improve their ranking. A precisely similar need and behavior was observed when studies were conducted with sellers on Facebook MarketPlace [16], freelancers on other platforms like TaskRabbit and Fiverr [9, 24], drivers using Uber and Lyft [29], and sellers on handmade product platform Etsy [38]. Several freelancers who Rahman [37] interviewed mentioned that owing to the black-box behavior of the algorithmic freelancing platform, they made frequent efforts to take the work offline or pause the work to pacify the algorithm, and some even quit. Jarrahi and Sutherland [24] had the same observations in their interviews with Upwork's freelancers. Providing recourses to content providers of a platform would help them understand what actions they can take to improve their product's ranking and get transparency into the current ranking.

A.2.2 Counterfactual Explanations in Recommender Systems There have been recent proposals for some approaches that seek to generate explanations for recommended item in a counterfactual manner [17, 26, 47, 51, 60]. All these approaches explain a recommendation by finding the smallest change in the user's interaction history that would replace the top recommendation with anything else. For example, an explanation for the top-recommended movie *The*

Godfather II is that the user had previously liked *Goodfellas* and *The Godfather*. Had the user not liked these two movies, *The God-Father II* would not be the top recommendation (it could still be recommended but at a different rank). RecRec is distinct from these works in several ways:

- RecRec provides recourse to the content providers of the recommender system, while the aforementioned approaches provide explanations to the users of the recommender system.
- RecRec provides a set of actions that can be executed to get a favorable rank for an item while the aforementioned approaches do not provide that. They only find the smallest change that would replace the top recommendation with anything else, not something the content provider or the user wants.
- RecRec can make suggestions to change features that are *not in* the user's history, while the aforementioned approaches only alter the user's history.
- RecRec is able to provide a recourse for items at any rank (in order to get them to an improved rank) in the recommended list, while the aforementioned approaches provide an explanation for only the top-ranked item.

Our work also has subtle similarity to the work by Dean et al. [14], where they define reachability as the feasibility of the end-user of a recommender system modifying their rating in order to get an item recommended. RecRec is distinct from Dean et al. [14]'s work in the following ways:

- Their work is concerned with only the end-user of a recommender system, while RecRec is targeted towards the content providers of the recommender system.
- The goal of their work is to audit the recommender system to understand whether it could cause polarization or filter bubbles, while the goal of RecRec is to provide recourse to the content providers of the recommender system.
- Their approach is limited to matrix factorization based recommender systems, while RecRec generalizes not only to all architectures of collaborative filtering recommender systems, but also to content based recommender systems.

References

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Recommender Systems as Multistakeholder Environments. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (Bratislava, Slovakia) *(UMAP '17)*. Association for Computing Machinery, New York, NY, USA, 2 pages. https://doi.org/10.1145/3079628.3079657

[2] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering* 17, 6 (jun 2005), 734–749.

[3] Charu C Aggarwal. 2016. Content-based recommender systems. In *Recommender systems*. Springer, 139–166.

[4] Muhammad Aurangzeb Ahmad, Ankur Teredesai, and Carly Eckert. 2018. Interpretable Machine Learning in Healthcare. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. 447–447. https://doi.org/10.1109/ICHI.2018.00095

[5] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. https://doi.org/10.48550/ARXIV.1606.06565

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[7] Thomas Blumensath and Mike E. Davies. 2009. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis* 27, 3 (2009), 265–274. https://www.sciencedirect.com/science/article/pii/S1063520309000384

[8] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based systems* 46 (2013), 109–132.

[9] Eliane Léontine Bucher, Peter Kalum Schou, and Matthias Waldkirch. 2020. Pacifying the algorithm - Anticipatory compliance in the face of algorithmic management in the gig economy. *Organization* 28 (2020), 44 – 67.

[10] Laurent Candillier, Kris Jack, Françoise Fessant, and Frank Meyer. 2009. State-of-the-art recommender systems. In *Collaborative and Social Information Retrieval and Access: Techniques for Improved User Modeling*. IGI Global, 1–22.

[11] Nicholas Carlini and David A. Wagner. 2017. Towards Evaluating the Robustness of Neural Networks. *2017 IEEE Symposium on Security and Privacy (SP)* (2017), 39–57.

[12] Alexandra Chouldechova and Aaron Roth. 2020. A Snapshot of the Frontiers of Fairness in Machine Learning. *Commun. ACM* 63, 5 (apr 2020), 82–89. https://doi.org/10.1145/3376898

[13] Anupam Datta, Matt Fredrikson, Klas Leino, Kaiji Lu, Shayak Sen, and Zifan Wang. 2021. *Machine Learning Explainability and Robustness: Connected at the Hip*. Association for Computing Machinery, New York, NY, USA, 4035–4036. https://doi.org/10.1145/3447548.3470806

[14] Sarah Dean, Sarah Rich, and Benjamin Recht. 2020. Recommendations and User Agency: The Reachability of Collaboratively-Filtered Information. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 10 pages. https://doi.org/10.1145/3351095.3372866

[15] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [stat.ML]

[16] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I Always Assumed That I Wasn't Really That Close to [Her]": Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 10 pages. https://doi.org/10.1145/2702123.2702556

[17] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. 2020. PRINCE: Provider-Side Interpretability with Counterfactual Explanations in Recommender Systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 9 pages. https://doi.org/10.1145/3336191.3371824

[18] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 80–89. https://doi.org/10.1109/DSAA.2018.00018

[19] Leif Hancox-Li. 2020. Robustness in Machine Learning Explanations: Does It Matter?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 640–647. https://doi.org/10.1145/3351095.3372836

[20] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. arXiv:1610.02413 [cs.LG]

[21] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872

[22] Maya Holikatti, Shagun Jhaver, and Neha Kumar. 2019. Learning to Airbnb by Engaging in Online Communities of Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019), 19 pages. https://doi.org/10.1145/3359330

[23] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. *Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?* Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300830

[24] Mohammad Hossein Jarrahi and Will Sutherland. 2019. Algorithmic Management and Algorithmic Competencies: Understanding and Appropriating Algorithms in Gig Work. In *Information in Contemporary Society*. Springer International Publishing, Cham, 578–589.

[25] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 12 pages. https://doi.org/10.1145/3173574.3173995

[26] Vassilis Kaffes, Dimitris Sacharidis, and Giorgos Giannopoulos. 2021. Model-Agnostic Counterfactual Explanations of Recommendations. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 6 pages. https://doi.org/10.1145/3450613.3456846

[27] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial Machine Learning at Scale. arXiv:1611.01236 [cs.CV]

[28] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5 (2018).

[29] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 10 pages. https://doi.org/10.1145/2702123.2702548

[30] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* 16, 3 (jun 2018), 31–57. https://doi.org/10.1145/3236386.3241340

[31] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. *Recommender systems handbook* (2011), 73–105.

[32] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[33] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019). https://doi.org/10.1016/j.artint.2018.07.007

[34] Christoph Molnar, Giuseppe Casalicchio, and B. Bischl. 2020. Interpretable Machine Learning - A Brief History, State-of-the-Art and Challenges. In *PKDD/ECML Workshops*.

[35] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. arXiv:1909.09223 [cs.LG]

[36] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, Rujun Long, and Patrick McDaniel. 2018. Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. arXiv:1610.00768 [cs.LG]

[37] Hatim A. Rahman. 2021. The Invisible Cage: Workers' Reactivity to Opaque Algorithmic Evaluations. *Administrative Science Quarterly* 66, 4 (2021), 945–988. https://doi.org/10.1177/00018392211010118

[38] Lubna Razaq, Beth Kolko, and Gary Hsieh. 2022. Making crafting visible while rendering labor invisible on the Etsy platform. In *Designing Interactive Systems Conference 2021* (Virtual Event, USA) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 15 pages.

[39] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2021. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *ArXiv* abs/2103.11251 (2021).

[40] Philipp Schmidt and Felix Biessmann. 2019. Quantifying Interpretability and Trust in Machine Learning Systems. arXiv:1901.08558 [cs.LG]

[41] Kumba Sennaar. 2019. Machine Learning for Recruiting and Hiring – 6 Current Applications. https://emerj.com/ai-sector-overviews/machine-learning-for-recruiting-and-hiring/. Accessed: 2020-10-15.

[42] Kunal Shah, Akshaykumar Salunke, Saurabh Dongare, and Kisandas Antala. 2017. Recommender systems: An overview of different approaches to recommendations. In *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. IEEE, 1–4.

[43] Guy Shani and Asela Gunawardana. 2011. *Evaluating Recommendation Systems*. Springer US, Boston, MA, 257–297. https://doi.org/10.1007/978-0-387-85820-3_8

[44] Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin T Vechev. 2018. Fast and Effective Robustness Certification. *NeurIPS* 1, 4 (2018), 6.

[45] Saurav Singla. 2020. Machine Learning to Predict Credit Risk in Lending Industry. https://www.aitimejournal.com/@saurav.singla/machine-learning-to-predict-credit-risk-in-lending-industry. Accessed: 2020-10-15.

[46] Özge Sürer, Robin Burke, and Edward C. Malthouse. 2018. Multistakeholder Recommendation with Provider Constraints. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240323.3240350

[47] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. 2021. Counterfactual Explainable Recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, New York, NY, USA, 10 pages.

[48] Jason Tashea. 2017. Courts Are Using AI to Sentence Criminals. That Must Stop Now. https://www.wired.com/2017/04/courts-using-ai-sentence-criminals-must-stop-now/. Accessed: 2020-10-15.

[49] Poonam B Thorat, RM Goudar, and Sunita Barve. 2015. Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications* 110, 4 (2015), 31–36.

[50] Tianchi. 2018. Ad Display/Click Data on Taobao.com. https://tianchi.aliyun.com/dataset/dataDetail?dataId=56

[51] Khanh Hiep Tran, Azin Ghazimatin, and Rishiraj Saha Roy. 2021. *Counterfactual Explanations for Neural Recommenders*. Association for Computing Machinery, New York, NY, USA, 1627–1631. https://doi.org/10.1145/3404835.3463005

[52] Kush R. Varshney. 2021. *Trustworthy Machine Learning*. Chappaqua, NY, USA. http://trustworthymachinelearning.com.

[53] Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual Explanations for Machine Learning: A Review. arXiv:2010.10596 [cs.LG]

[54] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness* (Gothenburg, Sweden) *(FairWare '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3194770.3194776

[55] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. arXiv:1711.00399 [cs.AI]

[56] Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*. ACM, 86–94. https://doi.org/10.1145/3240323.3240369

[57] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-Grained Spoiler Detection from Large-Scale Review Corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Association for Computational Linguistics, 2605–2610. https://doi.org/10.18653/v1/p19-1248

[58] Adrian Weller. 2017. Transparency: Motivations and Challenges. https://doi.org/10.48550/ARXIV.1708.01870

[59] Huan Xu and Shie Mannor. 2012. Robustness and generalization. *Machine learning* 86, 3 (2012), 391–423.

[60] Shuyuan Xu, Yunqi Li, Shuchang Liu, Zuohui Fu, and Yongfeng Zhang. 2020. Learning Post-Hoc Causal Explanations for Recommendation.

[61] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1171–1180. https://doi.org/10.1145/3038912.3052660

[62] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28* (Atlanta, GA, USA) *(ICML'13)*. JMLR.org, 325–333.

[63] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 7472–7482. https://proceedings.mlr.press/v97/zhang19p.html

[64] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Found. Trends Inf. Retr.* 14 (2020), 1–101.

[65] Yong Zheng. 2019. Multi-Stakeholder Recommendations: Case Studies, Methods and Challenges *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 2 pages. https://doi.org/10.1145/3298689.3346951