| | |
|---|---|
| **Document Title** | **Survey article** |
| **Project Title and acronym** | Cyprus Center for Algorithmic Transparency (CyCAT) |
| **H2020-WIDESPREAD-05-2017-Twinning** | Grant Agreement number: 810105 — CyCAT |
| **Deliverable No.** | D3.4 |
| **Work package No.** | WP3 |
| **Work package title** | Mitigating Bias in Algorithmic Systems: A Fish-Eye View of Problems and Solutions across Domains |
| **Authors (Name and Partner Institution)** | Fausto Giunchiglia (UNITN)<br>Kalia Orphanou (OUC)<br>Jahna Otterbacher (OUC) |
| **Contributors (Name and Partner Institution)** | Veronika Bogin (UH)<br>Alan Hartman (UH)<br>Styliani Kleanthous (OUC)<br>Tsvi Kuflik (UH)<br>Avital Shulner Tal (UH) |
| **Reviewers** | Maria Kasinidou (OUC) |
| **Status (D: draft; RD: revised draft; F: final)** | F |
| **File Name** | D3.4_Survey_article.docx |
| **Date** | 15 May 2020 |

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 810105.

| **Draft Versions - History of Document** | | | | |
|---|---|---|---|---|
| **Version** | **Date** | **Authors / contributors** | **e-mail address** | **Notes / changes** |
| V1.0 | 31/10/19 | Kalia Orphanou | kalia.orphanou@ouc.ac.cy | Initial version |
| V2.0 | 12/5/20 | Jahna Otterbacher | Jahna.otterbacher@ouc.ac.cy | Manuscript for review |
| V3.0 | 15/5/20 | Jahna Otterbacher | Jahna.otterbacher@ouc.ac.cy | Manuscript for submission |

| **Abstract** | |
|---|---|
| This deliverable constitutes a comprehensive survey of the literature on mitigating algorithmic bias. It represents the synthesis of all the work carried out in WP3. The survey develops a conceptual framework for understanding the problem and solution spaces of algorithmic bias, as well as the roles of various stakeholders. The manuscript was prepared as a submission to the journal ACM Computing Surveys. | |



| | |
|---|---|
| **Keyword(s):** | Algorithmic bias, ACM Computing Surveys, conceptual framework, literature review |

**Contents**

1. Executive Summary 5
2. Affected Attributes 5
3. Problem Space 6
4. Solution Space 8
5. Conclusion 11
6. References 12

# 1. Executive Summary

D3.4 is a survey paper detailed our understanding of the state-of-the-art in the emerging field of *Mitigating Bias in Algorithmic Systems*, based on 12 months of intensive, collaborative work with the existing published literature. Mitigating bias in algorithmic systems is a critical issue drawing attention across communities within the information and computer sciences. Given the complexity of the problem and the involvement of multiple stakeholders – including developers, end-users and third-parties – there is a need to understand the landscape of the sources of bias, and the solutions being proposed to address them. This deliverable provides a "fish-eye view" examining approaches across four areas of research: machine learning (ML), human-computer interaction (HCI), recommender systems (RecSys), and information retrieval (IR).

The literature describes three steps toward a comprehensive treatment – bias detection, fairness and explainability management – and underscores the need to work from within the system as well as from the perspective of stakeholders in the broader context. The survey aims to help the reader achieve a high-level understanding of the current state of bias in algorithmic systems across the four domains and to describe opportunities for cross-fertilization between communities. It presents a fish-eye view of the literature surrounding algorithmic bias, its problem and solution spaces in order the user to maintain perspective of the "big picture", but can still choose when to drill down into further details.

# 2. Affected Attributes

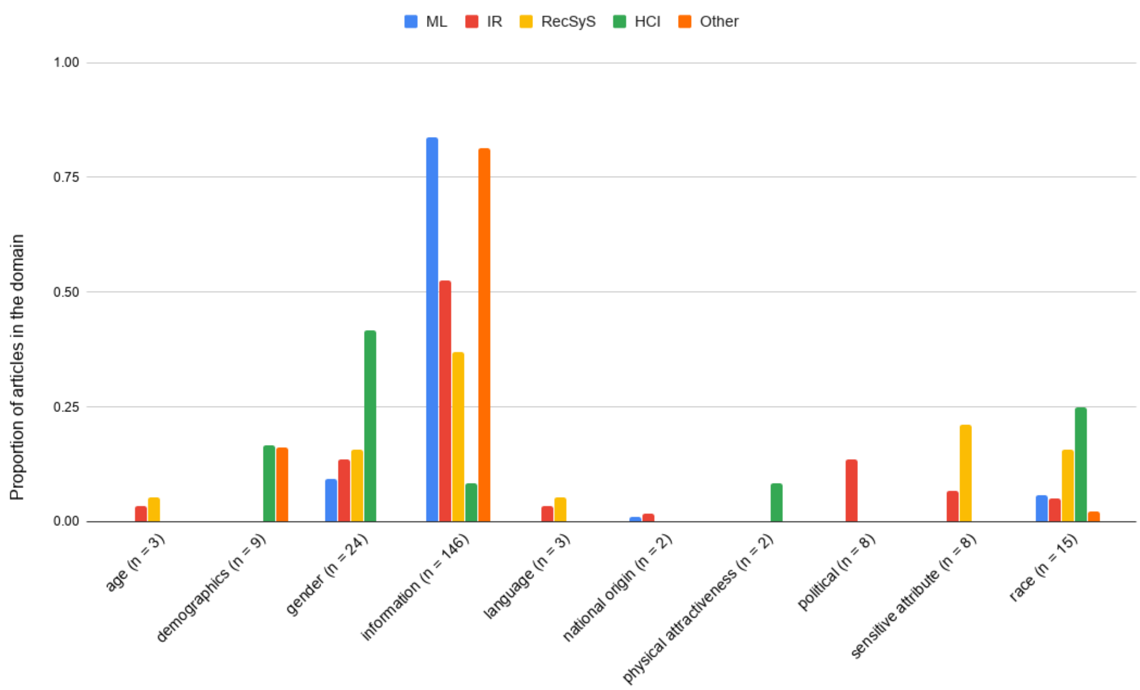


Fig. 1: Affected attributes in the surveyed articles

We also characterized, for a given article, the attribute(s) affected by the problematic system behavior. While early technical works often discussed generic sensitive attributes [108], we recorded the specific attribute of interest in the respective research. Thus, we follow the more recent work in socio-technical systems that considers how specific dimensions, such as the social, cultural, and political attributes of the content or person being processed, may be affected by algorithmic behaviors. Fig. 1 analyzes the frequency with which specific attributes were examined in the literature we surveyed, across the respective domains. In particular, the chart presents the proportion of articles within a given domain that discussed each attribute. As can be observed, across all articles, we find 10 attributes described; note that some researchers describe more/less specific attributes (e.g. demographics or sensitive attribute vs. gender, race or natural origin). Frequencies across the entire corpus are detailed on the horizontal axis.

Information is the most frequently studied attribute in our corpus, and is the primary dimension addressed in the ML literature. For instance, in the explainability literature, a primary concern is the extent to which information is effectively conveyed to the user. Likewise, IR articles often consider information as the affected dimension under study; here, the classic example is the large body of work on search engine biases. In contrast, the literature in HCI and RecSys do not often address information as an affected dimension. In these fields, articles on mitigating algorithmic biases more often consider social and cultural dimensions, such as demographics (generally), gender, and race, with a few studies on attributes such as age, language and physical attractiveness also emerging.

## 3. Problem Space

Fig. 2 provides a general characterization of an algorithmic system, with its macro components, which we have used to examine the problem space of algorithmic bias. In this generic architecture, the system receives input (I) for an instance of its operation. This is provided by a user (U), or another source (e.g., the result of an automated process). The algorithmic model (M) makes some computation(s) based on the inputs and produces an output (O). The model learns from a set of observations of data (D) from the problem domain. It may also receive constraints from third-party actors (T) and/or internal fairness criteria (F) which modify the operation of the algorithmic model (M). Finally, some systems have direct interaction with a user (U) who, as previously discussed, will bring her own knowledge, background and attitude when interpreting the system’s output.

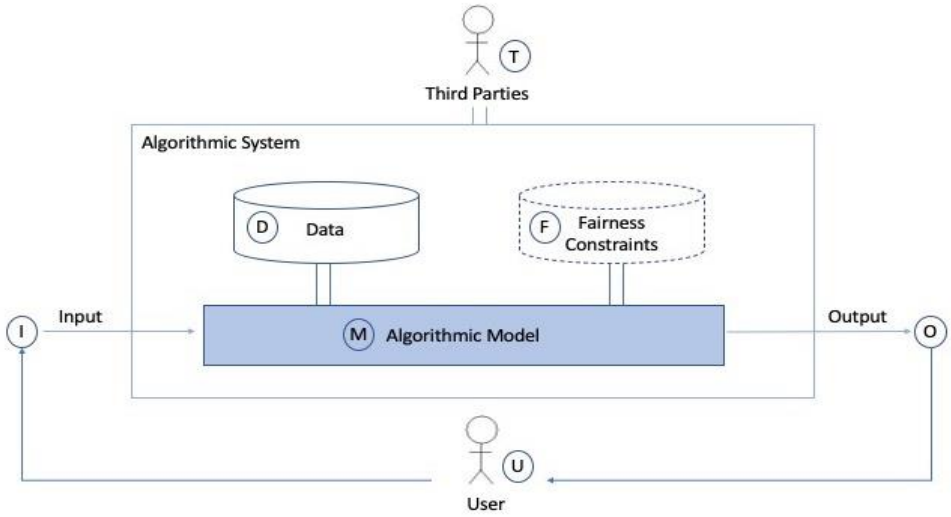


Fig. 2: Generic architecture of an algorithmic system

Thus, bias may manifest and/or be detected in one or more of these components:

- Input (I) - Bias may be introduced in the input data, e.g., as incorrect or incomplete information input by the user.
- Output (O) - Bias may be detected at the outcome (value(s)/label(s)) produced in response to the input.
- Algorithm (M) - Bias can manifest during the model's processing and learning.
- Training Data (D) - Training data may be inaccurate, imbalanced, and/or unrepresentative. Furthermore, it may contain information about sensitive attributes of people.
- Third Party Constraints (T) - Implicit and explicit constraints, given by third parties, may impact the design and performance of the algorithm and cause discrimination and fairness issues. These include operators of the system, regulators, and other bodies which influence the use and outcomes of the system.
- Fairness Constraints (F) - Fairness constraints may be introduced within the system, such that one interpretation of fairness is prioritized over others [77].
- User (U) - When users interact directly with a system, they may contribute to bias in a number of ways, such as through the inappropriate use of the system or misinterpretation of system output.

# 4. Solution Space

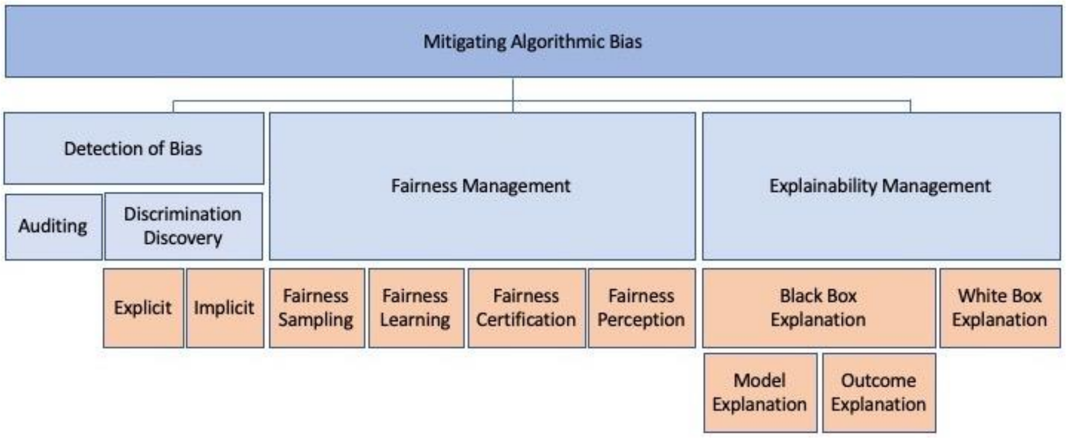


Fig. 3: The solution space for mitigating bias

The literature suggests that a comprehensive solution for mitigating algorithmic bias consists of three main steps (Fig. 3):

- **Detection of Bias:** This involves scrutinizing the system to detect any type of systematic bias. The two main approaches for detecting bias in an algorithmic system, which are described in the literature are: Auditing and direct/indirect Discrimination discovery. As Table 1 shows, in machine learning systems, discrimination detection is mostly done by implicit/explicit discrimination discovery methods which include measuring discrimination or using a causal Bayesian network. Auditing in ML systems is mostly done by a black-box auditing software tool or when auditors search for any bias through the dataset. In IR, HCI and RecSys systems, users mostly act as auditors by submitting different queries in search engines and social networks or by taking the role of crowdworker in the crowdsourcing conducted studies.

| Domain | Problem | Solution Space | Reference(s) |
|---|---|---|---|
| **Detection of Bias** | | | |
| ML | Data/Model | Auditing | Situational and testing auditing [91, 159]<br>Automatic auditing tool [121] |
| ML | Data<br>Data/Model/Output | Discrimination Discovery | Discrimination metrics [155, 164]<br>Data mining methods<br>[28, 32, 87, 108, 159] |
| IR | User/Data/Output | Auditing | Submit queries to search engines/ Twitter<br>[65, 73, 83, 86, 93, 104, 142] |
| IR | User/Data/Output<br><br>User/Third Party/Data<br>User/Third Party | Discrimination Discovery | Analysis of web logs<br>[8, 24, 106, 144, 147–149, 154]<br>Crowdsourcing studies[41, 89]<br>Direct discrimination of perceived bias<br>[6, 66, 145, 146] |
| HCI | Output/Model/User | Auditing | Analysing system behavior<br>[69, 95] |
| HCI | Data/User/Third Party | Discrimination Discovery | Crowdsourcing studies<br>[7, 31, 53, 110] |
| RecSys | Data/User | Auditing | Auditing application systems<br>[39, 44] |
| RecSys | User/Model/Output<br><br>Output/Model | Discrimination Discovery | Discrimination detection in advertising recommendation systems[3, 131, 134]<br>Discrimination detection in evaluation metrics<br>[10, 42] |

Table 1: Summary of the problem and bias detection solution space per domain

- **Fairness Management:** includes the techniques developers use to mitigate the detected bias and certify that the system is fairness-aware. Fairness management approaches can be classified into: Fairness sampling (or pre-processing), Fairness learning (or in-processing) and Fairness certification (or post-processing methods). Pre-processing methods handle bias in input data, in-processing methods concern the mitigation of bias in the algorithm and post-processing methods concern the elimination of bias in the outcome. As displayed in Table 2, in machine learning algorithmic systems, data mining techniques are used to mitigate bias either in the data, in the model processing or at the outcome decision. User-focus systems such as information retrieval, recommender systems and human-computer interface systems use mostly pre-processing approaches such as fairness sampling and feature selection to handle bias in data.

| Fairness Management | | | |
|---|---|---|---|
| **Domain** | **Problem** | **Solution Space** | **Reference(s)** |
| ML | Data | Fairness Sampling | Removing protected attributes<br>& transform the training data [18, 68]<br>Causal BN[70, 85, 159] |
| ML | Model/Third Party<br>Model/Output<br>Model/Output<br>Model/Output<br>Data/Model | Fairness Learning | Fairness constraints [21, 35, 77, 157]<br>Fairness metrics [54, 75, 151]<br>Regularization approach [71]<br>Counterfactual fairness [84]<br>Encrypted version of sensitive data [76] |
| IR | Data | Fairness Sampling | Data sampling [34, 36, 52, 127] |
| IR | User/Output<br><br>Model | Fairness Learning | User's interaction with system<br>[62, 100]<br>Mitigating search engine bias [96] |
| HCI<br><br>HCI | Data<br>Model<br>Model | Fairness Sampling<br><br>Fairness Learning | Data sampling [69]<br>Automated generated data [122]<br>Human in the loop approach[16, 74] |
| RecSys | Data | Fairness Sampling | Data sampling [72, 91] |
| RecSys | Model<br>Model<br>Model/Output | Fairness Learning | Error-based Fairness Criteria [82]<br>Fair top-k ranking algorithm [22, 130, 156]<br>Optimization approaches [98, 152] |
| ML | Third Party/Output<br>Output<br>User/Third Party | Fairness Certification | Altering of labels [58, 70]<br>Altering of confidence of classification rules [108]<br>Perceived fairness management [132] |
| IR | Output/User | Fairness Certification | Raise user awareness[43] |
| HCI | Output/User | Fairness Certification | Perceived fairness management [88, 150] |

Table 2: Summary of the problem and fairness management solution space per domain

- **Explainability Management:** is applied to the system to facilitate transparency and to build trust between Observers/Users and the system. Explainability approaches have primarily been developed in the context of ML algorithms and systems. However, there is a growing literature within the HCI and IR communities. These works suggest that explainability and judgement of the outcome or decision of the system should be provided in order to enhance the trust of the end user in the system. As displayed in Table 3, in ML systems, the explainability method is usually based on the algorithm used in the system, considering whether it is an interpretable algorithm (white-box) or a black-box model such as deep learning. The explainability approaches also concern either the explainability of how the algorithm works or of the algorithm's outcome. There are also the model-agnostic explanation approaches that explain the output of any classifier, regardless of the machine learning algorithm used to train it. Finally, explainability approaches have also been widely discussed in recommender systems. The difference between these approaches and the ones used in ML are that they take into consideration the user's perception and specific goal of increasing the trust of the end-user in the system. In RecSys literature, various explanation styles have been reviewed according to the purpose of providing explanations in a recommender system e.g., transparency, scrutability, trust, etc.

| Explainability Management | | | |
|---|---|---|---|
| **Domain** | **Problem** | **Solution Space** | **Reference(s)** |
| ML | Model<br>Model<br>Model<br>Data/Model | Black-box Model Expainability | Decision tree mimic black-box [15, 26, 37, 51, 81, 124, 135, 161]<br>Decision rules explaining black-box [30, 67, 90, 136, 162]<br>Deep weighted averaging classifier [20]<br>Feature-based explanation [60, 141] |
| ML | Output<br>Output<br>Output/User | Black-box Outcome Explainability | Model-agnostic explanations [32, 109, 113, 133]<br>LIME Explanations [55, 113, 114, 139]<br>Visualization methods [12, 47, 125, 129, 153, 161, 163]<br>Convert to a linear model [59] |
| HCI | Output/Data<br>User/Output<br>User/Output<br>User/Output | White & Black-box Outcome Explainability | Feature-based explanation [64]<br>Taxonomy of explanations[49]<br>Raise user awareness [111]<br>Explanation styles [11, 40] |
| RecSys | Model/User<br>Model/User<br>Output/User<br>Output/User<br>Output/User | Black-box Model Explainability | Taxonomy of concepts [102]<br>Based on user opinions [143]<br>Uniform explanation style [138]<br>Hybrid explanations [14, 80, 140]<br>Matrix-factorization [25] |

Table 3: Summary of the problem and explainability management solution space per domain

# 5. Conclusion

We provided a “fish-eye view" of research to date on the mitigation of bias in any type of algorithmic system. With the aim of raising awareness of biases in user-focused, and algorithmic-focus systems, we examined studies conducted in four different research communities: information retrieval (IR), human-computer interaction (HCI), recommender systems (RecSys) and machine learning (ML). We outlined a classification of the solutions described in the literature for detecting bias as well as for mitigating the risk of bias and managing fairness in the system. Multiple stakeholders, including the developer (or anyone involved in the pipeline of a system’s development), and various system observers (i.e., stakeholders who are not involved in the development, but who may use, be affected by, oversee, or even regulate the use of the system) are involved in mitigating bias.

Fig. 4 summarizes at a high-level, the problems, solutions and stakeholders involved in mitigating algorithmic bias. Basically, the Developers, as the only stakeholders with full access to the interworking of the system, they are the only ones who can be involved in all three steps: detection of bias, fairness management, and explainabiity management. On the other hand, Observers and Users are more limited in their access to the system’s interworkings. Therefore, in terms of detecting bias, they are typically involved in auditing the system. Explainability management is positioned between the inside of the system and the user interface. Therefore, it plays a very important role in winning the user’s trust in the system. Finally, Fairness Perception, which is somehow related to both fairness management and explainability management, is not yet depicted in this figure, and is food for our future research.

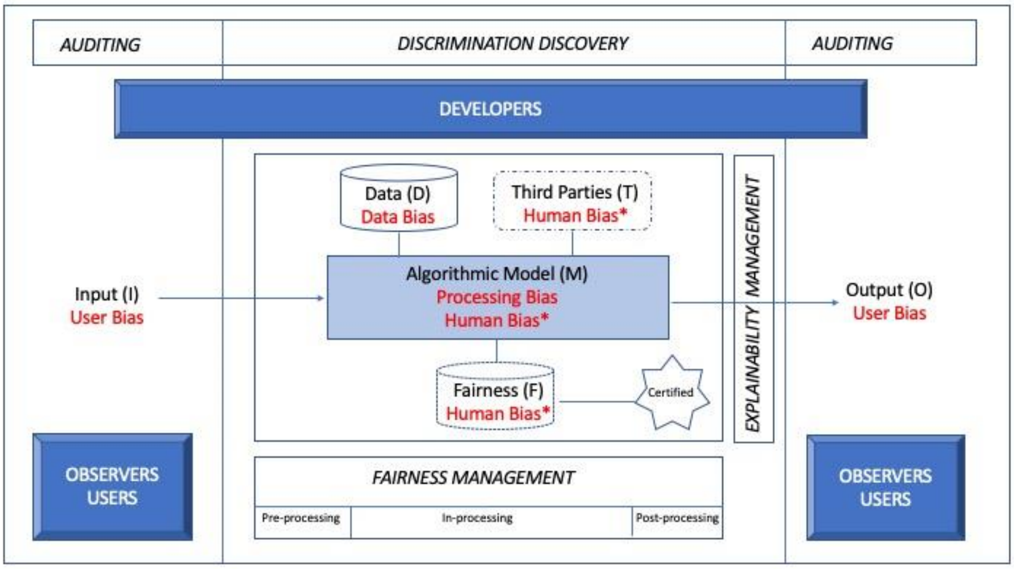


Fig. 4: High-level summary of the problems, solutions and stakeholders involved in mitigating bias in algorithmic systems.

## 6. References

[1] Behnoush Abdollahi and Olfa Nasraoui. 2018. Transparency in Fair Machine Learning: the Case of Explainable Recommender Systems. In Human and Machine Learning. Springer, Cham, pp 21–35. https://doi.org/10.1007/978-3-319-90403-0_2

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access 6 (2018), 52138–52160.

[3] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook’s ad delivery can lead to skewed outcomes. arXiv:1904.02095 [cs] (April 2019). http://arxiv.org/abs/1904.02095 arXiv: 1904.02095.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. ProPublica, May 23 (2016), 2016.

[5] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical Bias Removal for Hate Speech Detection Task Using Knowledge-based Generalizations. In The World Wide Web Conference (WWW ’19). ACM, New York, NY, USA, 49–59. https://doi.org/10.1145/3308558.3313504 event-place: San Francisco, CA, USA.

[6] Judit Bar’ Ilan, Kevin Keenoy, Mark Levene, and Eti Yaari. 2009. Presentation bias is significant in determining user preference for search results—A user study. Journal of the American Society for Information Science and Technology 60, 1 (2009), 135–149. https://doi.org/10.1002/asi.20941

[7] Pinar Barlas, Styliani Kleanthous, Kyriakos Kyriakou, and Jahna Otterbacher. 2019. What Makes an Image Tagger Fair?. In Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization (UMAP ’19). ACM, NewYork, NY, USA, 95–103. https://doi.org/10.1145/3320435.3320442 event-place: Larnaca, Cyprus.

[8] Shariq Bashir and Andreas Rauber. 2011. On the relationship between query characteristics and IR functions retrieval bias. Journal of the American Society for Information Science and Technology 62, 8 (2011), 1515–1532. https://doi.org/10.1002/asi.21549

[9] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv:1810.01943 [cs] (Oct. 2018). http://arxiv.org/abs/1810.01943 arXiv: 1810.01943.

[10] Alejandro Bellogin, Pablo Castells, and Ivan Cantador. 2017. Statistical Biases in Information Retrieval Metrics for Recommender Systems. Inf. Retr. 20, 6 (Dec. 2017), 606–634. https://doi.org/10.1007/s10791-017-9312-z

[11] Reuben Binnis, Max Van Kleek, Veale Michael, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. ’It’s Reducing a Human Being to a Percentage’: Perceptions of Justice in Algorithmic Decisions. In CHI ’18 Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM New York, NY, USA ’©2018. https://doi.org/10.1145/3173574.3173951

[12] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry Jackel, Urs Muller, and Karol Zieba. 2016. Visualbackprop: visualizing cnns for autonomous driving. arXiv:1611.05418 [cs] (Nov. 2016). http://arxiv.org/abs/1611.05418 arXiv: 1611.05418.

[13] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In In Advances in neural information processing systems. Curran Associates Inc. , USA ’©2016, Barcelona, Spain, pp. 4349–4357.

[14] Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, and Claudia Hauff. 2019. SIREN: A Simulation Framework for Understanding the Effects of Recommender Systems in Online News Environments. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* ’19). ACM, New York, NY, USA, 150–159. https://doi.org/10.1145/3287560.3287583 event-place: Atlanta, GA, USA.

[15] Olcay Boz. 2002. Extracting Decision Trees from Trained Neural Networks. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’02). ACM, New York, NY, USA, 456–461. https://doi.org/10.1145/775047.775113 event-place: Edmonton, Alberta, Canada.

[16] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI ’19). ACM, New York, NY, USA, 41:1–41:12. https://doi.org/10.1145/3290605.3300271 event-place: Glasgow, Scotland Uk.

[17] C. E. Buckley, Darrin L. Dimmick, Ian M. Soboroff, and Ellen M. Voorhees. 2007. Bias and the Limits of Pooling for Large Collections | NIST. Information Retrieval (July 2007). https://www.nist.gov/publications/bias-and-limits-pooling-large-collections

[18] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery 21, 2 (2010), 277–292.

[19] Ewa S. Callahan and Susan C. Herring. 2011. Cultural bias in Wikipedia content on famous persons. Journal of the American Society for Information Science and Technology 62, 10 (2011), 1899–1915. https://doi.org/10.1002/asi.21577

[20] Dallas Card, Michael Zhang, and Noah A. Smith. 2019. Deep Weighted Averaging Classifiers. ACM New York, NY, USA '©2019, Atlanta, GA, USA, pp. 369–378. https://doi.org/10.1145/3287560.3287595

[21] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. ACM New York, NY, USA ©2019, Atlanta, GA, USA, Pp. 319–328. https://doi.org/10.1145/3287560.3287586

[22] Abhijnan Chakraborty, Johnnatan Messias, Fabricio Benevenuto, Saptarshi Ghosh, Niloy Ganguly, and Krishna P. Gummadi. 2017. Who Makes Trends? Understanding Demographic Biases in Crowdsourced Recommendations. In Eleventh International AAAI Conference on Web and Social Media. pp. 22–31. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15680

[23] Abhijnan Chakraborty, Gourab K. Patro, Niloy Ganguly, Krishna P. Gummadi, and Patrick Loiseau. 2019. Equality of Voice: Towards Fair Representation in Crowdsourced Top-K Recommendations. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). ACM, New York, NY, USA, 129–138. https://doi.org/10.1145/3287560.3287570 event-place: Atlanta, GA, USA.

[24] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the Impact of Gender on Rank in Resume Search Engines. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, 651:1–651:14. https: //doi.org/10.1145/3173574.3174225.

[25] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C. Kanjirathinkal, and Mohan Kankanhalli. 2019. MMALFM: Explainable Recommendation by Leveraging Reviews and Images. ACM Trans. Inf. Syst. 37, 2 (Jan. 2019), 16:1–16:28. https://doi.org/10.1145/3291060.

[26] Andrea Chipman, Edward I. George, R. E.F, and McCullochDepartment. 2007. Making sense of a forest of treesH.

[27] Junghoo Cho and Sourashis Roy. 2004. Impact of Search Engines on Page Popularity. In Proceedings of the 13th International Conference on World Wide Web (WWW '04). ACM, New York, NY, USA, 20–29. https://doi.org/10.1145/988672.988676 event-place: New York, NY, USA.

[28] Bo Cowgill and Catherine Tucker. 2017. Algorithmic bias: A counterfactual perspective. Technical Report. Working Paper: NSF Trustworthy Algorithms. 3 pages.

[29] Mark Craven and JudeW. Shavlik. 1994. Using Sampling and Queries to Extract Rules from Trained Neural Networks. In Proceedings of the Eleventh International Conference on International Conference on Machine Learning (ICML'94). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 37–45. http://dl.acm.org/citation.cfm?id=3091574.3091580 event-place: New Brunswick, NJ, USA.

[30] Mark Craven and Jude W. Shavlik. 1996. Extracting Tree-Structured Representations of Trained Networks. In Advances in Neural Information Processing Systems 8. MIT Press, 24–30. http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf

[31] Maitraye Das, Brent Hecht, and Darren Gergle. 2019. The Gendered Geography of Contributions to OpenStreetMap: Complexities in Self-Focus Bias. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). ACM, New York, NY, USA, 563:1–563:14. https://doi.org/10.1145/3290605.3300793 event-place: Glasgow, Scotland Uk.

[32] A. Datta, S. Sen, and Y. Zick. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In 2016 IEEE Symposium on Security and Privacy (SP). 598–617. https://doi.org/10.1109/SP.2016.42

[33] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In Eleventh International AAAI Conference on Web and Social Media. Montreal, Canada, 512 – 515. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665

[34] Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing Age-Related Bias in Sentiment Analysis.

In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, 412:1–412:14. https://doi.org/10.1145/3173574.3173986 event-place: Montreal QC, Canada.

[35] Christos Dimitrakakis, Yang Liu, David Parkes, and Goran Radanovic. 2018. Bayesian Fairness. https://hal.inria.fr/hal-01953311

[36] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). ACM, New York, NY, USA, 67–73. https://doi.org/10.1145/3278721.3278729 event-place: New Orleans, LA, USA.

[37] Pedro Domingos. 1998. Knowledge discovery via multiple models. Intelligent Data Analysis 2, 1-4 (Jan. 1998), 187–202. https://doi.org/10.1016/S1088-467X(98)00023-7

[38] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference. 214–226.

[39] Benjamin Edelman, Michael Luca, and Dan Svirsky. 2017. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. American Economic Journal: Applied Economics 9, 2 (April 2017), 1–22. https://doi.org/10.1257/app.20160213

[40] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In 23rd International Conference on Intelligent User Interfaces (IUI '18). ACM, New York, NY, USA, 211–223.

[41] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18). ACM, New York, NY, USA, 162–170. https://doi.org/10.1145/3159652.3159654 event-place: Marina Del Rey, CA, USA.

[42] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In Conference on Fairness, Accountability and Transparency. 172–186. http://proceedings.mlr.press/v81/ekstrand18b.html

[43] Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the Search Engine Manipulation Effect (SEME). Proc. ACM Hum.-Comput. Interact. 1, CSCW (Dec. 2017), 42:1–42:22. https://doi.org/10.1145/3134677

[44] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful; Things can beworse than they appear" – Understanding biased algorithms and users' behavior around them in rating

platforms. In Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017. AAAI Press, 62–71.

[45] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes Towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). ACM, New York, NY, USA, 494:1–494:14. https://doi.org/10.1145/3290605.3300724 event-place: Glasgow, Scotland UK.

[46] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). ACM, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311 event-place: Sydney, NSW, Australia.

[47] Ruth Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In 2017 IEEE International Conference on Computer Vision (ICCV). 3449–3457. https://doi.org/10.1109/ICCV.2017.371 arXiv: 1704.03296.

[48] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. ACM Transactions on Information Systems (TOIS) 14, 3 (1996), 330–347.

[49] Gerhard Friedrich and Markus Zanker. 2011. A taxonomy for generating explanations in recommender systems. AI Magazine 32, 3 (2011), 90–98.

[50] George W Furnas. 2006. A fisheye follow-up: further reflections on focus+ context. In Proceedings of the SIGCHI conference on Human Factors in computing systems. 999–1008.

[51] Robert D. Gibbons, Giles Hooker, Matthew D. Finkelman, David J. Weiss, Paul A. Pilkonis, Ellen Frank, Tara Moore, and David J. Kupfer. 2013. The CAD-MDD: A Computerized Adaptive Diagnostic Screening Tool for Depression. The Journal of clinical psychiatry 74, 7 (July 2013), 669–674. https://doi.org/10.4088/JCP.12m08338.

[52] Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. 2010. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. 1–9. https://doi.org/10.1109/INFCOM.2010.5462078.

[53] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19). ACM, New York, NY, USA, 90–99. https://doi.org/10.1145/3287560.3287563 event-place: Atlanta, GA, USA.

[54] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In Proceedings of the 2018 World Wide Web Conference (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 903–912. https://doi.org/10.1145/3178876.3186138 event-place: Lyon, France.

[55] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. Technical Report. http://arxiv.org/abs/1805.10820 arXiv: 1805.10820.

[56] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51, 5 (Aug. 2018), 93:1–93:42. https://doi.org/10.1145/3236009.

[57] Anik3 Hannk, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW ’17). ACM, New York, NY, USA, 1914–1933. https://doi.org/10.1145/2998181.2998327 event-place: Portland, Oregon, USA.

[58] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. Advances in neural information processing systems (Oct. 2016). http://arxiv.org/abs/1610.02413 arXiv: 1610.02413.

[59] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 87 (Feb. 2014), 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067.

[60] Andreas Henelius, Kai Puolamaki, Henrik Bostrom, Lars Asker, and Panagiotis Papapetrou. 2014. A peek into the black box: exploring classifiers by randomization. Data Mining and Knowledge Discovery 28, 5-6 (Sept. 2014), 1503–1529. https://doi.org/10.1007/s10618-014-0368-8.

[61] Birger Hjørland. 2002. Domain analysis in information science. Journal of documentation (2002).

[62] Robert R Hoffman and Gary Klein. 2017. Explaining explanation, part 1: theoretical foundations. IEEE Intelligent Systems 32, 3 (2017), 68–73.

[63] Kajta Hofmann, Bhaskar Mitra, Filip Radlinski, and Milad Shokouhi. 2014. An Eye-tracking Study of User Interactions with Query Auto Completion. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM ’14). ACM, New York, NY, USA, 549–558. https://doi.org/10.1145/2661829.2661922 event-place: Shanghai, China.

[64] Benjamin D. Horne, Dorit Nevo, John O’Donovan, Jin-Hee Cho, and Sibel Adali. 2019. Rating Reliability and Bias in News Articles: Does AI Assistance Help Everyone?. In In Proceedings of the International AAAI Conference on Web and Social Media (1), Vol. 13. 247 – 256.

[65] Desheng Hu, Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2019. Auditing the Partisanship of Google Search Snippets. In The World Wide Web Conference (WWW ’19). ACM, New York, NY, USA, 693–704. https://doi.org/10.1145/3308558.3313654 event-place: San Francisco, CA, USA.

[66] Bernard J. Jansen and Marc Resnick. 2006. An Examination of Searcher’s Perceptions of Nonsponsored and Sponsored Links During Ecommerce Web Searching. J. Am. Soc. Inf. Sci. Technol. 57, 14 (Dec. 2006), 1949–1961. https://doi.org/10.1002/asi.v57:14.

[67] Ulf Johansson and Lars Niklasson. 2009. Evolving decision trees using oracle guides. In 2009 IEEE Symposium on Computational Intelligence and Data Mining. IEEE, Nashville, TN, USA, 238–244. https://doi.org/10.1109/CIDM.2009.4938655

[68] James E. Johndrow and Kristian Lum. 2019. An algorithm for removing sensitive information: Application to race-independent recidivism prediction. The Annals of Applied Statistics 13, 1 (March 2019), 189–220. https://doi.org/10.1214/18-AOAS1201

[69] Isaac Johnson, Connor McMahon, Johannes Schoning, and Brent Hecht. 2017. The Effect of Population and "Structural" Biases on Social Media-based Algorithms: A Case Study in Geolocation Inference across the Urban-Rural Spectrum. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI ’17). ACM, New York, NY, USA, 1167–1178. https://doi.org/10.1145/3025453.3026015 event-place: Denver, Colorado, USA.

[70] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In 2009 2nd International Conference on Computer, Control and Communication. IEEE, 1–6.

[71] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In Machine Learning and Knowledge Discovery in Databases (Lecture Notes in Computer Science). Springer Berlin Heidelberg, 35–50.

[72] Chen Karako and Putra Manggala. 2018. Using Image Fairness Representations in Diversity-Based Re-ranking for Recommendations. In Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP '18). ACM, New York, NY, USA, 23–28. https://doi.org/10.1145/3213586.3226206 event-place: Singapore, Singapore.

[73] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15). ACM, New York, NY, USA, 3819–3828. https://doi.org/10.1145/2702123.2702520 event-place: Seoul, Republic of Korea.

[74] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. Proc. ACM Hum.-Comput. Interact. 2, CSCW (Nov. 2018), 88:1–88:22. https://doi.org/10.1145/3274357.

[75] Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. 2019. Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality. The World Wide Web Conference on - WWW '19 (2019), 2907–2914. https://doi.org/10.1145/3308558.3313559 arXiv: 1903.11719.

[76] Niki Kilbertus, Adrià Gascón, Matt J. Kusner, Michael Veale, Krishna P. Gummadi, and Adrian Weller. 2018. Blind Justice: Fairness with Encrypted Sensitive Attributes. arXiv:1806.03281 [cs, stat] (June 2018). http://arxiv.org/abs/1806.03281 arXiv: 1806.03281.

[77] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Proceedings of Innovations in Theoretical Computer Science (ITCS), 2017. http://arxiv.org/abs/1609.05807

[78] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. 2015. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In Proceedings of the 2015 Internet Measurement Conference (IMC '15). ACM, New York, NY, USA, 121–127. https://doi.org/10.1145/2815675.2815714 event-place: Tokyo, Japan.

[79] Christie Kodama, Beth St Jean, Mega Subramaniam, and Natalie Greene Taylor. 2017. There'sa creepy guy on the other end at Google!: engaging middle school students in a drawing activity to elicit their mental models of Google. Springer Netherlands 20, 5 (Oct. 2017), 403–432. https://doi.org/10.1007/s10791-017-9306-x

[80] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized Explanations for Hybrid Recommender Systems. In Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19). ACM, New York, NY, USA, 379–390.

https://doi.org/10.1145/ 3301275.3302306 event-place: Marina del Ray, California.

[81] R. Krishnan, G. Sivakumar, and P. Bhattacharya. 1999. Extracting decision trees from trained neural networks. Pattern Recognition 32, 12 (Dec. 1999), 1999–2009. https://doi.org/10.1016/S0031-3203(98)00181-2

[82] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. FARE: Diagnostics for Fair Ranking Using Pairwise Error Metrics. In The World Wide Web Conference (WWW '19). ACM, New York, NY, USA, 2936–2942. https://doi.org/10.1145/3308558.3313443 event-place: San Francisco, CA, USA.

[83] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17). ACM, New York, NY, USA, 417–432. https://doi.org/10.1145/2998181.2998321 event-place: Portland, Oregon, USA.

[84] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In Advances in Neural Information Processing Systems 30. Curran Associates, Inc., 4066–4076.

[85] Rodrigo L. Cardoso, Wagner Meira Jr., Virgilio Almeida, and Mohammed J. Zaki. 2019. A Framework for Benchmarking Discrimination-Aware Models in Machine Learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). ACM, New York, NY, USA, 437–444. https://doi.org/10.1145/3306618.3314262 event-place: Honolulu, HI, USA.

[86] Huyen Le, Raven Maragh, Brian Ekdale, Andrew High, Timothy Havens, and Zubair Shafiq. 2019. Measuring Political Personalization of Google News Search. In The World Wide Web Conference (WWW '19). ACM, New York, NY, USA, 2957–2963. https://doi.org/10.1145/3308558.3313682 event-place: San Francisco, CA, USA.

[87] Susan Leavy. 2018. Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. In Proceedings of the 1st International Workshop on Gender Equality in Software Engineering (GE '18). ACM, New York, NY, USA, 14–16. https://doi.org/10.1145/3195570. 3195580 event-place: Gothenburg, Sweden.

[88] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion- Based Social Division. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17). ACM, New York, NY, USA, 1035–1048. https://doi.org/10.1145/2998181.2998230 event-place: Portland, Oregon, USA.

[89] Y. L. Lin, C. Trattner, P. Brusilovsky, and D. He. 2015. The impact of image descriptions on user tagging behavior: A study of the nature and functionality of crowdsourced tags. Journal of the Association for Information Science and Technology 66 (Sept. 2015), 1785–1798. http://dscholarship.pitt.edu/25927/

[90] Jianjun Lu, Shozo Tokinaga, and Yoshikazu Ikeda. 2006. Explanatory rule extraction based on the trained neural network and the genetic programming. https://doi.org/10.15807/jorsj.49.66

[91] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. K-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11). Association for Computing Machinery, New York, NY, USA, 502–510. https://doi.org/10.1145/2020408.2020488

[92] Nishtha Madaan, Sameep Mehta, Taneea Agrawaal, Vrinda Malhotra, Aditi Aggarwal, Yatin Gupta, and Mayank Saxena. 2018. Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies. In Conference on Fairness, Accountability and Transparency. 92–105. http://proceedings.mlr.press/v81/madaan18a.html

[93] Gabriel Magno, Camila Souza Araujo, Wagner Meira Jr., and Virgilio Almeida. 2016. Stereotypes in Search Engine Results: Understanding The Role of Local and Global Factors. arXiv:1609.05413 [cs] (Sept. 2016). http://arxiv.org/abs/1609.05413 arXiv: 1609.05413.

[94] David Martens, Jan Vanthienen, Wouter Verbeke, and Bart Baesens. 2011. Performance of classification models from a user perspective. Decision Support Systems 51, 4 (2011), 782–793.

[95] Maria Matsangidou and Jahna Otterbacher. 2019. What Is Beautiful Continues to Be Good. In Human-Computer Interaction ’€“ INTERACT 2019 (Lecture Notes in Computer Science). Springer International Publishing, 243–264.

[96] David Maxwell, Leif Azzopardi, and Yashar Moshfeghi. 2019. The impact of result diversification on search behaviour and performance. Information Retrieval Journal (May 2019), 1 – 25. https://doi.org/10.1007/s10791-019-09353-0

[97] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019).

[98] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the Trade-off Between Relevance, Fairness & Satisfaction in Recommendation Systems. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM ’18). ACM, New York, NY, USA, 2243–2251. https://doi.org/10.1145/3269206.3272027 event-place: Torino, Italy.

[99] Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1. 2930–2939. https://doi.org/10.1109/CVPR.2016.320

[100] Bhaskar Mitra, Milad Shokouhi, Filip Radlinski, and Katja Hofmann. 2014. On user interactions with query auto-completion. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR ’14. ACM Press, Gold Coast, Queensland,Australia, 1055–1058. https://doi.org/10.1145/2600428.2609508

[101] Abbe Mowshowitz and Akira Kawaguchi. 2005. Measuring Search Engine Bias. Inf. Process. Manage. 41, 5 (Sept. 2005), 1193–1205. https://doi.org/10.1016/j.ipm.2004.05.005

[102] Ingrid Nunes and Dietmar Jannach. 2017. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. User Modeling and User-Adapted Interaction 27, 3-5 (Dec. 2017), 393–444. https://doi.org/10.1007/s11257-017-9195-0

[103] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Frontiers in Big Data 2, 3 (2019). https://doi.org/10.3389/fdata.2019.00013

[104] J. Otterbacher, J. Bates, and P. D. Clough. 2017. Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3025453.3025727

[105] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating User Perception of Gender Bias in Image Search: The Role of Sexism. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR ’18). ACM, New York, NY, USA, 933–936. https://doi.org/10.1145/3209978.3210094 event-place: Ann Arbor, MI, USA.

[106] Aditya Pal, F. Maxwell Harper, and Joseph A. Konstan. 2012. A Exploring Question Selection Bias to Identify Experts and Potential Experts in Community Question Answering. ACM Transactions on Information Systems (TOIS) 30, 2 (Jan. 2012), 10. https://doi.org/10.1145/0000000.0000000

[107] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. Integrating Induction and Deduction for Finding Evidence of Discrimination. In Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL ’09). ACM, New York, NY, USA, 157–166. https://doi.org/10.1145/1568234.1568252 event-place: Barcelona, Spain.

[108] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. Measuring discrimination in socially-sensitive decision records. In Proceedings of the 2009 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics. 581–592.

[109] Brett Poulin, Roman Eisner, Duane Szafron, Paul Lu, Russ Greiner, D S Wishart, Alona Fyshe, Brandon Pearcy, Cam MacDonell, and John Anvik. 2006. Visual Explanation of Evidence in Additive Classifiers. In In Proceedings of the National Conference on Artificial Intelligence, Vol. 21. 8.

[110] Giovanni Quattrone, Licia Capra, and Pasquale De Meo. 2015. There’s No Such Thing As the Perfect Map: Quantifying Bias in Spatial Crowdsourcing Datasets. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW ’15). ACM, New York, NY, USA, 1021–1032. https://doi.org/10.1145/2675133.2675235 event-place: Vancouver, BC, Canada.

[111] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations As Mechanisms for Supporting Algorithmic Transparency. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18). ACM, New York, NY, USA, 103:1–103:13. https://doi.org/10.1145/3173574.3173677 event-place: Montreal QC, Canada.

[112] Abdelhalim Rafrafi, Vincent Guigue, and Patrick Gallinari. 2012. Coping with the Document Frequency Bias in Sentiment Classification. In ICWSM.

[113] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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. ACM, San Francisco, CA, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778 arXiv: 1602.04938.

[114] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High Precision Model-Agnostic Explanations. In In Thirty-Second AAAI Conference on Artificial Intelligence. 9.

[115] Greg Ridgeway. 2020. Transparency, Statistics, and Justice System Knowledge Is Essential for Science of Risk Assessment. Harvard Data Science Review 2, 1 (31 1 2020). https://doi.org/10.1162/99608f92.cb0f8674 https://hdsr.mitpress.mit.edu/pub/vu6rc1yv.

[116] Ronald E. Robertson, Lisa Friedland, Kenneth JOSEPH, David Lazer, Christo Wilson, and Shan Jiang. 2018. Auditing Partisan Audience Bias within Google Search. In Proceedings of the ACM on Human-Computer Interaction, Vol. 2. 1–22. https://doi.org/10.1145/3274417

[117] Ronald E. Robertson, Shan Jiang, David Lazer, and Christo Wilson. 2019. Auditing Autocomplete: Suggestion Networks and Recursive Algorithm Interrogation. In Proceedings of the 10th ACM Conference on Web Science (WebSci ’19). ACM, New York, NY, USA, 235–244. https://doi.org/10.1145/3292522.3326047 event-place: Boston, Massachusetts, USA.

[118] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29, 5 (Nov. 2014), 582–638. https://doi.org/10.1017/S0269888913000039

[119] Alex Rosenblat and Luke Stark. 2016. Algorithmic Labor and Information Asymmetries: A Case Study of Uber'€™s Drivers. SSRN Scholarly Paper ID 2686227. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=2686227

[120] Cynthia Rudin, Caroline Wang, and Beau Coker. 2020. The Age of Secrecy and Unfairness in Recidivism Prediction. Harvard Data Science Review 2, 1 (31 1 2020). https://doi.org/10.1162/99608f92.6ed64b30 https://hdsr.mitpress.mit.edu/pub/7z10o269.

[121] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, and Rayid Ghani. 2018. Aequitas: A Bias and Fairness Audit Toolkit. arXiv:1811.05577 [cs] (Nov. 2018). http://arxiv.org/abs/1811.05577 arXiv: 1811.05577.

[122] Joni Salminen, Soon-Gyo Jung, and Bernard J. Jansen. 2019. Detecting Demographic Bias in Automatically Generated Personas. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19). ACM, New York, NY, USA, LBW0122:1–LBW0122:6. https://doi.org/10.1145/3290607.3313034 event-place: Glasgow, Scotland Uk.

[123] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In Data and Discrimination: Converting Critical Concerns into Productive Inquiry," a preconference at the 64th Annual Meeting of the International Communication Association. Seattle, WA.

[124] Vitaly Schetinin, Jonathan E. Fieldsend, Derek Partridge, Timothy J. Coats, Wojtek J. Krzanowski, Richard M. Everson, Trevor C. Bailey, and Adolfo Hernandez. 2007. Confident Interpretation of Bayesian Decision Tree Ensembles for Clinical Applications. IEEE Transactions on Information Technology in Biomedicine 11, 3 (May 2007), 312–319. https://doi.org/10.1109/TITB.2006.880553

[125] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, Venice, 618–626. https://doi.org/10.1109/ICCV.2017.74

[126] Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of Extractive Text Summarization. In Companion Proceedings of the The Web Conference 2018 (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 97–98. https://doi.org/10.1145/3184558.3186947 event-place: Lyon, France.

[127] Judy Hanwen Shen, Lauren Fratamico, Iyad Rahwan, and Alexander M. Rush. 2018. Darling or Babygirl? Investigating Stylistic Bias in Sentiment Analysis. In 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning. Stockholm, Sweden.

[128] Donghee Shin and Yong Jin Park. 2019. Role of fairness, accountability, and transparency in algorithmic affordance. Computers in Human Behavior 98 (Sept. 2019), 277–284. https://doi.org/10.1016/j.chb.2019.04.019

[129] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034 [cs] (Dec. 2013). http://arxiv.org/abs/1312.6034 arXiv: 1312.6034.

[130] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). ACM, New York, NY, USA, 2219–2228. https://doi.org/10.1145/3219819.3220088 event-place: London, United Kingdom.

[131] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Ribeiro, George Arvanitakis, Fabriccio Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Potential for Discrimination in Online Targeted Advertising. In FAT 2018 - Conference on Fairness, Accountability, andTransparency, Vol. 81. New-York, United States, 1–15. https://hal.archives-ouvertes.fr/hal-01955343

[132] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD ’19). Association for Computing Machinery, New York, NY, USA, 2459–2468. https://doi.org/10.1145/3292500.3330664

[133] Erik Strumbelj and Igor Kononenko. 2010. An Efficient Explanation of Individual Classifications using Game Theory. Journal of Machine Learning Research 11 (Jan. 2010), 1–18. https://doi.org/10.1145/1756006.1756007

[134] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. Queue 11, 3 (March 2013), 10:10–10:29. https://doi.org/10.1145/2460276.2460278

[135] Hui Fen Tan, Giles Hooker, and Martin T.Wells. 2016. Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable. arXiv:1611.07115 [cs, stat] (Nov. 2016).

[136] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2017. Detecting Bias in Black-Box Models Using Transparent Model Distillation. ArXiv abs/1710.06169 (2017).

[137] Mike Thelwall and Nabeil Maflahi. 2015. Are scholarly articles disproportionately read in their own country? An analysis of mendeley readers. Journal of the Association for Information Science and Technology 66, 6 (June 2015), 1124–1135. https://doi.org/10.1002/asi.23252

[138] N. Tintarev and J. Masthoff. 2007. A Survey of Explanations in Recommender Systems. In 2007 IEEE 23rd International Conference on Data Engineering Workshop. 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

[139] Ryan Turner. 2016. A model explanation system. In 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). 1–6. https://doi.org/10.1109/MLSP.2016.7738872

[140] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. 2013. Visualizing recommendations to support exploration, transparency and controllability. In Proceedings of the 2013 international conference on Intelligent user interfaces. 351–362.

[141] Marina M.-C. Vidovic, Nico Görnitz, Klaus-Robert Muller, and Marius Kloft. 2016. Feature Importance Measure for Non-linear Learning Algorithms. In NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems. Barcelona, Spain. http://arxiv.org/abs/1611.07567 arXiv: 1611.07567.

[142] Nicholas Vincent, Isaac Johnson, Patrick Sheehan, and Brent Hecht. 2019. Measuring the Importance of User-Generated Content to Search Engines. In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 13. 505–5016.

[143] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable Recommendation via Multi-Task Learning in Opinionated Text Data. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR ’18). ACM, New York, NY, USA, 165–174. https://doi.org/10.1145/3209978.3210010 event-place: Ann Arbor, MI, USA.

[144] Ingmar Weber and Carlos Castillo. 2010. The demographics of web search. In In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA, 523–530. https://doi.org/10.1145/1835449.1835537

[145] Ryen W. White. 2014. Belief dynamics in web search. Association for Information Science and Technology 65, 11 (Nov. 2014), 2165–2178. https://doi.org/10.1002/asi.23128

[146] Ryen W. White and Eric Horvitz. 2015. Belief Dynamics and Biases in Web Search. ACM Transactions on Information Systems 33, 4 (May 2015), 1–46. https://doi.org/10.1145/2746229

[147] Colin Wilkie and Leif Azzopardi. 2014. Best and Fairest: An Empirical Analysis of Retrieval System Bias. In In European Conference on Information Retrieval (LNCS), Vol. 8416. Springer, Cham, 13–25. https://doi.org/10.1007/978-3-319-06028-6_2

[148] Colin Wilkie and Leif Azzopardi. 2014. A Retrievability Analysis: Exploring the Relationship Between Retrieval Bias and Retrieval Performance. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. Shanghai, China, 81–90. https://doi.org/10.1145/2661829.2661948

[149] Colin Wilkie and Leif Azzopardi. 2017. Algorithmic Bias: Do Good Systems Make Relevant Documents More Retrievable?. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore, 2375–2378. https://doi.org/10.1145/3132847.3133135

[150] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI ’18). ACM, New York, NY, USA, 656:1–656:14. https://doi.org/10.1145/3173574.3174230 event-place: Montreal QC, Canada.

[151] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. On Convexity and Bounds of Fairness-aware Classification. In The World Wide Web Conference. ACM New York, NY, USA ©2019, San Francisco, CA, USA, 3356–3362. https://doi.org/10.1145/3308558.3313723

[152] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. 2017. Fairness-Aware Group Recommendation with Pareto-Efficiency. In Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys ’17). ACM, New York, NY, USA, 107–115. https://doi.org/10.1145/3109859.3109887 event-place: Como, Italy.

[153] Kelvin Xu, Jimmy Ba Lei, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, S. Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on International Conference on Machine Learning, Vol. 37. Lille, France, 2048–2057. https://dl.acm.org/citation.cfm?id=3045336

[154] Elad Yom-Tov. 2019. Demographic differences in search engine use with implications for cohort selection | SpringerLink. Springer Netherlands (2019), 1–11. https://doi.org/10.1007/s10791-018-09349-2

[155] 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 (WWW ’17). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1171–1180. https://doi.org/10.1145/3038912.3052660

[156] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM ’17). ACM, New York, NY, USA, 1569–1578. https://doi.org/10.1145/3132847.3132938 event-place: Singapore, Singapore.

[157] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in Decision-Making: The Causal Explanation Formula. In Thirty-Second AAAI Conference on Artificial Intelligence. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16949

[158] Lu Zhang, Yongkai Wu, and Xintao Wu. 2016. Situation Testing-based Discrimination Discovery: A Causal Inference Approach. In Proceedings ofthe Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16). AAAI Press, 2718–2724. http://dl.acm.org/citation.cfm?id=3060832.3061001 event-place: New York, New York, USA.

[159] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A causal framework for discovering and removing direct and indirect discrimination. (2017).

[160] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 2). http://arxiv.org/abs/1804.06876 arXiv: 1804.06876.

[161] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Olivia, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2921–2929. https://arxiv.org/abs/1512.04150

[162] Zhi-Hua Zhou, Yuan Jiang, and Shi-Fu Chen. 2003. Extracting symbolic rules from trained neural network ensembles. AI Communications - Artificial Intelligence Advances in China 16, 1 (Jan. 2003), 3–15. https://dl.acm.org/citation.cfm?id=1218644

[163] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. 2017. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. 12. https://arxiv.org/abs/1702.04595

[164] Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. arXiv:1511.00148 [cs, stat] (Oct. 2015). http://arxiv.org/abs/1511.00148 arXiv: 1511.00148.

[165] Indr˙e Žliobait˙e and Bart Custers. 2016. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artificial Intelligence and Law 24, 2 (June 2016), 183–201. https://doi.org/10.1007/s10506-016-9182-5